BILKENT NEWS PORTAL:

A SYSTEM WITH NEW EVENT DETECTION AND TRACKING CAPABILITIES

A THESIS
SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING
AND THE INSTITUTE OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

By
Hüseyin Çağdaş Öcalan
May, 2009

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Fazlı Can (Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Seyit Koçberber (Co-Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Fabio Crestani (University of Lugano)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. H. Murat Karamüftüoğlu

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Özgür Ulusoy

Approved for the Institute of Engineering and Science:

Prof. Dr. Mehmet Baray

Director of the Institute

ABSTRACT

BILKENT NEWS PORTAL:
A SYSTEM WITH NEW EVENT DETECTION AND TRACKING CAPABILITIES

Hüseyin Çağdaş Öcalan
M.S. in Computer Engineering
Supervisors
Prof. Dr. Fazlı Can,
Asst. Prof. Dr. Seyit Koçerber
May, 2009

News portal services such as browsing, retrieving, and filtering have become an important research and application area as a result of information explosion on the Internet. In this work, we give implementation details of Bilkent News Portal that contains various novel features ranging from personalization to new event detection and tracking capabilities aiming at addressing the needs of news-consumers. The thesis presents the architecture, data and file structures, and experimental foundations of the news portal. For the implementation and evaluation of the new event detection and tracking component, we developed a test collection: BilCol2005. The collection contains 209,305 documents from the entire year of 2005 and involves several events in which eighty of them are annotated by humans. It enables empirical assessment of new event detection and tracking algorithms on Turkish. For the construction of our test collection, a web application, ETracker, is developed by following the guidelines of the TDT research initiative. Furthermore, we experimentally evaluated the impact of various parameters in information retrieval (IR) that has to be decided during the implementation of a news portal that provides filtering and retrieval capabilities. For this purpose, we investigated the effects of stemming, document length, query length, and scalability issues.

Keywords: Content based information filtering, Information retrieval (IR), New event detection and tracking, News portal, Test collection construction, TDT.

ÖZET

BİLKENT HABER PORTALI:
YENİ OLAY BELİRLEME VE İZLEME YETENEKLERİ OLAN BİR SİSTEM

Hüseyin Çağdaş Öcalan
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticileri
Prof. Dr. Fazlı Can
Yrd. Doç. Dr. Seyit Koçberber
Mayıs, 2009

Internet'deki bilgi patlaması sonucu gezinme, erişim ve süzme gibi haber portalı servisleri önemli araştırma ve uygulama alanları haline gelmiştir. Bu çalışmada, Internet'deki haber tüketicilerinin ihtiyaçlarına yönelik, kişiselleştirmeden yeni olay belirleme ve izlemeye kadar geniş bir yelpazede çeşitli özgün yetenekleri olan Bilkent Haber Portalı'nın tasarım ve geliştirme detayları verilmektedir. Tez, bu sistemin mimari tasarımını, veri ve dosya yapılarını ve deneysel temellerini sunmaktadır. Portalın yeni olay belirleme ve izleme bileşeninin geliştirilmesi ve değerlendirilmesi için bir deney derlemi oluşturulmuştur: BilCol2005. Bu deney derlemi 2005 yılına ait 209,305 haber ve seksen adedi kullanıcı tarafından değerlendirilmiş birçok olay içermektedir. Bu derlem, Türkçede yeni olay berlirleme ve izleme algoritmalarının deneysel olarak ölçülebilmesine olanak sağlamaktadır. Deney derleminin hazırlanabilmesi için TDT araştırma programının yönergeleri takip edilerek bir web uygulaması, ETracker, geliştirilmiştir. Ayrıca, bilgi erişimi ve süzme yetenekleri olan bu haber portalının gerçekleştirilmesinde kullanılacak çeşitli parametrelerin, bilgi erişimi üzerine etkileri deneysel olarak ölçülmüştür. Bu amaçla, kök bulma yöntemlerinin, doküman uzunluğunun, sorgu uzunluğunun etkileri ve ölçeklenebilirlik konuları incelenmiştir.

*Anahtar Kelimeler:*İçerik tabanlı bilgi süzme, Bilgi erişimi, Yeni olay belirleme ve izleme, Haber portalı, Deney derlemi oluşturulması, TDT

Acknowledgements

I want to express my deepest gratitude to Prof. Dr. Fazlı Can for his guidance and support during my research and study at Bilkent University. He provided us a distinctive research environment with his wisdom and appreciation. Without his leadership, encouragement, ideas and great personality this work would not have been possible. I am very glad to have a chance of working with him.

I would like to address my special thanks to Asst. Prof. Dr. Seyit Koçberber for his valuable suggestions and comments throughout this study.

I would like to thank Prof. Dr. Fabio Crestani (of University of Lugano), Asst. Prof. Dr. H. Murat Karamüftüoğlu, and Prof. Dr. Özgür Ulusoy for their valuable pointers.

Many thanks to my friends Özgür Bağlıoğlu, Süleyman Kardaş and Erkan Uyar who shared their three years in the way that we walked together. I would also like to thank Cihan Kaynak and Erman Balçık especially for their friendship and also for their valuable comments and contributions to this work.

I am grateful to Bilkent University for providing me research assistant scholarship for my MS study. I would also like to thank TÜBİTAK for its partial financial support of my thesis work under the grant number 106E014 ("Instant new event detection and tracking and retrospective event clustering in Turkish news using Web resources").

Above all, I am deeply thankful to my parents, who supported me in every decision that I made. Without their love and encouragement, this thesis would have never been completed.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The printing press technology enabled provision of news to many people in bulk in the 15th century. When the radio was born in the late 19th century, it changed the way of accessing the news. The television technology facilitated the animated news delivery in the middle of the 20th century. All of these technologies provided one way communication by delivering the news without caring for the individual needs of news-consumers. In all of these technologies reaching news is sequential. Towards the end of the 20th century, the computer and communication technologies came together in the form of the Internet and Web (see the timeline in Figure 1.1). With this new technology, a tremendous amount of information has become accessible throughout the world. Development of this technology also shaped the way of delivering news. It provided a true interactive media by enabling a request and response mechanism and made the news-consumers active participants in the news delivery process. Furthermore, these ICT (Information and Communication Technology) facilities make the current news immediately available with no time delay. Especially due to this reason, news-consumers refer to news portals rather than the traditional media.

Figure 1-1: News delivery technologies in years from printing press to the Internet.

1.1 Motivations

The amount of online news sources has increased dramatically in the last decade. The rapid increase in the number of online news sources is obviously due to the increased demand and the decreased cost of investment. However, the availability of large amount of news overwhelms news-consumers. Personalized news portals are introduced as a solution to this problem. They aim to deliver news in an effective (according to consumer needs) and efficient (effortless) manner. Their search facilities allow accessing not only current news but also news archives. Their filtering capabilities deliver latest news according to consumers' interest. As a final outcome, such systems aim to allow news-consumers creating their own virtual aggregated newspapers to receive news from multiple sources. Their inclusive, diverse, and neutral manner is another reason for the popularity of news portals.

1.2 Contributions

In this work, we present the implementation issues and the foundations of Bilkent News Portal [BIL2009] which provides unique capabilities when it is compared with similar systems. It is based on research on information retrieval, information filtering, duplicate elimination, and new event detection and tracking. The portal was built by cooperative work of a team. This thesis covers design issues, experimental foundations, and recommendations for its future large-scale implementation. The other functionalities provided in the portal such as near-duplicate news detection, and new event detection and tracking are briefly discussed in this thesis and covered in detail in the complementing works [BAG2009, KAR2009, UYA2009].

One of the major components of the Bilkent News Portal is the new event detection and tracking capability. For the implementation of this component we constructed a test collection to measure the performance of the algorithms developed for the implementation of this part of the news portal. For this purpose, a system for test collection preparation is implemented: ETracker. It is used to construct a test collection (BilCol2005) according to the traditions of the TDT (Topic Detection and Tracking) research program [TDT2002, TDT2004]. BilCol2005 is one of the significant contributions of this thesis, we plan to share it with other researchers.

Search engines have many parameters and concerns for providing effective and efficient retrieval services. Some of these, such as stemming, are language dependent, and others are general, such as scalability and the effects of document and query lengths. The best parameters for a small document collection may not give the best results when the size of the collection grows. Therefore, for implementing a system with a desirable effectiveness, we have performed a series of IR experiments. The results, supported by statistical tests, show that

- a stopword list has no influence on effectiveness;
- a simple word truncation approach and an elaborate lemmatizer-based stemmer provide similar performances in terms of effectiveness;
- longer queries improve effectiveness; however, it saturates as the query lengths become longer; and
- longer documents provide higher effectiveness.

1.3 Overview of the Thesis

The thesis is organized as follows; in Chapter 2, we provide a short survey of news portals. In Chapter 3, we describe the system architecture in terms of processes involved in its implementation; namely they are content extraction, indexing, new event detection and tracking, information retrieval, information filtering, news categorization, personalization, latest news selection, multi-document summarization, and near-duplicate detection processes. The data and file structures of the system are described in Chapter 4. This chapter presents incremental indexing which is essential in a news portal that provides news crawling and query processing facilities at the same time. It also illustrates RDBMS table and XML file structures. These technical details are explained to provide a starting point for the future enhancements on the portal. In the following two chapters, we explain the experimental foundations of the news portal: Chapter 5 provides ETracker application and gives a detailed description of BilCol2005; Chapter 6 reports the results of several information retrieval experiments. Chapter 7 covers the future pointers, which are our suggestions for building a large-scale implementation of the news portal. Chapter 8 concludes the work with a brief summary and a description of the contributions.

Chapter 2

Related Work

To attract Internet users many news portals have been developed. These systems solve the problem caused by information explosion and provide useful services. In this chapter, the systems that provide similar technologies are introduced and compared with Bilkent News Portal.

2.1 NewsBlaster from Columbia University

NewsBlaster from Columbia University crawls the web for news articles, clusters them, and generates summaries of these clusters by using a multi-document summarization algorithm [MCK2003].

NewsBlaster classifies the news articles into three levels. The system keeps tf-idf vectors for each category and compares them with the input documents' vectors to determine whether the processed news belongs to one of the predefined categories. Then the clusters are classified to which the largest number of news in the cluster are assigned. By this way, the clusters become hierarchically bounded together. Then, the agglomerative clustering with a groupwise average similarity function finds the similarities between these clusters. Finally, the level of classification is generated as cluster of articles (events), cluster of events, categories [MCK2002].

The constructors of the system state that although their clustering approach is similar to Topic Detection and Tracking (TDT) style, it uses learned, weighted combination of features to determine similarity of stories. In terms of originality, NewsBlaster represents a good combination of TDT and summarization. Summarization capability highlights this system among the similar ones, but it is not enough to make this system competitive. Changing user needs requires enhancements on existing systems. In that sense NewsBlaster does not contain popular features such as filtering and personalization. Bilkent News Portal provides personal solutions such as profiling and information filtering in addition to TDT.

2.2 NewsInEssence from University of Michigan

NewsInEssence is a system which is specialized on clustering and summarization as NewsBlaster. In the generic scenario that they explained, the user selects a single news story from a news web site, and then the system searches live sources for related stories. The findings are summarized according to a user defined compression ratio. Additionally, instead of selecting a story from a web site, the user can enter keywords or a URL to create a cluster [RAD2001].

The major advantage of NewsInEssence is allowing the user to create personalized clusters and summaries. The user interface of NewsInEssence is not user friendly, it is very complex and hard to use. When the complexity of the user interface is considered and the efficiency and stability problems are taken into account, NewsInEssence is far behind NewsBlaster, although it turns the major weakness of NewsBlaster into an advantage by providing personalization.

2.3 AllInOneNews of webscalers.com

The federated search systems can be grouped into two major categories in terms of searching methodology. The regular search engines crawls the documents from the Web and build a local index. During searching, they use the local index and display the results. The second type of search engines, called metasearch engines, send user's query to the external search engines and gather the results to display. Metasearch engines do not need a local index and search mechanism. These two different design approaches describe the general structure of the current search engines. AllInOneNews is a metasearch engine [LIU2007].

The significant advantage of metasearch engines is the capability to search large amount of news sources. It is obvious that with a good merging strategy they increase the accuracy of the results. However, it must be emphasized that the quality of the results is inevitably related to the quality of the duplicate detection and data fusion algorithms that are used in the system.

Crawling delay is one of the important concerns of regular news portals. If the data is not crawled frequently, the immediate access to the latest news cannot be possible. Metasearch engine-based news portals can resolve this problem by providing instant access to the search engines with fresh data. In metasearch engines the requirement of searching large amount of news sources increases query response time, since the total response time depends on the loads of the searched systems. Long response times reduce the user's interest. Frequently updated regular search engine can overcome these problems since it uses a local index.

The systems which use a metasearch engine cannot provide additional functionalities for searching. If the aim is serving multiple functionalities, using local index becomes a necessity. Bilkent News Portal works as a regular search engine and provides additional services such as filtering, categorization, and new event detection and tracking.

AllInOneNews is the largest news metasearch engine which uses more than 1,000 news sites over 150 countries. In terms of the capacity and the quality of the results, it is a significant contribution. It allows users to search news from multiple news sources. However, if the sources use forms with heavy Javascript, automatically connecting cannot be possible. So, if we consider the recent development in web programming technologies, it is not certain how long this approach will serve its purpose.

2.4 Europe Media Monitor from European Commission

European Media Monitor (EMM) was developed for the European Commission's Directorate General Communication to replace their traditional and expensive manual media monitoring services. It consists of three public web portals NewsBrief, NewsExplorer, MedlSys which receive 1.2 Million hits per day. EMM approximately monitors 1,200 news portals from all over the world and retrieves over 40,000 reports per day in 35 languages. NewsBrief (http://press.jrc.it/) shows the hottest topics of multiple news sources. It gathers and groups related articles as stories at every ten minutes. The Medical Information System MedISys (http://medusa.jrc.it/) displays health related articles by grouping into disease or disease type categories. NewsExplorer application (http://press.jrc.it/NewsExplorer/) analyses the EMM news articles, extracts information about people, organisations and locations, and links related news items over time and across languages [KHU2007].

The powerful part of this system is generating and visualizing the statistics about people, organizations and locations in the articles by using text mining. While the location of the event can be seen on the world map, the information about the people or organizations can be accessed.

2.5 Google News

Google News is one of the most popular systems. It generates a frequently updated local index by crawling news sources. While this thesis is being written, Google News crawls 25,000 news sources from all over the world in different languages. The number of sources in English is specified as more than 4,500 and the number sources in Turkish is given as at least 200 [GOG2009].

The information about Google News is limited. According to information that is available at its website, their system gathers news from the sources at every 15 minutes and then categorizes them to provide relevant news in story groups. This approach shows that it uses a new event detection and tracking algorithm. The crawled news articles are used for only ranking and evaluation purposes. When a user wants to read a news article, it redirects the user to the original news source. Ranking and evaluating process is based on certain characteristics of news content such as freshness, location, relevance, and diversity. These are the global parameters to all sites, but additionally Google News uses different weights for different sources using their "page rank algorithm."

Google personalized news access is based on collaborative filtering. Two features can be found in Google News when the user login the system. First one is "Search History" and the second one is "Recommended news for the active user." Search history allows the user to browse the news that is read in the past [DAS2007]. Although, the idea is to help the user, keeping track of every input could be annoying. Unwanted tracking is also possible for the case of news recommendation. Google News accepts each user click as a positive vote and performs filtering on new documents by using clicked ones. However note that counting each click as a positive vote is not as reliable as intentional positive votes. In our system, the user manually chooses the news articles to be used for profile update and does not feel as if he has been tracked by the system.

2.6 Microsoft Personalized Search

A group of researchers from Microsoft addresses the problem of personalization and study if personalization is effective on different queries for different users under different search contexts [DOU2007]. In their work, two types of the personalization are discussed and compared. They evaluated different methods of profile based personalization. They group the users which have similar interests, and personalize the query results according to users' inputs and browsing behaviors. This approach is obviously useful to recommend results to the users who have done a few searches. Otherwise, when the inputs are not enough to learn interest of users, filtering process could not be successful.

The evaluation results show that profile-based personalization have significant improvements over common web search on queries with larger click entropy. However, on the queries with small click entropy, they observed similar, or even worse performance than common web search. Then they concluded that, all queries should not be handled in the same manner.

In general, they emphasize that although profile-based personalized search strategies improve the quality of the results, click-based ones are more stable and straightforward. And also their suggestion is in the direction of using click-based personalization strategies. Bilkent News Portal is also using click-based personalization for filtering purpose, and we are planning to provide personalized search results in the near future.

Chapter 3

System Architecture

Information services have been growing rapidly with the rise of new technological innovations. The latest trend on the web is the collaboration of users. One common approach in this direction is personalization. The aim of personalization is making a user feel as an integral part of a website. Bilkent News Portal is designed with this in mind.

Bilkent News Portal is a collaborative work of Bilkent IR Group. Many applications have been developed as the results of comprehensive researches. This work covers combining these tools in a harmony within a real application. Not only user interfaces but also synchronization of the services has been designed as part of this work.

The members of the Bilkent IR group have developed the algorithms of Bilkent News Portal. The details of new event detection and tracking component are reported in the thesis of Süleyman Kardaş and Özgür Bağlıoğlu [KAR2009, BAG2009]. The near-duplicate detection component is presented in the thesis of Erkan Uyar [UYA2009]. Content extraction component is developed by Levent… and retrospective incremental clustering component is implemented by İsmet Zeki Yalnız. Multi document summarization component is designed by Gönenç Ercan and the experiments are published in [ERC2009]. In addition to integration of the services, information retrieval, information filtering and user interface applications are my responsibilities. For the indexing and document matching we used the Lemur Toolkit. Information filtering component is developed by using information retrieval component, it uses local index and matching functions of IR system. Similarly, incremental clustering component uses local index, and required modifications has been added to integrate the component. This chapter provides description of the system architecture by explain the functionalities of these components.

3.1 Design Overview of the System

Multi-source news portals, a relatively new technology, receive and gather news from several web news providers. These systems aim to make news more accessible, especially by providing event-based information organization. New event detection and tracking applications aim to prevent overwhelming of news-consumers by too many unconnected items [ALL2002]. We present the first personalizable Turkish news portal with such capabilities. The architecture of the system can be divided into two parts; web component and core component.

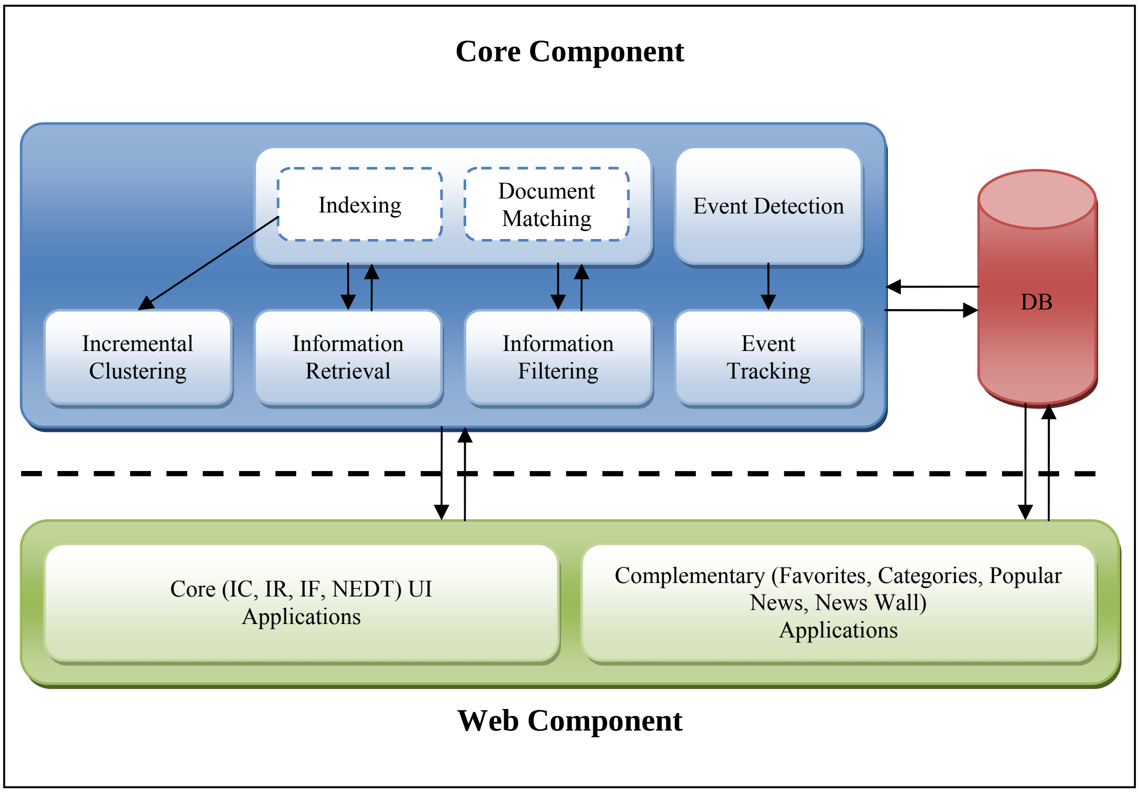
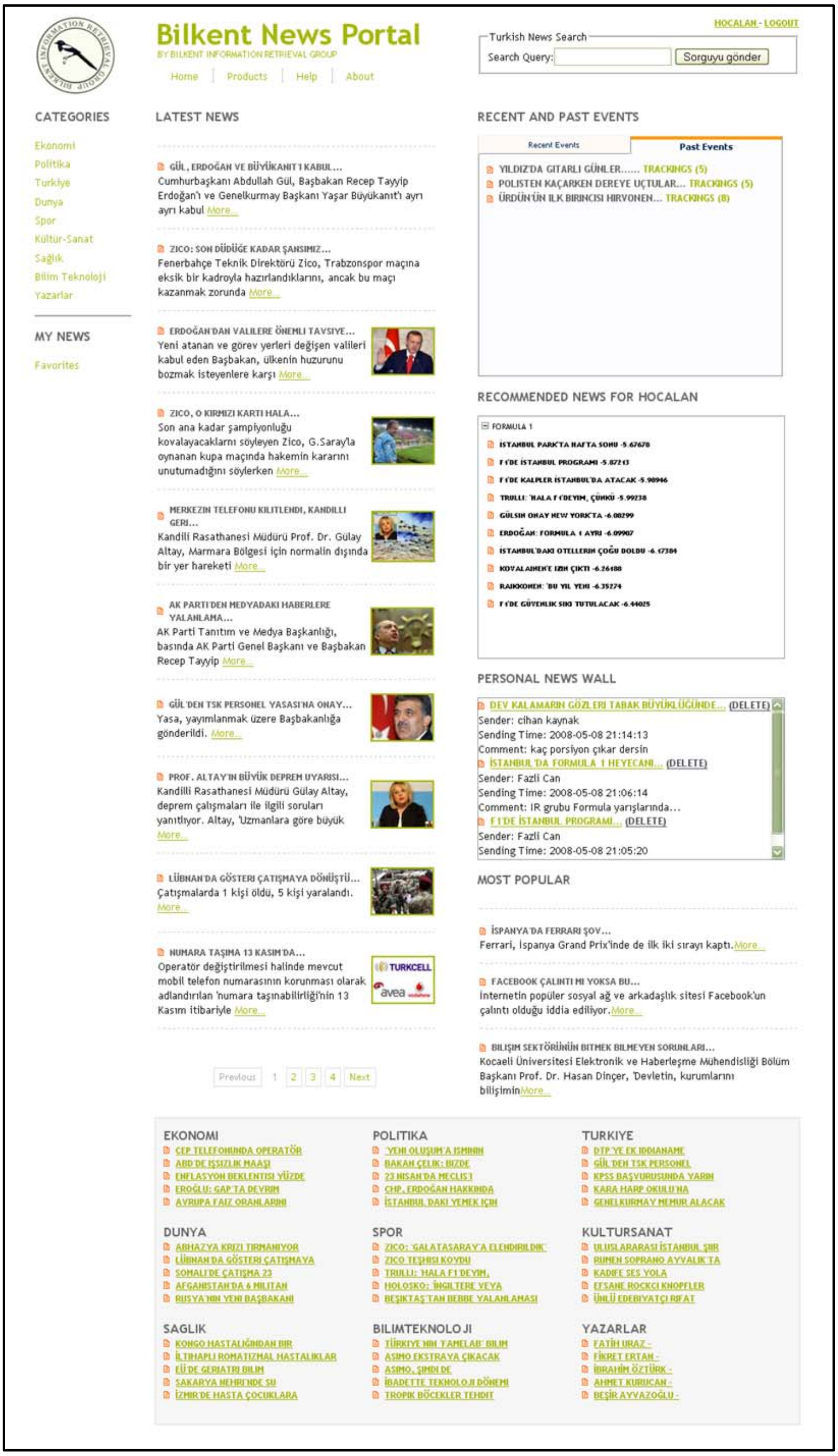


Figure 3-1: General system overview.

The web component receives inputs from the users and display the outputs formed upon user demands. It is developed by using PHP and AJAX (Asynchronous JavaScript and XML) technology. Additionally, to improve the user interface visualization, some tools from the X library is added to the system. The appearance settings are set by CSS (Cascade Style Sheets). The web component has direct access to the news database over PHP.

A minimalist, i.e., a simple design approach is followed while designing the portal. So, functionalities and ease of use are the key issues of the system.



Bilkent News Portal
BY BILKENT INFORMATION RETRIEVAL GROUP

Home | Products | Help | About

HOCALAN - LOGOUT

Turkish News Search
Search Query: [Sorguyu gönder]

CATEGORIES

Ekonomi
Politika
Turkiye
Dunya
Spor
Kültür-Sanat
Sağlık
Bilim Teknoloji
Yazarlar

MY NEWS

Favorites

LATEST NEWS

GÜL, ERDOĞAN VE BÜYÜKANIT'I KABUL...
Cumhurbaşkanı Abdullah Gül, Başbakan Recep Tayyip Erdoğan'ı ve Genelkurmay Başkanı Yaşar Büyükanıt'ı ayrı ayrı kabul More...

ZICO: SON DÜDÜĞE KADAR ŞANSIMIZ...
Fenerbahçe Teknik Direktörü Zico, Trabzonspor maçına eksik bir kadroyla hazırlandıklarını, ancak bu maçı kazanmak zorunda More...

ERDOĞAN'DAN VALILERE ÖNEMLI TAVSIYE...
Yeni atanan ve görev yerleri değişen valileri kabul eden Başbakan, ülkenin huzurunu bozmak isteyenlere karşı More...

ZICO, O KIRMIZI KARTI HALA...
Son ana kadar şampiyonluğu kovalayacaklarnı söyleyen Zico, G.Saray'la oynanan kupa maçında hakemin kararını unutumadığını söylerken More...

MERKEZIN TELEFONU KILITLENDI, KANDILLI GERI...
Kandili Rasathanesi Müdürü Prof. Dr. Gülay Altay, Marmara Bölgesi için normalin dışında bir yer hareketi More...

AK PARTI'DEN MEDYADAKI HABERLERE YALANLAMA...
AK Parti Tanıtım ve Medya Başkanlığı, basında AK Parti Genel Başkanı ve Başbakan Recep Tayyip More...

GÜL'DEN TSK PERSONEL YASASI'NA ONAY...
Yasa, yayımlanmak üzere Başbakanlığa gönderildi. More...

PROF. ALTAY'IN BÜYÜK DEPREM UYARISI...
Kandilli Rasathanesi Müdürü Gülay Altay, deprem çalışmaları ile ilgili soruları yanıtlıyor. Altay, 'Uzmanlara göre büyük More...

LÜBNAN'DA GÖSTERI ÇATIŞMAYA DÖNÜŞTÜ...
Çatışmalarda 1 kişi öldü, 5 kişi yaralandı. More...

NUMARA TAŞIMA 13 KASIM'DA...
Operatör değiştirilmesi halinde mevcut mobil telefon numarasının korunması olarak adlandırılan 'numara taşınabilirliği'nin 13 Kasım itibariyle More...

Previous 1 2 3 4 Next

RECENT AND PAST EVENTS

Recent Events | Past Events

YILDIZ'DA GITARLI GÜNLER...... TRACKINGS (5)
POLISTEN KAÇARKEN DEREYE UÇTULAR... TRACKINGS (5)
ÜRDÜN'ÜN ILK BIRINCISI HIRVONEN... TRACKINGS (8)

RECOMMENDED NEWS FOR HOCALAN

⊟ FORMULA 1
İSTANBUL PARK'TA HAFTA SONU -5.67678
F1'DE İSTANBUL PROGRAMI -5.87213
F1'DE KALPLER İSTANBUL'DA ATACAK -5.98946
TRULLI: 'HALA F1'DEYIM, ÇÜNKÜ -5.99238
GÜLSIN ONAY NEW YORK'TA -6.08299
ERDOĞAN: FORMULA 1 AYRI -6.09907
İSTANBUL'DAKI OTELLERIN ÇOĞU DOLDU -6.17384
KOVALAINEN'E IZIN ÇIKTI -6.26188
RAIKKONEN: 'BU YIL YENI -6.35274
F1'DE GÜVENLIK SIKI TUTULACAK -6.44025

PERSONAL NEWS WALL

DEV KALAMARIN GÖZLERI TABAK BÜYÜKLÜĞÜNDE... (DELETE)
Sender: cihan kaynak
Sending Time: 2008-05-08 21:14:13
Comment: kaç porsiyon çıkar dersin
İSTANBUL'DA FORMULA 1 HEYECANI... (DELETE)
Sender: Fazli Can
Sending Time: 2008-05-08 21:06:14
Comment: IR grubu Formula yarışlarında...
F1'DE İSTANBUL PROGRAMI... (DELETE)
Sender: Fazli Can
Sending Time: 2008-05-08 21:05:20

MOST POPULAR

İSPANYA'DA FERRARI ŞOV...
Ferrari, İspanya Grand Prix'inde de ilk iki sırayı kaptı.More...

FACEBOOK ÇALINTI MI YOKSA BU...
İnternetin popüler sosyal ağ ve arkadaşlık sitesi Facebook'un çalıntı olduğu iddia ediliyor.More...

BILIŞIM SEKTÖRÜNÜN BITMEK BILMEYEN SORUNLARI...
Kocaeli Üniversitesi Elektronik ve Haberleşme Mühendisliği Bölüm Başkanı Prof. Dr. Hasan Dinçer, 'Devletin, kurumlarını bilişiminMore...

EKONOMI
CEP TELEFONUNDA OPERATÖR
ABD'DE IŞSIZLIK MAAŞI
ENFLASYON BEKLENTISI YÜZDE
EROĞLU: GAP'TA DEVRIM
AVRUPA FAIZ ORANLARINI

POLITIKA
'YENI OLUŞUM'A ISMININ
BAKAN ÇELIK: BIZDE
23 NISAN'DA MECLIS'I
CHP, ERDOĞAN HAKKINDA
İSTANBUL'DAKI YEMEK IÇIN

TURKIYE
DTP'YE EK IDDIANAME
GÜL'DEN TSK PERSONEL
KPSS BAŞVURUSUNDA YARIN
KARA HARP OKULU'NA
GENELKURMAY MEMUR ALACAK

DUNYA
ABHAZYA KRIZI TIRMANIYOR
LÜBNAN'DA GÖSTERI ÇATIŞMAYA
SOMALI'DE ÇATIŞMA 23
AFGANISTAN'DA 6 MILITAN
RUSYA'NIN YENI BAŞBAKANI

SPOR
ZICO: 'GALATASARAY'A ELENDIRILDIK'
ZICO TEŞHISI KOYDU
TRULLI: 'HALA F1'DEYIM,
HOLOSKO: 'İNGILTERE VEYA
BEŞIKTAŞ'TAN BEBBE YALANLAMASI

KULTURSANAT
ULUSLARARASI İSTANBUL ŞIIR
RUMEN SOPRANO AYVALIK'TA
KADIFE SES YOLA
EFSANE ROCKCI KNOPFLER
ÜNLÜ EDEBIYATÇI RIFAT

SAGLIK
KONGO HASTALIĞINDAN BIR
İLTIHAPLI ROMATIZMAL HASTALIKLAR
EÜ'DE GERIATRI BILIM
SAKARYA NEHRI'NDE SU
İZMIR'DE HASTA ÇOCUKLARA

BILIMTEKNOLOJI
TÜRKIYE'NIN 'FAMELAB' BILIM
ASIMO EKSTRAYA ÇIKACAK
ASIMO, ŞIMDI DE
İBADETTE TEKNOLOJI DÖNEMI
TROPIK BÖCEKLER TEHDIT

YAZARLAR
FATİH URAZ -
FİKRET ERTAN -
İBRAHİM ÖZTÜRK -
AHMET KURUCAN -
BEŞİR AYVAZOĞLU -

Figure 3-2: Bilkent News Portal main page.

3.2 Description of Processes

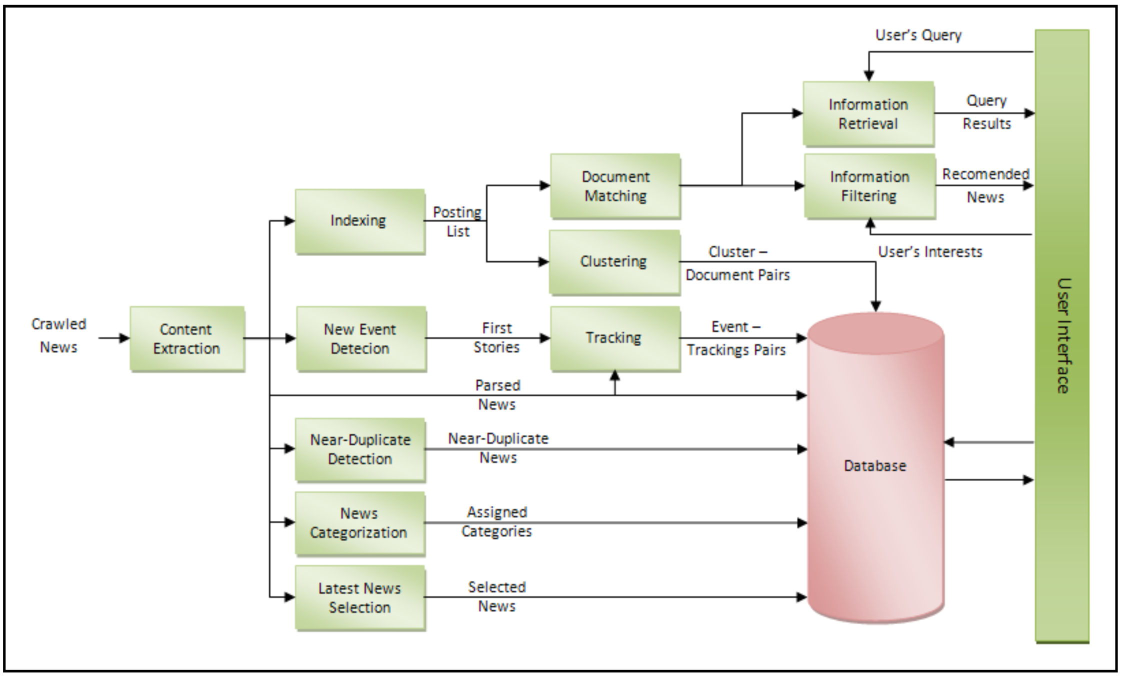


Figure 3-3: Process flow diagram of the components

3.2.1 Content Extraction

The content extraction process performs crawling and parsing. The structure of the web pages is volatile, since news sources frequently change their web page designs or link addresses. Writing resource specific parsers requires frequent updates in coding. Our automated parser resolves this problem by using an application that pays attention to the text densities of the web pages.

The content extraction application is developed after having difficulties in the crawling and parsing phases. It parses the RSS feeds of the predefined web news sources and downloads the html pages containing the news. After crawling, the program parses the raw data and determines the word chunks along with text and tag densities. These chunks are overlapped units, the size of the chunks and overlap ratio are defined as system parameters and they may be changed during execution if necessary. Each chunk's text and tag density are used to determine the text beginnings and ends.

Previous technique used at the initial phases of the portal development was news source specific. For each source we had to write different parsers, and any change in the websites had to be reflected to the parsers, because badly parsed documents significantly affect the performance of each component in the system. The source-dependent parser provided a temporary solution for the system has been changed with the new technique that uses the text densities. Our observations and experimental results reported in [KOC2009] show that this technique solves the source dependency problem.

3.2.2 Indexing

The system keeps three different indexes: index for retrieval, index for filtering, and index for clustering. The retrieval index (full index) keeps the index of all documents that have been parsed until that time, the filtering index (can be seen as a daily index) keeps the index of the last 24 hours, and finally the clustering index (short index) keeps the index of all documents for only topic and description fields. We observe that the topic and descriptions of the documents are a better descriptor for the documents when they are clustered. It is observed that this approach in clustering prevents the distortions caused by noise words.

The indexes are generated by the Lemur Toolkit [LEM2009]. Although Lemur is a powerful kit, it does not support a stemmer which is suitable for Turkish language. By using the experience from Turkish information retrieval experiments presented in Chapter 6, we expanded the Lemur Toolkit according to our needs.

3.2.3 New Event Detection & Tracking

The new event detection algorithm detects the first stories (seed news) of new events among the parsed news and prepares appropriate input for the tracking algorithm. At the next step tracking algorithm runs and finds the tracking news for the previously defined first stories [BAG2009, KAR2009].

The web component lists these events and their trackings under recent and past event tab menus (see Figure 3.2). Recent Events are the events which are detected in last update time (currently done every two hours). The past events contain all events which are alive for a period of time. Our system allows multiple seed for tracking news, so when a user opens any tracking news for reading, he can also view all the events which tracks the current news.

3.2.4 Information Retrieval (IR)

Information retrieval component enables users to search the entire collection. The IR component is based on the experimental results of our research [CAN2008a], also partly presented in Chapter 6. It uses a stopword list to eliminate the common words which do not discriminate documents from each other. After this, the first five characters of the words are used as a stem of the document and query words. This method is very simple to implement and provides comparable effectiveness results with complex stemming methods. We extended the Lemur Toolkit, and wrote mapping application to reflect the results to the web interface.

3.2.5 Information Filtering (IF)

Information filtering enables registered users follow the daily news of their interests. The filtering approach is based on supervised training, since supervised training gives better results than unsupervised training. Users define their categories (profiles) to track their interests. The system offers three different ways to define the categories.

- In the first one, user adds at least one news to the category. The filtering component extracts terms from the documents added to the category according to tf-idf value and uses these terms as a query on the news of last 24 hour.

- The second option works similar to information retrieval process. User creates a category and defines keywords for this category. The system uses these keywords to filter the latest news. This approach does not use an automated term extraction algorithm, so possible mistakes that are caused by term extraction algorithm are prevented.

- The third approach uses tracking algorithm of new event detection and tracking component. After a category is created user chooses a document. Our algorithm tracks this document within the latest news.

Both IR and IF use the same stemming and query-document matching techniques. Most relevant ten documents are displayed as the filtered content for the user defined categories.

3.2.6 News Categorization (NC)

In general, resources provide news category information. Although it is a precise tagging, the same category may have different tags and may have a different name in different sources. Our system solves this problem by mapping source categories to group of categories defined by us. News categorization in our system can be named as mapping of source categories. Finally, the news articles are listed in these categories on the web interface.

3.2.7 Retrospective Incremental News Clustering (RINC)

Retrospective incremental news clustering provides the functionality of browsing among the new and old news by separating the news into groups. To perform this, cover coefficient-based incremental clustering methodology (C2ICM) has been used [CAN1993]. The algorithm is defined for dynamic environments and produces partitions that do not overlap with each other. The partitioned document collection which is generated by using this algorithm also gives effective results in information retrieval. These results are already shown in previous works [CAN1993, CAN2004, ALS2008].

In the context of a news portal, clusters are used to browse the news. In the user interface user can reach the documents in the same cluster by following the link "view cluster." The observations show that the clusters may have too many documents and this situation may be overwhelming for the user. To prevent this problem, the news articles are given in the order of their similarity to the current document selected by the user to come to that cluster.

3.2.8 System Personalization (SP)

It is verified that user-adapted systems are more effective and usable than non-adaptive system in several areas. The basic principal of these systems is providing a service by accepting each user as unique with different needs. In the case of news portal, presenting news according to users' interest is the fundamental idea behind personalization. Information filtering component that we presented above is also a part of system personalization.

With the concept of Web 2.0, new trends have been announced such as collaboration, information sharing and social networking. The main idea was using the web as a platform that people can share information and communicate with each other. Thus, this concept shifts the expectations to a higher level.

While designing the news portal we give importance to these expectations and develop the system accordingly. The system allows user to create their filtering profiles, and at the same time add other users as their friends. By this way, the users become connected to the users in their friend list. If they wish, they can send news stories (optionally with their comments on these stories) to selected users in their friend list. By this way suggested news can be accessed without requiring extra time by friends.

The main idea was developing a user-oriented news site. In this way, user would manage and access the news easily. “Favorites” is one of the applications that serve for this purpose. By adding news to the favorites, user can store his “favorite news” for later access.

3.2.9 Latest News Selection (LNS)

The incoming news from different news sources are presented to the user according to their relative relevance to today's news agenda. This is presented to the user automatically by using C3M approach which groups news according to their relative relevance to today's news [CAN1990, CAN1993]. According to this approach, if a news article is different from the today's previous news stories its coupling (overlap) is low. The news articles with high coupling with the current news articles are presented on the main page of the news portal (see Figure 3.3).

1. There is a window that has 500 latest news.
2. Coupling coefficient of the incoming news is calculated according to this window.
3. Then, the news are sorted in descending order by their coupling coefficient value and presented to the users.

Figure 3-4: Steps of latest news selection.

3.2.10 Near-duplicate Detection (NDD)

Near-duplicate elimination is inevitable if you are working with multiple sources, because the same news arrives frequently from different sources that uses the same news agencies. In this case duplicates of the news decreases the result of the services such as retrieval, novelty detection and tracking.

The news portal provides a unique near-duplicate elimination algorithm to detect the duplicate news. The algorithm generates signatures of each crawled news. The news with the same signature are marked as near-duplicate and stored, then clustered in the system. The first arrived document is accepted as the cluster head of the duplicate clusters, and primarily cluster head is presented to the user while generating the results by the services. However, if the user whishes, it is possible to see the duplicates to examine the differences of these similar news in the same cluster. It is obvious that this component increases the readability of the results. The details of the algorithm are provided in [UYA2009].

3.2.11 Multi-Document Summarization (MDS)

Multi document summarization (MDS) is one of the interesting research topics of NLP and IR [ERC2009]. In general, MDS is designed to summarize the result of retrieval and tracking components. It can provide the users a general view about the results without browsing them. In most cases of tracking results, MDS can give a perfect outline to see how the event evolved. But in some cases of tracking results and retrieval results, the documents could be dissimilar in terms of meaning, although it contains similar terms. The algorithm has to be empowered by semantic process to solve this problem. The prototype implementation is still developing to increase the performance to solve such semantic problems.

Chapter 4

System Data and File Structures

Data and file structures of information retrieval systems have a significant impact in efficiency. Well organized data decreases response time by reducing the unnecessary disk access. The news portal is developed by considering these issues to serve large amount of people simultaneously. In this chapter, data and file structures of the system are explained in detail. Firstly, incremental indexing is described, then RDBMS tables, and finally XML structures are presented.

4.1 Incremental Indexing

Indexing is the major part of the system that affects the performance significantly. In this project we use the Lemur Toolkit, information modeling and retrieval library [LEM2009]. This library is developed with the collaboration of research groups in Carnegie Mellon University and University of Massachusetts and provided under the open-source license. To reflect the benefits of this library to the news portal, it is expended according to the results of our information retrieval experiments are presented in Chapter 6.

Incremental indexing is inevitable in real time information retrieval systems. Indri indexing module of the Lemur Toolkit is an appropriate choice for our system, since our system receives new documents frequently and requires regular updates.

Our indexing component powered by indri indexing can index more than terabytes of textual data incrementally. Its flexible parsing opportunity allows us to parse plain text, HTML, XML, and PDF documents which use UTF-8 encoding. Only the newly parsed documents are used to build the index incrementally. Building index on memory and writing it to the disk by different threads keeps disk I/O at the minimum. Additionally, we use tf-idf model of Lemur Toolkit as the matching function of IR component.

4.2 RDBMS Tables

News Portal uses background services working behind the interface, and these services process the data in retrospective manner. For this purpose, the data must be stored in files or in the tables of a database. For efficiency and manageability, using a database is advantageous over the file system. In this part of the chapter, database tables will be provided and described in detail.

Portal Tables

Users: User registered to the portal are added to the users tables. In addition to the username and the password, some addition of the information is requested such as name, lastname and e-mail. Type attribute of new user is set as ordinary “user” as default, the admin users can set the type as “admin” to allow the user to see some statistical information about the system.

Queries: This table stores queries that are searched by the users for statistical purpose. However, in the near future the queries will be used to characterize the user groups. So, the system will be able to provide content with respect to users' general interest.

Main_News: Meta data of the news such as source, genre, language, title, description, and the references to the content are kept in this table. The user interface programs can display the news without accessing the whole content. The minimum access to the operating systems file structure provides significant improvement in terms of efficiency.

VisitorLogs: This table stores the logs of each user whether it is signed in or not. It is used for statistical purpose, but in the released version, the system does not keep this information to respect the privacy.

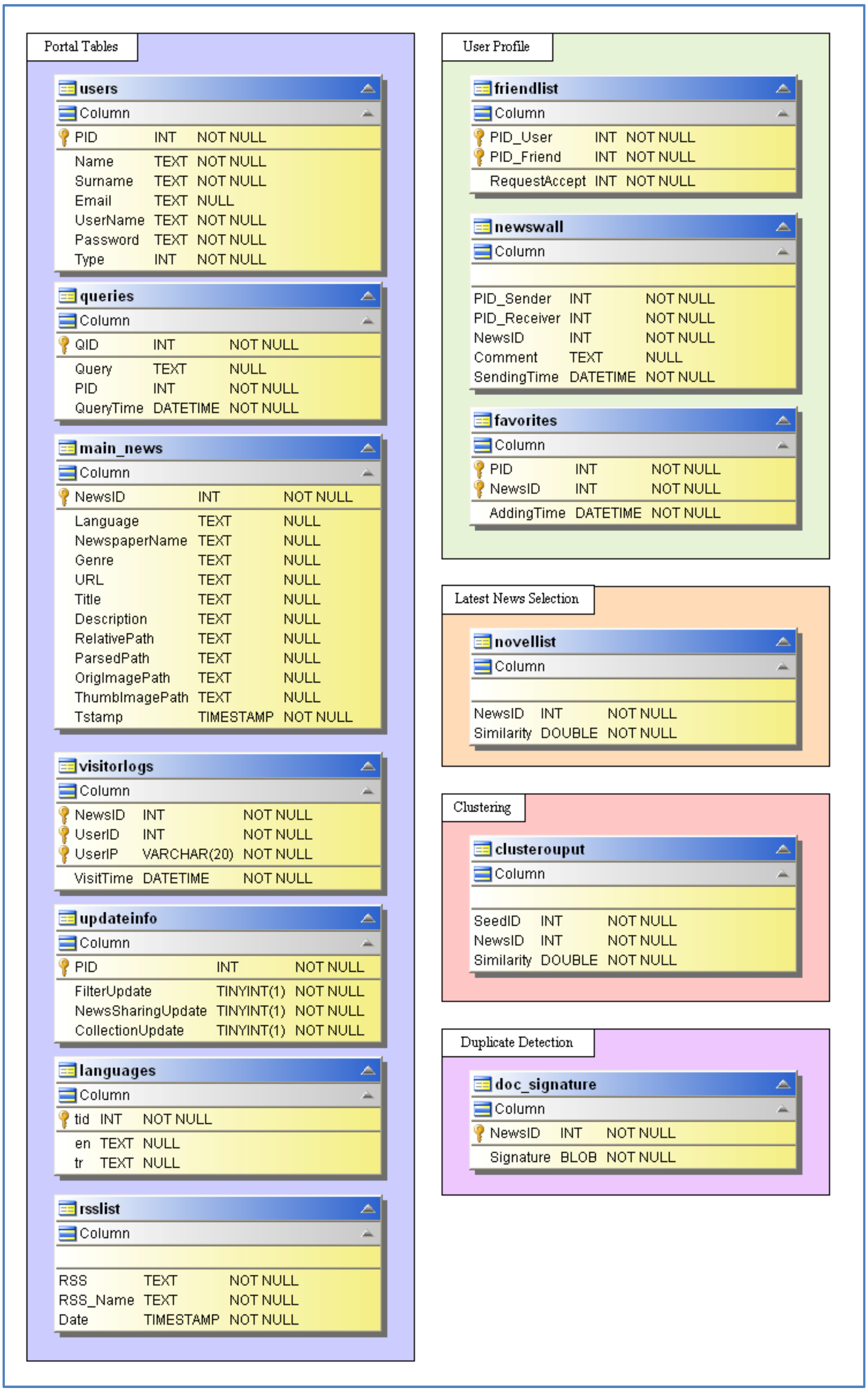
UpdateInfo: Our system uses separate caching mechanism for each user. This decreases the loading time of the personalized pages, and a user that visited portal before does not wait for the load of unchanged content. To provide updated information, the cached content is dismissed when one of the following conditions occurs:

- Change in filtering content,
- Receiving a shared news from another user,
- Update of the general collection with latest news.

UpdateInfo table stores the changes of these conditions for each user to decrease the loading time.

Languages: The user interface of our portal allows using different languages. This table keeps the definition of each label in different languages.

RSSList: The crawling module collects from the RSS of news providers. This table keeps complete list of our RSS sources.



Portal Tables

users

Column		
🔑 PID	INT	NOT NULL
Name	TEXT	NOT NULL
Surname	TEXT	NOT NULL
Email	TEXT	NULL
UserName	TEXT	NOT NULL
Password	TEXT	NOT NULL
Type	INT	NOT NULL

queries

Column		
🔑 QID	INT	NOT NULL
Query	TEXT	NULL
PID	INT	NOT NULL
QueryTime	DATETIME	NOT NULL

main_news

Column		
🔑 NewsID	INT	NOT NULL
Language	TEXT	NULL
NewspaperName	TEXT	NULL
Genre	TEXT	NULL
URL	TEXT	NULL
Title	TEXT	NULL
Description	TEXT	NULL
RelativePath	TEXT	NULL
ParsedPath	TEXT	NULL
OrigImagePath	TEXT	NULL
ThumbImagePath	TEXT	NULL
Tstamp	TIMESTAMP	NOT NULL

visitorlogs

Column		
🔑 NewsID	INT	NOT NULL
🔑 UserID	INT	NOT NULL
🔑 UserIP	VARCHAR(20)	NOT NULL
VisitTime	DATETIME	NOT NULL

updateinfo

Column		
🔑 PID	INT	NOT NULL
FilterUpdate	TINYINT(1)	NOT NULL
NewsSharingUpdate	TINYINT(1)	NOT NULL
CollectionUpdate	TINYINT(1)	NOT NULL

languages

Column		
🔑 tid	INT	NOT NULL
en	TEXT	NULL
tr	TEXT	NULL

rsslist

Column		
RSS	TEXT	NOT NULL
RSS_Name	TEXT	NOT NULL
Date	TIMESTAMP	NOT NULL

User Profile

friendlist

Column		
🔑 PID_User	INT	NOT NULL
🔑 PID_Friend	INT	NOT NULL
RequestAccept	INT	NOT NULL

newswall

Column		
PID_Sender	INT	NOT NULL
PID_Receiver	INT	NOT NULL
NewsID	INT	NOT NULL
Comment	TEXT	NULL
SendingTime	DATETIME	NOT NULL

favorites

Column		
🔑 PID	INT	NOT NULL
🔑 NewsID	INT	NOT NULL
AddingTime	DATETIME	NOT NULL

Latest News Selection

novellist

Column		
NewsID	INT	NOT NULL
Similarity	DOUBLE	NOT NULL

Clustering

clusterouput

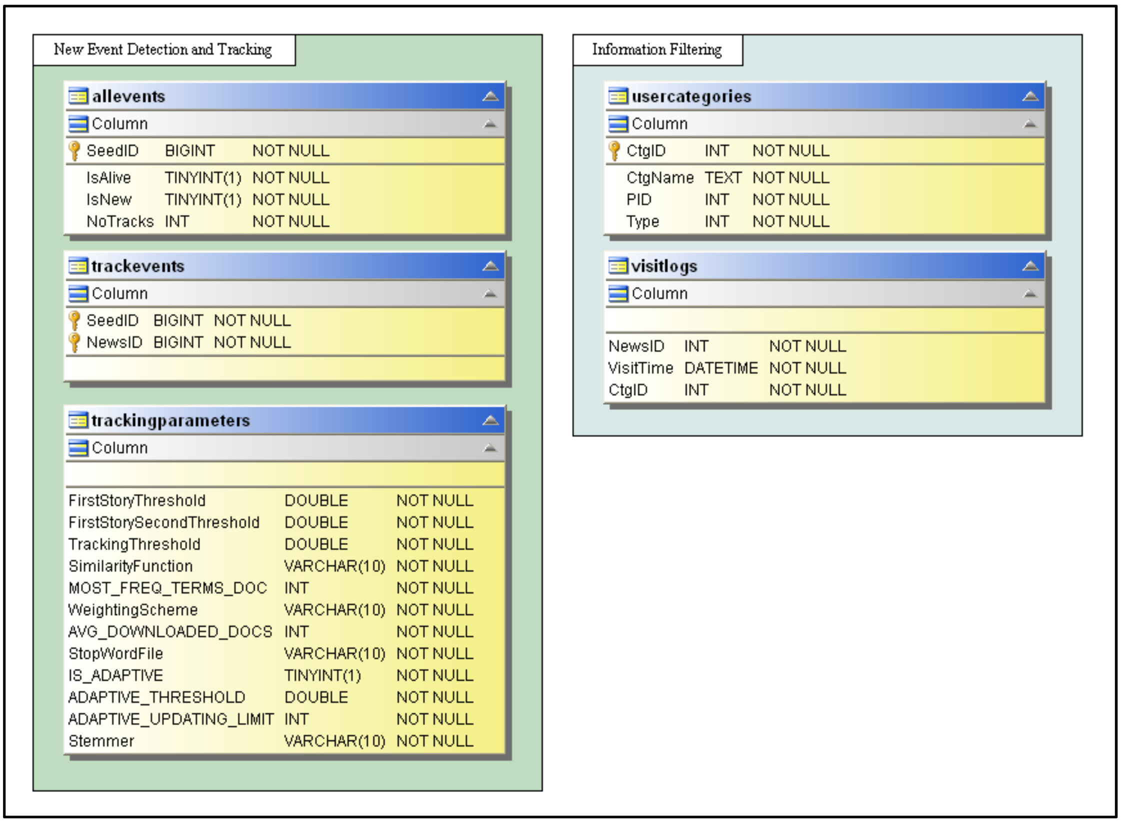
Column		
SeedID	INT	NOT NULL
NewsID	INT	NOT NULL
Similarity	DOUBLE	NOT NULL

Duplicate Detection

doc_signature

Column		
🔑 NewsID	INT	NOT NULL
Signature	BLOB	NOT NULL

Figure 4-1: Database tables.



New Event Detection and Tracking

allevents

Column		
SeedID	BIGINT	NOT NULL
IsAlive	TINYINT(1)	NOT NULL
IsNew	TINYINT(1)	NOT NULL
NoTracks	INT	NOT NULL

trackevents

Column		
SeedID	BIGINT	NOT NULL
NewsID	BIGINT	NOT NULL

trackingparameters

Column		
FirstStoryThreshold	DOUBLE	NOT NULL
FirstStorySecondThreshold	DOUBLE	NOT NULL
TrackingThreshold	DOUBLE	NOT NULL
SimilarityFunction	VARCHAR(10)	NOT NULL
MOST_FREQ_TERMS_DOC	INT	NOT NULL
WeightingScheme	VARCHAR(10)	NOT NULL
AVG_DOWNLOADED_DOCS	INT	NOT NULL
StopWordFile	VARCHAR(10)	NOT NULL
IS_ADAPTIVE	TINYINT(1)	NOT NULL
ADAPTIVE_THRESHOLD	DOUBLE	NOT NULL
ADAPTIVE_UPDATING_LIMIT	INT	NOT NULL
Stemmer	VARCHAR(10)	NOT NULL

Information Filtering

usercategories

Column		
CtgID	INT	NOT NULL
CtgName	TEXT	NOT NULL
PID	INT	NOT NULL
Type	INT	NOT NULL

visitlogs

Column		
NewsID	INT	NOT NULL
VisitTime	DATETIME	NOT NULL
CtgID	INT	NOT NULL

Figure 4.1: (Cont.) Database tables.

User Profile

The tables in this group stores user dependent information such as friend list, news wall, and favorite news.

FriendList: The table stores the user pairs who are added as friends after request and confirmation process. The users who are recorded as friends can share the news with each other.

NewsWall: The shared news for each user are recorded in this table. When a user logs into system, the news that are shared by its friends are displayed in the users personal page.

FavoriteNews: Users can save their favorite news to reach them later.

Latest News Selection

NovelList: Latest News Selection component selects the news which will be displayed in the main page. The results are written to this table with their similarity values. The user interface displays the news which have higher similarities than threshold value.

Clustering

ClusterOutput: Our system clusters the documents to make the browsing process easy. The output of the clustering component is written to this table. The user can browse news clusters to see the similar documents.

Duplicate Detection

DocSignature: Our duplicate detection component stores the signature of each news document in this table. The news which have the same signature are accepted as duplicate in the system.

New Event Detection and Tracking

New event detection and tracking component uses database tables to store its data. There are three tables in this group, AllEvents, TrackEvents, and TrackingParameters which keep the events, trackings, and settings respectively.

AllEvents: New event detection algorithm uses this table to write the seed document of the events. Additionally, this table keeps information about events whether they are new or old and alive or not. Only the events which are alive are displayed on the user interface.

TrackEvents: Tracking algorithm reads the events from AllEvents table and writes their trackings as SeedID-NewsID pairs to this table. Additionally, tracking algorithm decides whether the events are new or old and alive or not by modifying AllEvents table.

TrackingParameters: Tracking algorithm uses this table to set its parameters.

Information Filtering

Information filtering component uses two tables which are UserCategories and VisitLogs to store user interests and users' positive votes.

UserCategories: Users can create categories for their interests, and these categories are feed by information filtering component when the latest news arrives. The categories of the users are stored in this table. Type attribute defines the type of filtering operation as we mentioned in the related chapter.

VisitLogs: If the user has decided to filter the news with type 1 filtering, at least one document has to be added to this table for the selected category. Consequently, filtering component extracts the terms that will be used in filtering process. So, this table keeps the positive voted news for categories.

4.3 XML Data Structures

In our news portal about two thousands of articles are processed daily, and this number is increasing day by day with the addition of new resources. The documents that are processed are stored in the IR collection according to the TREC standard to provide a large test document collection to the Turkish researchers.

The system uses two types of simple XML files to store news and queries. The Figure 4.3 shows a sample document from the system. `<DOC>` tag identifies the beginning of the document content. <DOCNO> element keeps the identification number of the news given by the parser. <SOURCE> element specifies the source of the news, and <URL> element gives the original location of the document in the source. <DATE> and <TIME> elements keep the publication time of the news. If the news does not have this information, the parser sets these values as crawling date and time. <AUTHOR/> element is used if the author of the news is provided by the source. <HEADLINE> and <TEXT> keeps the content of the news.

```
<DOC>
      <DOCNO> 697516 </DOCNO>
      <SOURCE> TRT </SOURCE>
      <URL> http://www.trt.net.tr/wwwtrt/hdevam.aspx?hid=199971&k=0
      </URL>
      <DATE> 2008/04/17 </DATE>
      <TIME> 00:11 </TIME>
      <AUTHOR/>
      <HEADLINE>
            Para Politikası Kurulu Toplanacak
      </HEADLINE>
      <TEXT>
            Toplantı, saat 13.00-17.00 arasında iki aşamalı olarak
            yapılacak.
            Merkez Bankası Para Politikası Kurulu, bu yılın dördüncü
            toplantısını bugün (17.4.2008) yapacak.
            ...
      </TEXT>
</DOC>
```

Figure 4-2: Sample news document from Bilkent News Portal.

Our information retrieval component uses indri indexing of the Lemur Toolkit and it has its own query language definition. So, our system creates input query files for lemur toolkit. In this file, <parameters> tag defines the beginning of the query and <query> tag keeps the content of the query. The sample query in Figure 4.4 includes four keywords, and three of them are typed as phrase. So, retrieval component searches by combining the phrase “Para Politikası Kurulu” with the keyword “toplanacak.”

```
<parameters>
      <query>
            #combine(#1(Para Politikası Kurulu) toplanacak)
      </query>
</parameters>
```

Figure 4-3: Sample indri query file from Bilkent News Portal.

Chapter 5

Experimental Foundations I: BilCol2005 Test Collection

In the scope of this work, new event detection and tracking system is developed for the first time for Turkish resources. Although, this is a brand new service for Turkish news portals, it is also very rare for other languages. It requires comprehensive research and development effort for providing such system to the users. The performance of the algorithms tested to improve effectiveness. Bilkent News Portal based on a series of experiments which measure the services in terms of effectiveness and efficiency. In this chapter, the tools developed to prepare a test collection, ETracker system and BilCol2005 collection will be mentioned in detail.

5.1 Test Collection Creation for NED

In TDT, a test collection contains several news articles in temporal order and first stories and tracking news of a set of events that are identified by human annotators. In this section we describe the contents of the test collection in terms of news resources, topic profiles, annotation process, and finally in terms of the characteristics of the annotated events.

News Sources

We use five different Turkish news web sources for the construction of the news story collection.

- CNN Türk [CNN2009],
- Haber 7 [HAB2009],
- *Milliyet Gazetesi* [MIL2009],
- TRT [TRT2009],
- *Zaman Gazetesi* [ZAM2009].

It can be claimed that these sources have different worldviews. CNN Türk has an American style approach to news delivery; *Milliyet* is a high circulation newspaper in Turkey and by some people considered as a progressive newspaper; TRT (Turkish Radio and Television) is a state organization, and reflects the state views; *Zaman* is a conservative newspaper; and finally Haber 7 provides variety.

From these sources, we download all articles of the year 2005 that have a timestamp in terms of day, hour, and minute. The downloading is completed in the second half of 2006 by using the archives of these news sources. Duplicate or near-duplicate documents of this initial collection are eliminated by using a simple method: stories with the same timestamp coming from the same source and with identical initial three words are assumed as duplicate. We eliminated about 16,000 stories by this way. Such documents were due to interrupted crawling or multiple identical postings of the news providers.

Table 5.1: TDT5 corpus content [TDT2004].

Language	*No. of News Sources*	*No. of News Stories*
Arabic	4	72.91
Mandarin	4	56.486
English	7	278.109

Table 5.2: Information about distribution of stories among news sources in BilCol2005.

News Source	No. of News Stories	Percent of All Stories	Download Amount (MB)	Net Amount (MB)	Avg. No. of Words per Document
CNN Türk	23,644	11.3	1,008.3	66.8	271
Haber 7	51,908	24.8	3,629.5	107.9	238
Milliyet Gazetesi	72,233	34.5	508.3	122.5	218
TRT	18,990	9.1	937.9	18.3	121
Zaman Gazetesi	42,530	20.3	45.3	33.7	97
All together	209,305	100.0	6,129.3	349.2	196*

* Different from the weighted sum of the average word lengths due to rounding error.

Having different news sources provides variety, different viewpoints to news-consumers during news tracking, and a significant volume in terms of the number of stories. The size of our test collection is comparable to those of the TDT research initiative. Summary information regarding TDT5 corpus and more detailed information about our corpus, BilCol2005 (Bilkent TDT Collection for the year 2005), are provided in Tables 5.1 and 5.2, respectively. Our collection contains 209,305 documents and *Milliyet* provides the maximum number of news articles (72,233 stories or 34.5% of all of the stories) among the five sources. For content extraction, we download the entire news pages (6,129.3MB) and extract the parts that correspond to news texts (349.2MB). On the average, there are about 573 stories per day or about 24 stories per hour, the average time distance between two stories is about 2.5 minutes. Figure 5.1 shows the number of stories observed on each day of the year 2005. The distribution of news among the days is nonuniform: during weekends, we observe lesser number of stories. In a similar way, during the summer months (year day numbers approximately from 180 to 240, i.e., in July and August) the number of stories per day is smaller. Table 5.2

shows that the textual lengths of stories show variation among the news sources, on the average each story contains 196 words (tokens).

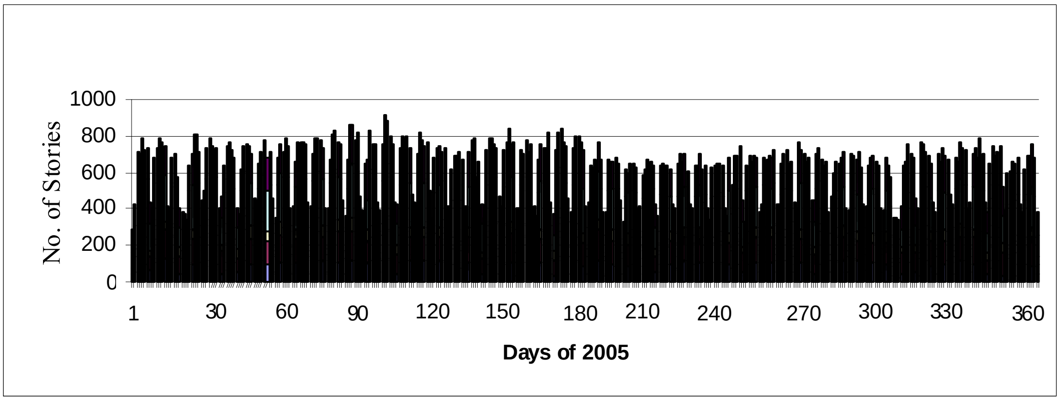
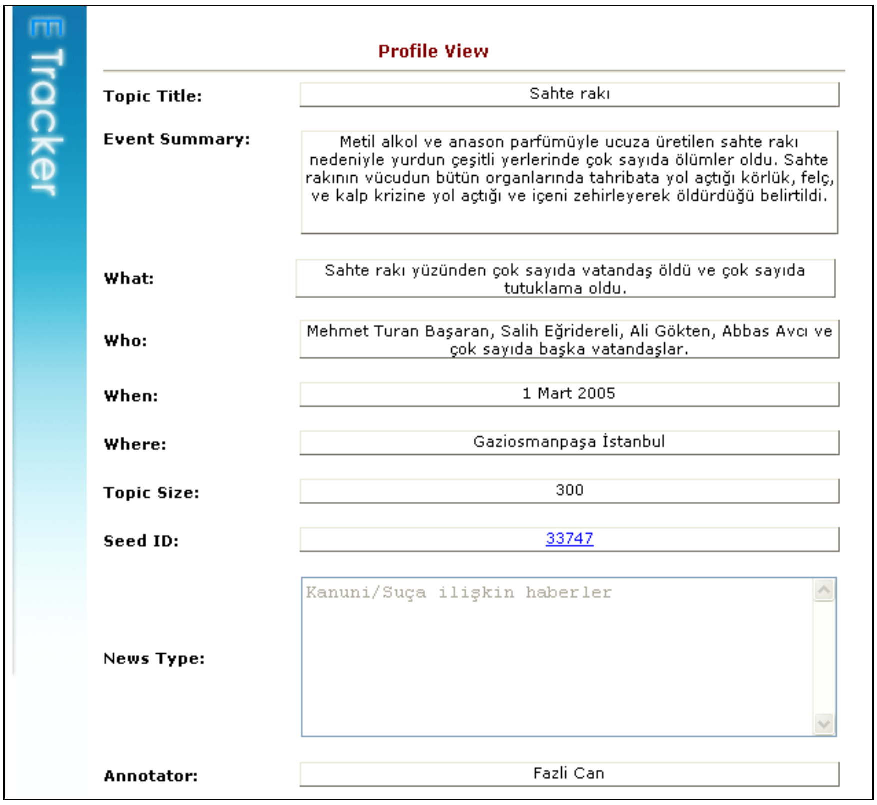


Figure 5-1: Distribution of news stories in 2005.

Topic Profiles

In order to construct a TDT test collection the initial task for each event is preparation of an event profile. An event profile has the following items (an example profile is shown in Figure 5.2).

- Topic Title: A brief phrase which is easy to recall and reminds the topic,
- Event Summary: A summary of the seminal event with 1 or 2 sentences,
- What: What happened during the seminal event,
- Who: Who was involved (people, organization etc.) during the seminal event,
- When: When the seminal event occurred,
- Where: Where the seminal event happened,
- Topic Size: Predicted number of stories for the event,
- Seed: The first story about the seminal event (the document number of the news in the collection),
- Event Topic Type: The news type (the annotators are allowed to mark more than one topic type).



Profile View

Topic Title:	Sahte rakı
Event Summary:	Metil alkol ve anason parfümüyle ucuza üretilen sahte rakı nedeniyle yurdun çeşitli yerlerinde çok sayıda ölümler oldu. Sahte rakının vücudun bütün organlarında tahribata yol açtığı körlük, felç, ve kalp krizine yol açtığı ve içeni zehirleyerek öldürdüğü belirtildi.
What:	Sahte rakı yüzünden çok sayıda vatandaş öldü ve çok sayıda tutuklama oldu.
Who:	Mehmet Turan Başaran, Salih Eğridereli, Ali Gökten, Abbas Avcı ve çok sayıda başka vatandaşlar.
When:	1 Mart 2005
Where:	Gaziosmanpaşa İstanbul
Topic Size:	300
Seed ID:	33747
News Type:	Kanuni/Suça ilişkin haberler
Annotator:	Fazli Can

Figure 5-2: profile: "Sahte – counterfeit- rakı."

In our study, we use the news topic classification as defined by the TDT research initiave [CIE2002, TDT2004]. There are 13 topic types: 1) elections, 2) scandals/hearings, 3) legal/criminal cases, 4) natural disasters, 5) accidents, 6) acts of violence or war, 7) science and discovery news, 8) financial news, 9) new laws, 10) sports news, 11) political and diplomatic meetings, 12) celebrity/human interest news, and 13) miscalleneous news.

5.2 ETracker System and Topic Annotation

During annotation of a specific topic started with an event, annotators aim to identify the tracking stories of a seminal event initiated by its first story. By following the TDT tradition, a topic is defined as "an event or activity, along with all directly related events and activities" [CIE2002, TDT2004]. Here an activity is "a connected set of events that have a common focus or purpose, happening at a specific place and time" [TDT2004].

Our topic annotation method is search-based or search guided and inspired by the TDT research initiative [CIE2002, TDT2004]. In TDT, the initial test collections were constructed using a brute-force approach by carefully examining all the news for each profile. However, beginning with TDT-3 (2000) evaluation, this task is performed semi-automatically by using an IR system due to difficulty of manual annotations [TDT2004]. Furthermore, previous experiments done by the TDT researchers had shown that search guided annotation could produce results that are as good as brute-force manual annotation results [CIE2002].

In our case, we developed and used the topic annotation system ETracker to find the first stories of new events and their tracking news. ETracker is a web application and developed in Microsoft .NET with the C# language.

In ETracker, the news collection is accessed by using an IR system developed for Turkish. The design principles of the IR system are available in Chapter 6. For document indexing and searching, we use a tf-idf-based matching function and the first five prefix stemmer. It is shown that this "matching function-stemmer" combination provides an effective IR environment for Turkish [CAN2008a].

Annotators select their own topics, i.e., we have imposed no restrictions on their decisions; like that of TDT (2004) no effort is made to nurse equal representation of each news source or month in the final set of selected events/topics. They are allowed to see each other's profiles to prevent multiple annotation of the same event. The annotator who selects an event creates the associated event profile and performs the annotation task. The annotators are provided with example profiles and asked to experiment with the system by creating and discarding experimental profiles for learning purposes. A high majority of the annotators is trained in a work-through style tutorial presentation followed by a question and answer session. A small set of annotators are also trained remotely by e-mail.

For identifying the first story of the selected event, the associated annotator may need to perform multiple passes over the corpus with queries using the ETracker's IR system. During this process, ETracker displays the documents in chronological order rather than relevance order. Annotators first create an event profile after finding the seed (first) story by interacting with ETracker and after reading not only the first but also some tracking stories of the seed story. The correct selection of the first story is important, otherwise during the experiments, a) for that particular event the correct first story not detected by the annotator could be detected by the system as the first story, and b) incorrectly chosen first story could become a tracking news of the first story identified by the system. If both of these cases happen during the experiments, case-a will be classified as a "false alarm," and case-b will be classified as a "miss". During the experimental evaluation process, these two cases would incorrectly lower the NED performance and lead to an incorrect measurement of the true system performance. For this reason, a senior annotator makes sure for each event profile the first story has been correctly identified. Therefore, junior annotators wait for the approval of a senior annotator regarding the correctness of the first story of the chosen event. If the first story of an event is not approved, the process of selecting the first story and generating the associated event profile is repeated until the senior annotator approves the first story. After identifying the first story, four annotation steps are performed for identifying the tracking stories of an event. They are followed by a quality control performed by a senior annotator. The annotation steps are the following.

Step-1 (Search with the seed): The annotation system searches the collection for tracking news by using the seed (first) story as a query. ETracker shows the results according to their relevance to the seed document. In all steps, the listed documents have a timestamp newer than that of the seed story. The annotator decides if the results coming from ETracker is on topic or not, and labels the results as "Yes: on-topic" or

"No: off -topic." In this and the following step if the annotator is "unsure" about a story he/she can mark it in both ways, i.e., "yes" and "no" at the same time and can change it later to "yes" or "no." If a story remains like that after the completion of all annotation steps, it is defined as "off-topic." There were only a few cases like that.

Step-2 (Search with the profile information): ETracker ranks the collection documents using the profile information (words, etc. used in the profile) as a query. In all steps the links of the newly retrieved documents are shown in blue. In this and the following steps, the links of the already labeled stories are shown in red (if labeled as "no"), green (if labeled as "yes"), or orange (if labeled as "yes" and "no" at the same time). The annotator can read a story multiple numbers of times and change their labels.

Step-3 (Search with on-topic stories): ETracker uses the first three on-topic stories of step-1 and step-2 and uses them as separate queries (if they are not distinct, the following on-topic stories of each step are selected to gather six distinct stories –if possible-). The final ranking of the stories retrieved by these queries is determined by using the reciprocal rank data fusion method [NUR2006]. They are presented to the annotator for labeling in the rank order determined by the data fusion process.

Step-4 (Search with queries): In this step, ETracker uses the annotator's queries for retrieving relevant stories. At this stage, the annotator has already become an expert on the event and may search using his/her own queries. The annotators may use any number of queries in this step.

Table 5.3: Information about ETracker search steps

Step	Search with	Max. No. of Documents Ranked for Annotation	Recommended Time Limit (minutes)
1	The seed document	200	60
2	The profile information	300	45
3	On-topic documents	400	45
4	Queries (for each attempt)	200	30

The number of listed stories is limited to 200, 300, 400, and 200, respectively, for the steps 1 to 4. In order to make the annotation process more efficient and effective, there is a recommended time limit for each step as shown in Table 5.3. As we go to the later steps of annotation, the number of stories that has already been labeled increases; therefore, in general the time allotment per listed document decreases. In each step, the annotator evaluates the listed stories and makes a decision. Annotators can spend more time than the recommended time limits. These time limits are given so that annotators would not spend too much or too little time and pay enough attention. Off-topic-threshold means that the last 10 stories evaluated by the annotator are off-topic and the ratio of "the number of on-topic stories found so far" to "number of off-topic stories" is ½. If annotators cannot find any on-topic stories in the top 50 stories of the first step, they are advised to drop the event and try another one. The senior annotator inspects the whole event if the annotator selected more than one diverse category, i.e., the event coverage, descriptions are completely different. If the coverage of the event is incorrectly interpreted by the annotator, the senior annotator deletes the event. Totally, 21 events are deleted by this way.

Quality Control: A senior annotator examines 20 documents from each of the following categories: documents labeled as on-topic ("yes"), documents labeled as off-topic ("no"), and documents brought to the attention of the annotator but not labeled (so all together 60 maximum). In this process, if any of the on/off-topic is labeled incorrectly, or any document not examined is actually an on-topic document, then the junior annotator is asked to redo the annotation from the very beginning. In such cases, annotators are allowed to change the topic and consider another event and its stories.

5.3 Characteristics of Annotated Events

All annotators are experienced web users: graduate and undergraduate students, faculty members, and staff. They are not required to have an expertise on the topic that they pick. All together, there were 39 native speaker annotators. The annotated topics and topic categories with the number of topics in each category are shown in the Appendix, Tables A.1 and Table A.2, respectively. The number of stories in each topic according to its type is shown in Figure 5.3. The news distribution of two sample event profiles, “Sahte Rakı (Counterfeit Rakı)” and “Türkiye’de Mortgage (Mortgage in Turkey),” are shown in Figure 5.4. Some additional information about the collection and information about some sample profiles are shown in Table 5.4. In this table for the “Onur Air'in Avrupa’da yasaklanması” event (event no. 2) there are total of 159 tracking events and it stays active for 203 days, and on the first 100 days of this event there are total of 154 news stories about this topic. The final list is obtained after eliminating 21 completed event profiles during quality control. In most of the eliminated profiles, the “eventness” of their topic was questionable. In some events the annotators violates event coverage. In the final test collection, there are 80 events and their associated annotations. On the average there are 73 (median: 32, minimum: 5, maximum: 454) tracking stories for each topic (event). On the average annotators spent, not counting the breaks that they may have taken, 109 (median: 80, minimum: 20, maximum: 825) minutes for their annotations. The average topic life is 92 (median: 59, minimum: 1, maximum: 357) days. The distribution of topic stories among the days of 2005 is shown in Figure 6.

Table 5.4: Information about sample profiles and some averages for BilCol2005

			No. of news on first n days			
Sample event (event no.)	**Tracking News**	**Life Span (days)**	**n= 100**	**n= 50**	**n= 25**	**n= 10**
Onur Air'in Avrupa’da yasaklanması (2)	159	203	154	154	148	105
Londra metrosunda patlama (6)	454	175	440	419	376	236
400 koyun intihar etti (10)	10	8	10	10	10	10
Mortgage Türkiye'de (14)	375	357	60	41	25	13
Attilâ İlhan vefat etti (21)	40	70	40	37	36	32
Sahte rakı (48)	323	182	316	291	255	197
İlk yüz nakli (61)	14	17	14	14	14	10
Averages for all 80 events	73	92	64	54	47	36

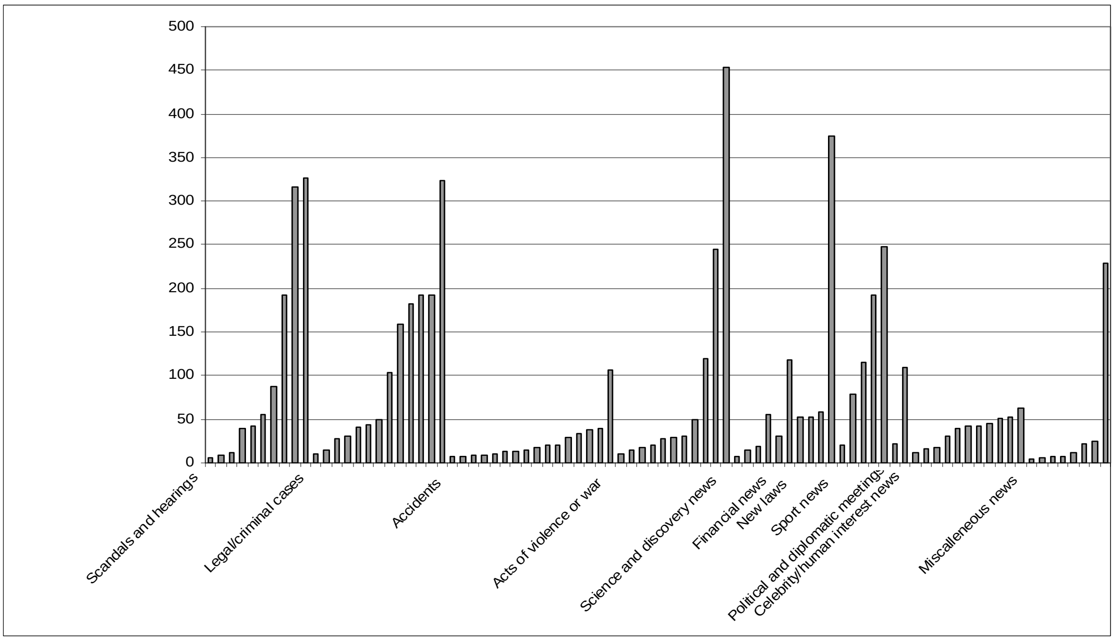


Figure 5-3: Distribution of news stories among news profiles of BilCol2005. There are total of 86 bars and each bar corresponds to an event/profile (for 6 events there are double event category assignments, they appear twice in this figure). The height of a bar indicates the number of tracking stories (shown on the y-axis) for that event.

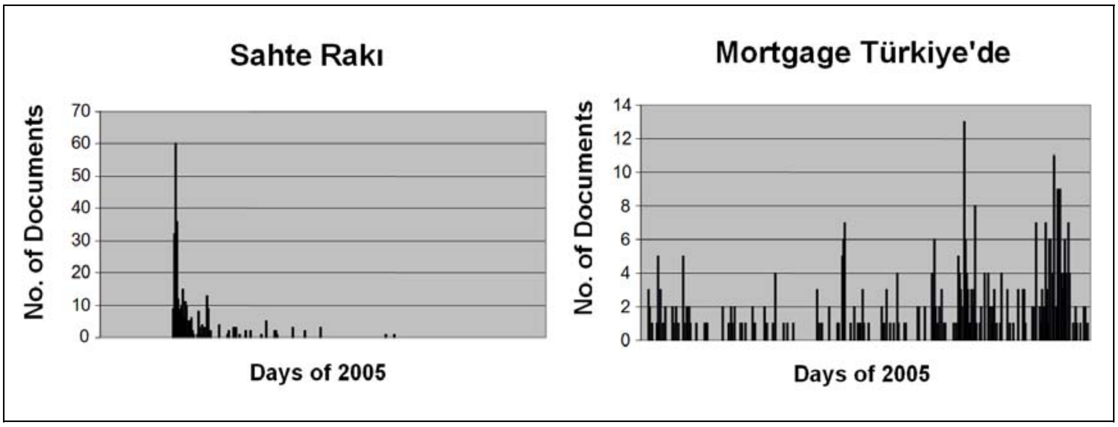


Figure 5-4: Distribution of news stories in the year 2005 for two sample events ("Counterfeit Rakı" –event no. 48- and "Mortgage in Turkey" – event no. 14 -), x-axis goes from Jan. 1 to Dec. 31, 2005.

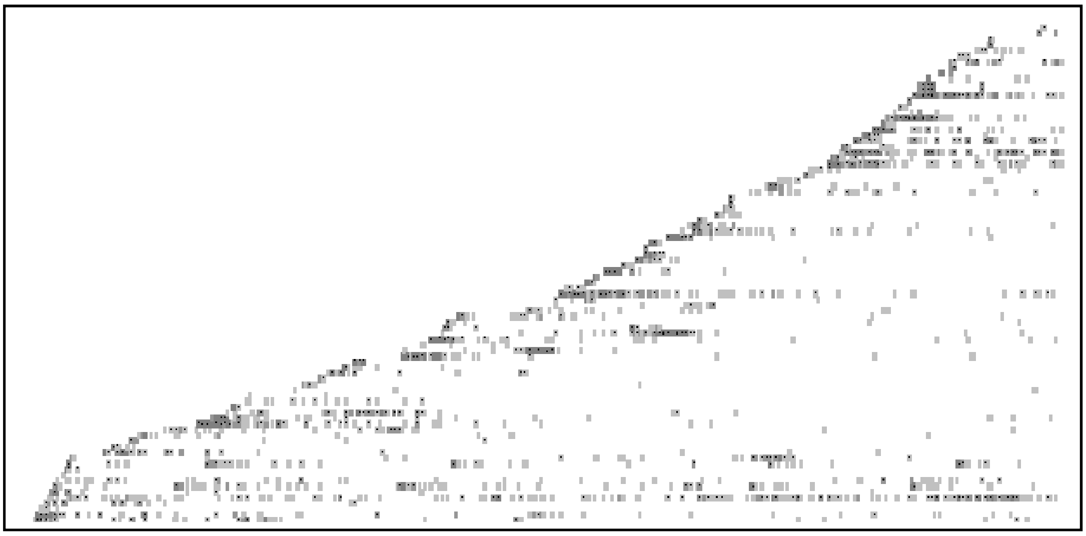


Figure 5-5: The distribution of BilCol2005 topic stories among the days of 2005. x axis goes from Jan.1 to Dec. 31, 2005. Each horizontal position represents a different event and there are 80 events. The gray level is proportional to the number stories on that day, darker gray spots indicate more stories. Days with 10 or more stories are shown with the same gray color.

Chapter 6

Experimental Foundations II: Information Retrieval Parameters

In this chapter, we investigate information retrieval (IR) on Turkish texts using a large-scale IR test collection that contains 408,305 documents and 72 ad hoc queries [CAN2008a]. We examine the effects of stemming options on retrieval performance. We show that a simple word truncation approach, and an elaborate lemmatizer-based stemmer provide similar retrieval effectiveness in Turkish IR. We investigate the effects of a range of search conditions on the retrieval performance; these include scalability issues, query and document length effects, and the use of stopword list in indexing. These experiments shed light on the implementation of the retrieving, filtering, and clustering components of the system.

The results of the stemming experiments show how to find the word stems for indexing and document matching. Query length effect experiments provide an important clue about implementation of the filtering component in terms of user profile lengths. Document length effect and scalability experiments present valuable observations to set the parameters such as stemmer by considering document length and documents collection size.

6.1 Stemming Effects

Turkish Language

Turkish is an agglutinative language similar to Finnish and Hungarian. Such languages carry syntactic relations between words or concepts through discrete suffixes and have complex word structures. Turkish words are constructed using inflectional and derivational suffixes linked to a root. In Turkish verbs can be converted into nouns and other forms as well as nouns can be converted into verbs and other grammatical constructs through affixation [LEW1988]. Turkish alphabet is based on Latin characters and has 29 letters consisting of 8 vowels and 21 consonants. The letters in alphabetical order are a, b, c, ç, d, e f, g, ğ, h, ı, i, j, k, l, m, n, o, ö, p, r, s, ş, t, u, ü, v, y, and z (in this list the vowels are: a, e, ı, i, o, ö, u, ü). In some words borrowed from Arabic and Persian, the vowels “a”, “i,” and “u” are made longer or softer by using the character ^ (circumflex accent) on top of them. In modern spelling, this approach is rarely used.

In Turkish prefixation is uncommon and usually used to intensify the meaning of adjectives (and less commonly of adverbs). Such as “dolu” (full) and “dopdolu,” “tamam” (complete) and “tastamam” [LEW1988]. Such intensive adjectives are more suitable for story telling, but not for news articles. Prefixation in old fashioned words (such as “bîperva” which means “fearless,” here “bî” stands for “less”) or prefixation coming from western languages such as “antisosyal” (anti social) are infrequent in the language. Like English, nouns in Turkish do not have a gender and the suffixes do not change depending on word type. However, there are some irregularities in adding suffixes to the words, some examples are provided in [CAN2008a].

In Turkish, the number of possible word formations obtained by suffixing can be as high as 11,313 [HAK2000]. Like other agglutinative languages, in Turkish it is possible to have words that would be translated into a complete sentence in non-agglutinative languages such as English. In the news, the average word (token) length is approximately 6.90 characters [CAN2008a]. The average word length in the Kucera and Francis English corpus is given as 4.74 characters [KUC1967]. However, note that in Turkish word length in news are longer than other types of texts such as novels [CAN2006b]. This may be due to the factual content of the newspaper collection; the descriptive nature of such content may require the usage of longer words.

Stemming for Turkish

In this study, we use three stemming methods in obtaining vectors used for document description. They are (1) no stemming, so called "austrich algorithm," (2) first n, n-prefix, characters of each word, and (3) a lemmatizer-based stemmer. There is another stemming algorithm, which is based on the "successor variety" (SV) concept [HAF1974]. In the SV approach, the root of a word is determined according to the number of distinct succeeding letters for each prefix of the word to be stemmed can have in a large corpus. The idea is intuitively appealing due to agglutinative nature of the Turkish language. Our recent IR work show that the SV-based method provides a performance similar to the n-prefix and the lemmatizer-based methods and therefore it is not considered in this study [CAN2008a].

No-Stemming (NS). The no-stemming (NS) option uses words as they are as an indexing term. The performance with this approach provides a baseline for comparison.

Fixed Prefix Stemming (F5, F6). The fixed prefix approach is a pseudo stemming technique. In this method, we simply truncate the words and use the first n (Fn) characters of each word as its stem; words with less than or equal to n characters are used with no truncation. In this study, we experiment with F5 and F6, which are experimentally shown to give the best performance in IR [CAN2008a]. The success of this method can be explained with the fact that Turkish word roots are not much affected with suffixes [EKM2000].

Lemmatizer & Successor Variety (LV). A lemmatizer is a morphological analyzer that examines inflected word forms and returns their dictionary forms. Lemmatizers are much more sophisticated than stemmers. They also provide the type (part of speech, POS, information) of these dictionary forms, and the number and type of suffixes (morphemes) that follow the matched forms [OFL1994]. Lemmatizers are not stemmers, since the latter obtains the root in which a word is based; in contrast, a lemmatizer tries to find the dictionary entry of a word.

The Successor Variety (SV) algorithm determines the root of a word according to the number of distinct succeeding letters for each prefix of the word can have in a large corpus [FRA1992, HAF1974]. It is based on the intuition that the stem of a word would be the prefix at which the maximum SV is observed. For the working principles of the algorithm, please refer to the example provided in [FRA1992]. Our SV implementation chooses the longest prefix corresponding to the highest SV value since longer stems would have a better reflection of the meaning of the complete word (note that the same SV value can be observed for various prefix sizes).

For various items, including misspelled and foreign words, which cannot be analyzed by the lemmatizer, we use the S**V** method for such words; this crossbreed is referred to as LV.

Being an agglutinative language, Turkish has different features from English. For English, stemming may possibly yield "stems" which are not real words. Lemmatization on the other hand tries to identify the "actual stem" or "lemma" of the word, which is the base form of the word that would be found in the dictionary. Due to the nature of English, sometimes words are mapped to lemmas, which apparently do not have any surface connection as in the case of “better” and “best” being mapped to “good”. However, Turkish does not have such irregularities and it is always possible to find the “stem” or “lemma” of any given word through application of grammar rules in removing the suffixes. In the thesis, we prefer the word “stemming” over lemmatization; as it is more commonly used, and the algorithm we use internally identify the suffixes and remove them in the stemming process.

In the lemmatization process, in most of the cases we obtain more than one result for a word. In such cases, the selection of the correct word stem (lemma) is done by using the following steps [ALT2007]. (1) Select the candidate whose length is closest to the average stem length for distinct words for Turkish; (2) If there is more than one candidate, then select the stem whose word type (POS) is the most frequent among the candidates.

6.2 Matching (Ranking) Function

We use a matching function, which is referred to as MF8 in this study, for query document matching [LON2003, WIT1999]. The details behind the use of this matching function are presented in [CAN2008a]. MF8 calculates matching value for document d_j with the query Q by using the following formula.

$$\sum_{t \in Q} \left(1 + \ln f_{dt})/\sqrt{D}\right) \cdot (f_{qt} \cdot \ln(1 + N/f_t))$$

In this formula, f_{dt} is the frequency of term t in document d_j, D is the total number of term occurrences in d_j, f_{qt} is the frequency of term t in the query Q, N is the total number of documents in the collection, and f_t is the frequency of term t in the entire document collection. Note that *MF*8 is especially suitable for dynamic environments since in dynamic collections one can easily reflect the effects of *idf* to the term weighting scheme via query term weights (the second product item of the MF8 formula).

6.3 Stopword List Effects on Retrieval Effectiveness

In this section, we analyze the effects of stopword list on retrieval effectiveness. In the first set of experiments we measure bpref values [BUC2004] using the semi manually constructed stopword list (of Appendix B.1) and without using a stopword list. The results presented in Table 6.1 along with two-tailed t-tests show that stopword list have no significant impact on performance. Note that the assessors (query owners) are told nothing about the use of frequent Turkish words; nevertheless, such words have not been used heavily in the queries. For example, in Q_M on the average a query contains 1.74 stopwords (The query characteristics are provided in Tables 6.2 and 6.3).

Table 6.1: Bpref values using Q_M with (NS, F5, LV) and without (NS', F5', LV') a stopword list

NS	*NS'*	*F5*	*F5'*	*LV*	*LV'*
0.3255	0.3287	0.4322	0.4330	0.4504	0.4524

In the above approach, we use the stopword list to eliminate words before entering them to the stemmers. As an additional experiment, we have used the stopword list after stemming. For this purpose, we first used F5 stemmer to find the corresponding stem and after that, we search the stemmed word in the stemmed stopword list. The experiments again show no statistically significant performance change.

In order to observe the possible effects of automatic stopword list generation we also use the automatically generated stopword lists of the most frequent 288 words, and 10 words of IR test collection. The IR effectiveness performance with them is not statistically significantly different from the case with no stopword list [CAN2008a].

From these observations, we conclude that the use of a stopword list have no significant effect on Turkish IR performance.

6.4 Query Length Effects

In an IR environment, depending on the needs of the users, we may have queries with different lengths. For this reason, we analyze the effects of query lengths on effectiveness. The query types according to their lengths are described in Tables 6.2 and 6.3. More details on the queries can be seen in [CAN2008a]

Table 6.2: Query statistics

Entity	Min.	Max.	Median	Avg.
No. of unique words in short queries (Q_S)*	1	7	3	2.89
No. of unique words in medium queries (Q_M)*	5	24	11	12
No. of unique words in long queries(Q_L)*	6	59	26	26.11

* With stopwords.

Table 6.3: Query word statistics

Entity	Q_S	Q_M	Q_L
No. of Words	208	1004	2498
No. of Unique Words	182	657	1359
Avg word Length	7.03	7.57	7.62
Avg Unique Word Length	7	7.75	8.04

The experimental results are summarized in Figure 6.1. The figure shows that as we go from Q_S to Q_M we have a statistically significant ($p < 0.01$) increase in performance using F5 and LV. Improvements in effectiveness for them, respectively, are 14.4%, and 13.5%. The tendency of performance increase can be observed as we go from Q_M to Q_L, but this time the increase is statistically insignificant. For all query cases, under the same query form the performance difference of F5 and LV is statistically insignificant. However, the performance difference of these stemmers with respect to NS is statistically significant ($p < 0.001$). In terms of NS, the performance increase is first 6.23% and then 14.59% as we increase the query lengths incrementally. The second increase is statistically significant ($p < 0.01$). In other words, NS gets more benefit from query length increase. Also, the negative impact of not being stemmed is partly recovered with the increase in query length. From the experiments, we observe that there is no linear relationship between query lengths and retrieval effectiveness. That is, as we increase the query length, first we have an improvement and after a certain length increase this effectiveness increase tends to saturate. However, the NS approach improves its performance as we increase the query length.

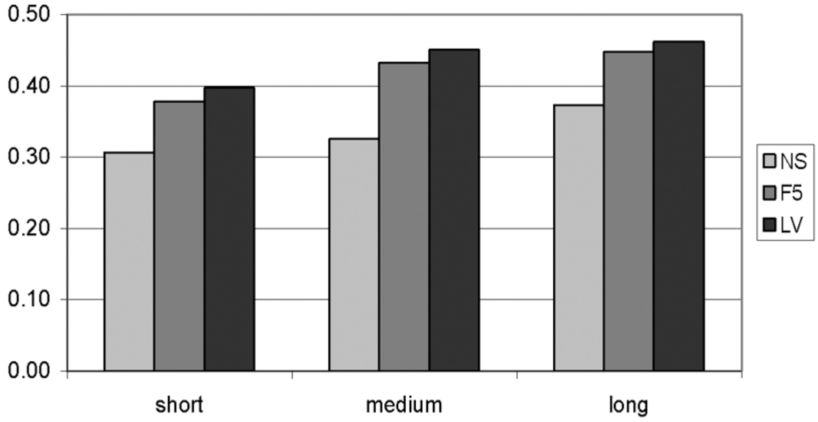


Figure 6-1: Bpref values with various query lengths.

The effectiveness improvement can be attributed to the fact that longer queries are more precise and provide better description of user needs. Similar results are reported in other studies regarding the effects of increasing query lengths. For example, Can, Altingovde, and Demir report similar results for increasing query lengths with the Financial Times TREC collection [CAN2004].

6.5 Document Length Effects

In different application environments, it is possible to have documents with different lengths. For this reason we divided the IR test collection according to document lengths and obtained three sub-collections that consist of short (documents with maximum 100 words), medium length (documents with 101 to 300 words), and long documents (documents with more than 300 words). In a similar fashion, we divided the relevant documents of the queries among these sub collections as we have done in the scalability experiments. Table 6.4 shows that most of the relevant documents are associated with the collection with medium sized documents. This can be explained by its size, it contains almost half of the full collection.

Table 6.4: Document collection characteristics for documents with different lengths

Collection Doc. Type	No. of Docs.	No. of Active Queries	Total No. of Unique Relevant Docs.	Avg. No. of Rel. Doc./Query	Median No. of Rel. Doc./Query
Short	139.13	72	1864	27.50	18.5
Medium	193.144	72	3447	52.14	45.0
Long	76.031	72	1612	24.67	21.0

In the experiments, we use the query form Q_M. The graphical representation of bpref values in Figure 6.2 shows that as the document sizes increase the effectiveness in terms of bpref values significantly increases ($p < 0.001$) and this is true for all stemming options. We have done no objective analysis of the average number of topics per news articles of the IR test collection. However, it is our anecdotal observation that in the overwhelming majority of the news articles only one topic is covered. Hence, the persistent increase in effectiveness as the document length increases can be attributed to the fact that longer documents provide better evidence about their contents and hence better discrimination during retrieval. Our result of having better performance with longer documents (news articles) is consistent with the findings of Savoy (1999, Tables 1a, 1b, A.1). However, note that when document size increases, document representatives could be more precise until a given limit. After this point, we may expect to see the inclusion of more details or non-relevant aspects (under the implicit assumption of a newspaper corpus). Thus longer documents could hurt the retrieval performance in such cases.

In the experiments, for all document length cases, the performance difference of F5 and LV is statistically insignificant. However, the performance difference of these stemmers with respect to NS is statistically significant ($p < 0.001$).

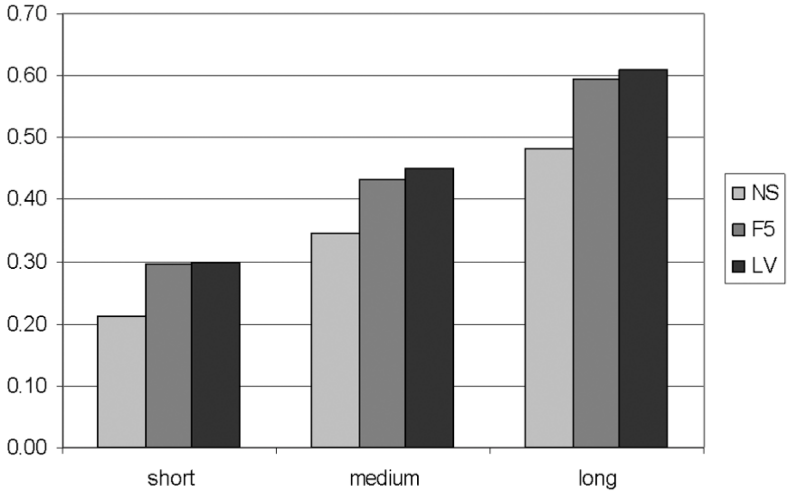


Figure 6-2: Query Characteristics and bpref values using Q_M for different document length collections.

6.6 Scalability Effects

Scalability is an important issue in IR systems due to the dynamic nature of document collections. For this purpose, we have created eight test collections in 50,000 document increments of the original IR test collection. The first one contains the initial 50,000 documents (in temporal order) of the IR collection, the second one contains the first 100,000 documents and is a superset of the first increment. The final step corresponds to the full version of the document collection.

For evaluation, we use the queries with at least one relevant document in the corresponding incremental collection. For example, for the first 50,000 documents, we have 57 active queries, (i.e., queries with at least one relevant document in the first 50,000 documents). Table 6.5 shows that each increment has similar proportional query set characteristics; for example, median number of relevant documents per query increases approximately 10 by 10 (11.0, 21.5, 34.0, etc.) at each collection size increment step. This means that experiments are performed in similar test environments.

Table 6.5: Query relevant document characteristics for increasing collection size.

No. of Docs	No. of Active Queries	Total No. Of Unique Relevant Docs.	Avg. No. Of Rel. Doc/Query	Median No. of Rel. Doc/Query
50,000	57	719	10.72	11.0
100,000	62	1380	21.08	21.5
150,000	63	2014	30.55	34.0
200,000	64	2944	44.33	45.5
250,000	68	3764	56.51	56.5
300,000	70	4794	71.45	66.0
350,000	71	5725	86.29	79.0
408,305	72	6923	104.30	93.0

Understanding the retrieval environments in more detail might be of interest. The characteristics of the collections as we scale up are shown graphically in Figures 6.3 and 6.4. Figure 6.3 shows that the number of unique words increases with the increasing collection size; however, F5 and LV show saturation in the increase of unique words as we increase the number of documents, and this is more noticeable with F5. Figure 6.4 shows that the number of postings (i.e., <document number, term weight> pairs or tuples) in the inverted files of NS, F5, and LV linearly increases as we increase the collection size. The graphical representation of posting list sizes of this figure indicates that with NS we have many short posting lists.

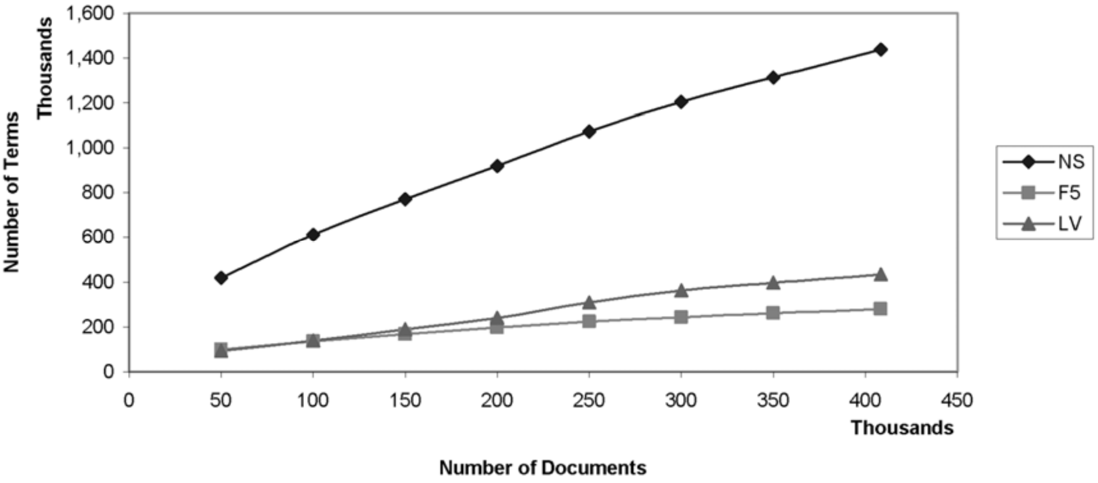


Figure 6-3: Indexing vocabulary size vs. collection size.

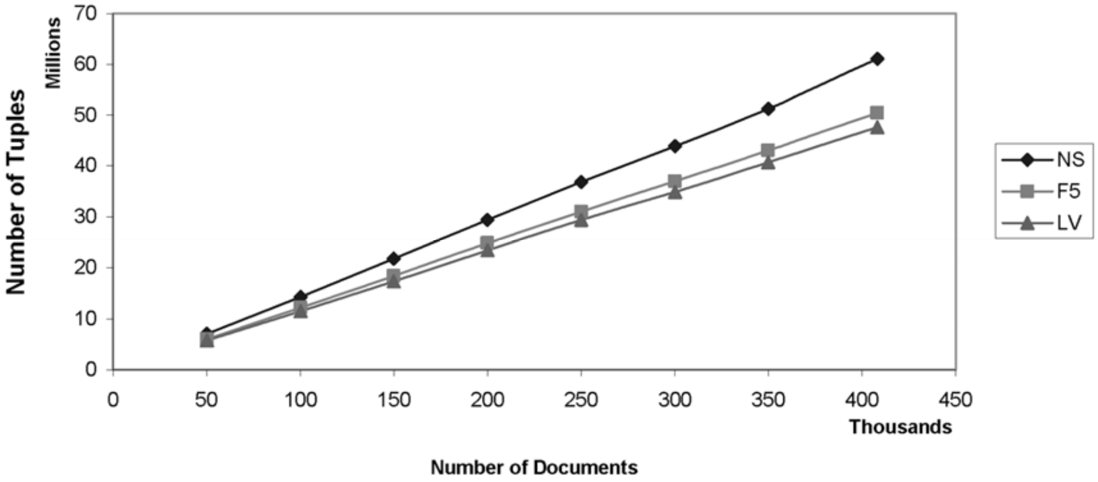


Figure 6-4: Number of posting list tuples vs. collection size.

The performance of NS, F5, and LV in terms of bpref as we scale up the collection is presented in Figure 6.5. With the first increment, we have a relatively better performance with respect the performances of the next three steps. In the second incremental step, i.e., with 100,000 documents, we have a decrease in performance, then we have a tendency of performance increase and beginning with 250,000 documents, we have a steady retrieval performance. This can be attributed to the fact that after a certain growth document collection characteristics reaches to a steady state.

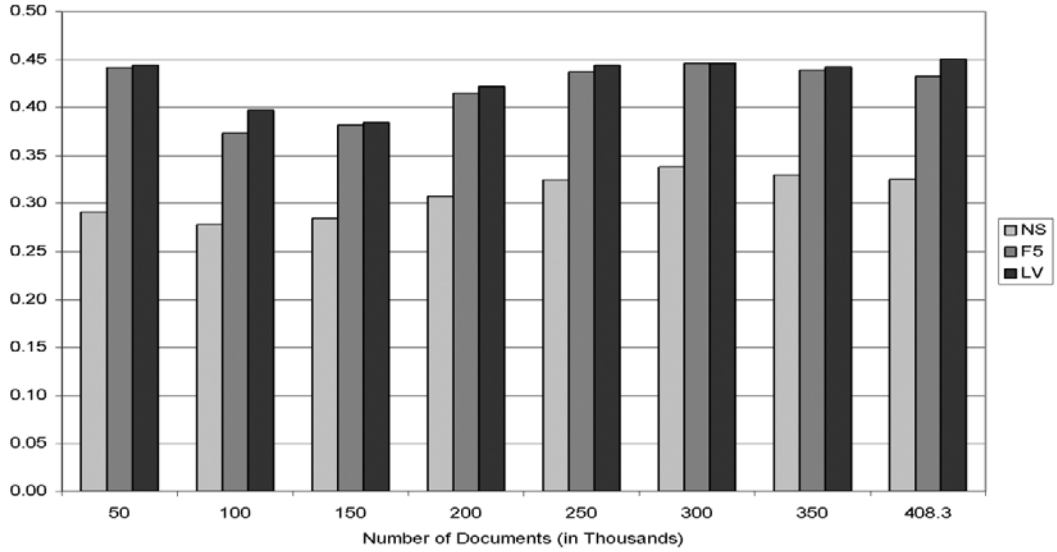


Figure 6-5: Bpref values for NS, F5, and LV using Q_M as collection size scales up.

The LV stemmer, which is designed according to the language characteristics, shows no improved performance as the collection size increases: simple term truncation method of F5 and LV are compatible with each other in terms of their retrieval effectiveness performances throughout all collection sizes. LV provides slightly better, but statistically insignificant performance improvement with respect to F5. However, the performance of F5 and LV with respect to NS is statistically significantly different (in all cases $p < 0.001$).

6.7 Chapter Summary and Recommendations

We show that within the context of Turkish IR,

- a stopword list has no influence on system effectiveness;
- a simple word truncation approach and an elaborate lemmatizer-based stemmer provide similar performances in terms of effectiveness;
- longer queries improve effectiveness; however, this increase is not linearly proportional to the query lengths; and

- longer documents provide higher effectiveness.

Therefore, we did not perform stopword list elimination in retrieval and filtering components in Bilkent News Portal. We added fixed prefix stemming to the Lemur Toolkit to use F5 stemmer on our queries and documents. During document indexing of news articles we keep the full text of the documents in the collection, and during information filtering, we aim to generate medium size queries for better performance.

Chapter 7
Future Pointers:
Large-Scale Implementation

News portals can be considered as living environments that interact with both users and news providers continuously. They are one of the major applications in information technology which is developing rapidly with the needs of the market. Adapting to these changes requires inevitably use the latest technology. In this chapter, new approaches and extensions are described in the scope of large-scale implementation of our news portal. Firstly, a new hardware design is introduced, then major structures of the software architecture are presented. Finally, the chapter is concluded with the explanation of new services and extensions.

7.1 New Approaches to the Current Hardware Design

The news portal is designed to be an efficient and effective news provider. Every component added to the system increases the processing complexity and requires more processing power and more main memory. We aim to provide service to may users at the same time, an increase in the number of users would necessitate a design change in the near future. In this chapter, a new design approach is introduced along with its advantages.

The current architecture does not provide load balancing for a large number of concurrent connections. For each user, the server shares its resources to display personal services and when the number of connections increases the system would not respond efficiently to the requests. Our three layered new approach that aims to handle this problem is presented in Figure 7.1 [CAN2009].

The proposed architecture has a layered structure. Each layer plays a different role in the system. When a user connects to the portal, the system accepts the user at the first layer, Dispatcher Layer. The only responsibility performed by this layer is distributing the connections to the workstations which run on business layer by considering their current workload.

The Business Layer consists of workstations that run identical copy of the portal. All business logic is performed by this layer. It is the layer that involves all the functional algorithms handling information exchange between Data and Dispatcher Layers. The major advantage of this parallel structure is the scalability. When the demand increases, increasing the capacity according to demand is possible by adding new nodes. Additionally, this structure also provides a guaranteed continuous service. Even if one of the nodes is completely out of service, others continue to serve. Another advantage is upgrading the system is possible without stopping the service.

The major challenge in Business Layer is keeping the nodes synchronized. This job is done by an application that updates the nodes automatically from a single source. It is also planned that one extra node will be added to the Business Layer to test the brand new services in the real environment.

The third layer, Data Layer, is responsible for storing and managing the data. We expect that one server will be sufficient. However, when extension becomes a necessity, the number of data servers can be increased.

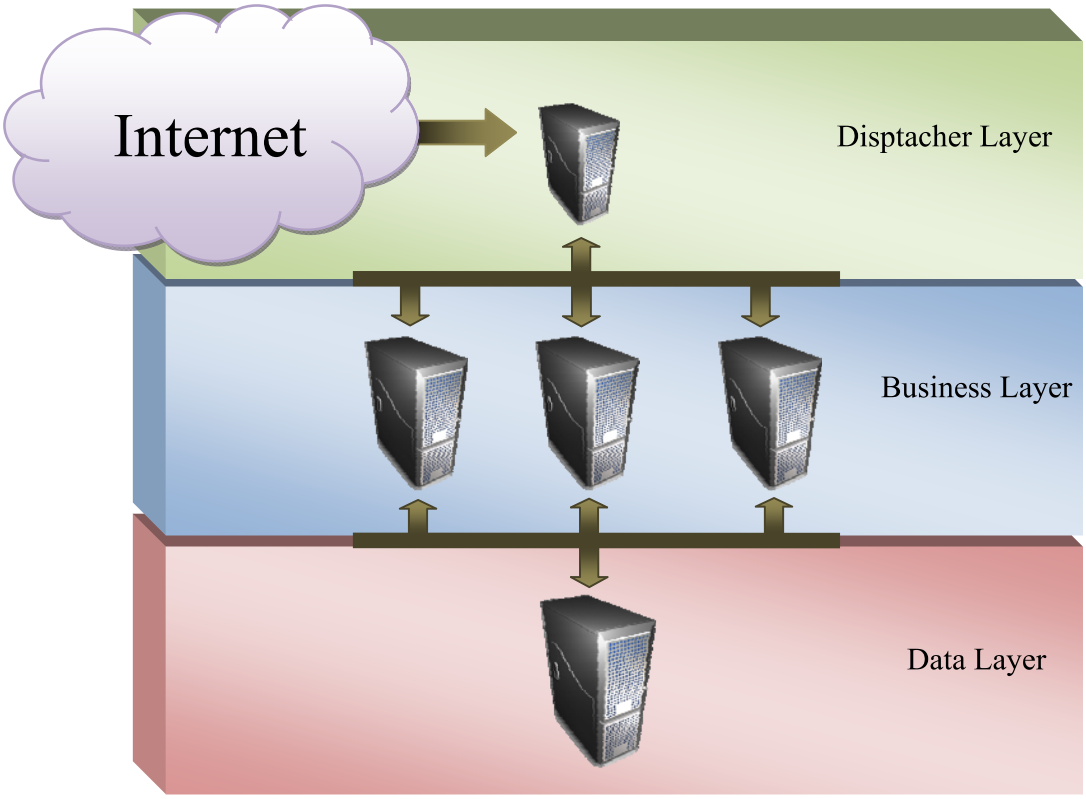


Figure 7-1: New hardware structure for Bilkent News Portal.

7.2 Web Services and SOAP

Web services become one of the most popular technologies that are widely used in information technologies to give opportunity to process data remotely. The foundation layer of a web services is called Simple Object Access Protocol (SOAP) that provides a basic messaging framework. The protocol relies on Extensible Markup Language (XML) as the message format, and the other Application Layer protocols (most notably Remote Procedure Call -RPC- and HTTP) for message negotiation and transmission. To use this protocol the system has to provide an interface which is called WSDL (Web

Service Definition Language). This interface defines how the functions running the services accept the input parameters and return the output.

We plan to provide such services to the public by using this software architecture. To enable this, WSDL documents of the services that are described in the previous chapters will be prepared. Content of the WSDL documents will be implemented and deployed to the workstations of the Business Layer. Our web interface needs to be modified to provide these services to the third party users.

7.3 New Services and Extensions

7.3.1 Topic Based Novelty Detection

The new topic based novelty detection (TBND) component aims to identify the new developments in the life time of a topic that consists of the first story of a new event and the tracking news articles, or simply trackings. The trackings may contain documents with no new information due to articles containing almost the same information provided by previous news articles. It is important to identify stories that include new information in topic tracking for facilitating easier topic perception by users. A topic based novelty detection service aims to identify these novel developments.

To the best of our knowledge, there is no TBND-related study for Turkish. This fact and expected positive impact of the 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 this service. 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 consumer trends in commercial data mining.

7.3.2 Faceted Interface for Information Retrieval

Improving query results with the relevance feedback is used in several systems to obtain the desired output in several areas including information retrieval. However, local news portals do not provide this feature because of its implementation complexity. This functionality requires cluster-based indexing of documents in the background. Simple search engines index the documents without generating clusters. However, if the documents are clustered according to its relevance to the cluster seed, then digging the query results in the same cluster could be possible with the relevance feedback.

The information retrieval component can be modified to perform cluster-based retrieval and the interface can display the terms that describes the clusters, and these terms can be chosen according to the query words. By selecting the terms one by one, users can expend their queries to access the results that are desired. This interface which works in this manner is called faced search interface which helps users to expand query results.

7.3.3 Communicating with the Services

The services which run on Business Layer will only be reachable by the operations (functions) which are defined in the WSDL document of the services. The user interface invokes the operations with the proper parameters and gets the result in XML. Also third party applications can integrate themselves as our user interface to use these services. The content that returns in XML format can be displayed by using (Extensible Stylesheet Language) XSL transformation.

The data layer can provide a web service which can be capable of performing all database and file operations for making an abstraction between layers. This approach can provide us an opportunity to expend the number of database server in data layer. The operations in the web service can allow performing SQL queries and basic file operations from a single interface.

7.3.4 Personalized News Search

For the near future, we are also planning to add this feature to our system. Our approach will be the combination of profile based and click based strategies. We are currently using click based personalization in information filtering, but profile based approach can be implemented. The major point is defining the thresholds of this hybrid system.

Chapter 8

Conclusions

The size of the digital universe is estimated as 281 exabytes (281 billion gigabytes) in 2007. It is expected that this amount will be 10 times larger in 2011 [GRA2008]. Although these indicators cover any type of digital content such as images and videos, we can approximately assume that the textual data growth also has a similar trend. Exponential growth of information makes the problem of managing and presenting digital data an interesting research problem. In this work, we present the design issues and experimental foundations of the first Turkish news portal with new event detection and tracking capabilities.

One of the primary requirements of IR and TDT applications is a test collection to measure the performance in a repeatable way. In Chapter 5, we present a web application (ETracker) to construct a test collection for new event detection and tracking applications for Turkish. Without a system like ETracker, test collection preparation requires manual evaluation. This is practically impossible due to the number of events to be annotated and the document collection size to be used in assessments. Since ETracker is a search-based system, we only look at some of the news articles. We also present the characteristics of the constructed test collection (BilCol2005). It is similar to the TDT test collections in terms of size and preparation procedures. It is used in the development of the new event detection and tracking components of Bilkent News Portal. BilCol2005, which will be shared with other researchers, is a significant contribution of this study.

In the development of Bilkent News Portal, IR tools are used for various purposes. As ETracker is a search-based system, in Bilkent News Portal the search and filtering facilities require an IR tool. In Chapter 6, we present experiments that help us to choose the right parameters in these tools. The experiments show that medium length queries and longer documents provide more effective result for Turkish IR. Based on these findings, during document indexing of news articles, we prefer to keep their full text in the news portal, and during information filtering for better effectiveness we prefer to use medium size user profiles.

In the thesis, we also present a large-scale implementation architecture with enhanced information processing capabilities. The proposed design provides practical pointers for future studies on the news portal.

References

[ALL2002] J. Allan, Introduction to topic detection and tracking. In J. Allan (Ed.), Topic Detection and Tracking Event-based Information Organization: (pp. 1-16). Norwell, MA: Kluwer Academic Publishers, 2002.

[ALS2008] I. S. Altingovde, E. Demir, F. Can, Ö. Ulusoy. Incremental cluster-based retrieval using compressed cluster-skipping inverted files. *ACM Transactions on Information Systems (TOIS)*, 26(3), 1-36, 2008.

[ALT2007] K. Altintas, F. Can, J. M. Patton. Language change quantification by time-separated parallel translations. *Literary and Linguistic Computing*, 22(4), 375-393, 2007.

[BAG2009] Ö. Bağlıoğlu. New Event Detection using Chronological Term Ranking. M.S. Thesis, Department of Computer Engineering, Bilkent University, Ankara, Turkey, 2009.

[BIL2009] Bilkent News Portal. (2009). Retrieved May 31, 2009, from http://newsportal.bilkent.edu.tr/PortalTest/.

[BEL1992] N. J. Belkin, W. C. Croft. Information filtering and information retrieval: two sides of the same coin?, *Communications of the ACM*, 35(12), 29-38, 1992.

[BUC2004] C. Buckley, E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th International Conference on Research and Development in Information Retrieval (ACM SIGIR '04),* pp. 25-32, 2004.

[CAN1990] F. Can, E. Ozkarahan. Concepts and effectiveness of the cover-coefficient-based clustering methodology for text databases, *ACM Transactions on Database Systems (TODS)*, 15(4), 483-517, 1990.

[CAN1993] F. Can. Incremental clustering for dynamic information processing, *ACM Transactions on Information Systems (TOIS)*, 11(2), 143-164, 1993.

[CAN2004] F. Can, I. S. Altingovde, E. Demir. Efficiency and effectiveness of query processing in cluster-based retrieval. *Information Systems*, 29(8), 697-717, 2004.

[CAN2008a] F. Can, S. Kocberber, E. Balcik, C. Kaynak, H. C. Ocalan, O. M. Vursavas. Information retrieval on Turkish texts. *Journal of the American Society for Information Science and Technology*, 59(2), 407-421, 2008.

[CAN2008b] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, E. Uyar. Bilkent News Portal: A personalizable system with new event detection and tracking capabilities [Demo paper]. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (ACM SIGIR'08),* pp. 885, 2008.

[CAN2009a] F. Can, S. Koçberber, Ö. Bağlıoğlu, G. Ercan, S. Kardaş, H. Ç. Öcalan, E. Uyar, L. Koç. Haber Portallarında Yenilikçi Yaklaşımlar. In *Proceedings of Akademik Bilisim 2009*.

[CAN2009b] F. Can, S. Koçberber, O. Bağlıoğlu, S. Kardaş, H. C. Ocalan, E. Uyar. Topic detection and tracking in Turkish. *Journal of the American Society for Information Science and Technology* (under revision).

[CIE2002] C. Cieri, S. Strassel, D. Graff, N. Martey, K. Rennert, M. Liberman. Corpora for topic detection and tracking. In J. Allan (Ed.), *Topic Detection and Tracking Event-based Information Organization*, pp. 33-66, 2002.

[CNN2009] CnnTürk. (2008). CNNTürk. Retrieved June 19, 2008, from http://www.cnnturk.com.

[CON2004] M. Connell, A. Feng, G. Kumaran, H. Raghavan, C. Shah, J. Allan. UMass at TDT 2004. Retrieved July 30, 2008, from http://maroo.cs.umass.edu/pub/web/getpdf.php?id=507.

[DAS2007] A. S. Das, M. Datar, A. Garg, S. Rajaram. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 271-280, 2007.

[DOU2007] Z. Dou, R. Song, J. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 581-590, 2007.

[EKM2000] F.C. Ekmekcioglu, P. Willett. Effectiveness of stemming for Turkish text retrieval. *Program*, 34, 195-200, 2000.

[ERC2009] G. Ercan, F. Can. Cover Coefficient-Based Multi-document Summarization. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval (ECIR)*, pp. 670 - 674, 2009.

[FRA1992] W.B. Frakes, R. Baeza-Yates. *Information Retrieval: Algorithms and Data Structures*. Englewood Cliffs, NJ: Prentice Hall. 1992.

[GOG2009] Google. (2009). Google News. Retrieved May 16, 2009, from http://news.google.com/.

[GRA2008] J. F. Gantz, C. Chute, A. Manfrediz, S. Minton, D. Reinsel, W. Schlichting, A. Toncheva. The diverse and exploding digital universe. *An International Data Cooperation White Paper*. March 2008, Retrieved May 31, 2009, from http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

[HAB2009] Haber7. (2009). Haber 7. Retrieved May 16, 2009, from http://www.haber7.com.

[HAF1974] M. A. Hafer, S. F. Weiss. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10, 371-385, 1974.

[HAK2000] D. Z. Hakkani-Tür. Statistical language modeling for agglutinative languages. Ph.D. Thesis, Department of Computer Engineering, Bilkent University, Ankara, Turkey, 2000.

[KAR2009] S. Kardaş. New Event Detection and Tracking in Turkish. M.S. Thesis, Department of Computer Engineering, Bilkent University, Ankara, Turkey, 2009.

[KHU2007] D. A. Khudhairy. Multilingual news gathering, information mining and analysis. *European Communities*, 2007. Retrieved June 9, 2009, from http://press.jrc.it.

[KOC2009] L. Koc, F. Can, S. Kocberber. Generic content and image extraction from web news. 2009, submitted for publication.

[KUC1967] H. Kucera, W. N. Francis. *Computational Analysis of Present-Day American English*. Rhode Island: Brown University Press, 1967.

[LEM2009] Lemur Project. The Lemur Toolkit for Language Modeling and Information Retrieval. Retrieved May 16, 2009, from http://www.lemurproject.org, 2009.

[LEW1988] G. L. Lewis. *Turkish Grammar, 2nd ed*. Oxford University Press, Oxford, 1988.

[LIU2007] K. Liu, W. Meng, J. Qiu, C. Yu, V. Raghavan, Z. Wu, Y. Lu, H. He, H. Zhao. AllInOneNews: development and evaluation of a large-scale news metasearch engine. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, ACM SIGMOD'07,* pp. 1017 – 1028, 2007.

[LON2003] X. Long, T. Suel. *Optimized query execution in large search engines with global page ordering*. In *Proceedings of the 29th Very Large Data Bases Conference (VLDB 2004)* pp. 129–140, 2003.

[MCK2002] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, A. Nenkova, C. Sable, B. Schiffman, S. Sigelman. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of HLT 2002, Second International Conference on Human Language Technology Research*, pp. 280-285, 2002.

[MCK2003] K. McKeown, R. Barzilay, J. Chen, D. Elson, D. Evans, J. Klavans, A. Nenkova, B. Schiffman, S. Sigelman. Columbia's newsblaster: new features and future directions. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations*, pp.15-16, 2003.

[MIL2009] Milliyet. (2009). Milliyet Gazetesi. Retrieved May 16, 2009, from http://www.milliyet.com.tr.

[NUR2006] R. Nuray, F. Can. Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management*, 42(3), 595-614, 2006.

[OFL1994] K. Oflazer. Two-level description of Turkish morphology. *Literary and Linguistic Computing*, 9(2), 137-148, 1994.

[RAD2001] D. R. Radev, S. Blair-Goldensohn, Z. V. Zhang, R. S. Raghavan. NewsInEssence: a system for domain-independent, real-time news clustering and multi-document summarization. In *Proceedings of the first International Conference on Human Language Technology Research*, pp.1-4, 2001.

[RAD2005] D. R. Radev, J. Otterbacher, A. Winkel, S. Blair-Goldensohn. NewsInEssence: summarizing online news topics. *Communications of the ACM*, 48(10), 95-98, 2005.

[TDT2002] TDT. The 2002 topic detection and tracking (TDT2002) task definition and evaluation plan. *Technical Report Version 1.1, National Institute of Standards and Technology*, 2002.

[TDT2004] Annotation manual: Version 1.2 – August 4, 2004. Retrieved January 9, 2007, from http: projects.ldc.upenn.edu TDT5 Annotation TDT2004V1.2.pdf.

[TDT2008] TDT. Topic Detection and Tracking Evaluation. Retrieved June 18, 2008, from http://www.itl.nist.gov/iaui/894.01/tests/tdt/.

[TRT2009] TRT. Turkish Radio and Television. Retrieved May 16, 2009, from http://www.trt.net.tr.

[UYA2009] E. Uyar. Near-duplicate News Detection Using Named Entities. M.S. Thesis, Department of Computer Engineering, Bilkent University, Ankara, Turkey, 2009.

[WIK2009] Wikipedia. Portal: current events. Retrieved May 16, 2009, from http://en.wikipedia.org/wiki/Portal:Current_events.

[WIT1999] I. H. Witten, A. Moffat, T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images (2nd ed.)*. San Francisco: Kaufmann. 1999.

[ZAM2009] Zaman. Zaman Gazetesi. Retrieved May 16, 2009, from http://www.milliyet.com.tr.

Appendix A

Tables for BilCol2005Test Collection

Table A.1: News categories and number of annotated events in each category

Categ. No.	News Category	No. of Events	Categ. No.	News Category	No. of Events*
1	Elections	0	8	Financial News	2
2	Scandals/Hearings	10	9	New Laws	4
3	Legal/Criminal Cases	13	10	Sports News	5
4	Natural Disasters	0	11	Political and Diplomatic Meetings	2
5	Accidents	16	12	Celebrity Human Interest News	11
6	Acts of Violence or War	11	13	Miscalleneous news	8
7	Science and Discovery News	4	-	-	-

* Due to double category assignment to 6 events (number 42, 55, 57, 58, 64, 75) there are 86 events.

Table A.2: Summary information for annotated events

Event No.	Brief Description of Event (Event Type)	No. of Tracking Stories	Time Span (days)	Event Days Month/Day
1	Kars'ta Trafik Kazası 7 ölü (5)	20	203	05/28 - 12/16
2	Onur Air'in Hollanda'ya inişi yasaklandı (3)	159	203	05/12 - 11/30
3	Koreli bilim adamının kök hücre araştırması sahte (7)	8	11	12/19 - 12/29
4	Nema karşılığı kredi (8)	31	280	02/08 - 11/14
5	Tokyo'da trenlerde haremlik selamlık (13)	8	263	04/04 - 12/22
6	Londra metrosunda patlama (6)	454	175	07/07 - 12/28
7	Barbaros Çocuk Köyü'nde çocuk tacizi skandalı (2)	88	275	01/26 - 10/27
8	Formula G (10)	20	58	07/04 - 08/30
9	Karamürsel kaymakamı intihar etti (2)	6	7	01/04 - 01/10
10	400 koyun intihar etti (5)	10	8	07/08 - 07/15
11	Şemdinli olayları (2)	317	53	11/09 - 12/31
12	Türkiye'de kuş gribi (13)	229	83	10/10 - 12/31
13	Şampiyon Fenerbahçe (10)	115	222	05/22 - 12/29
14	Mortgage Türkiye'de (9)	375	357	01/07 - 12/29
15	2005 Avrupa Basketbol Şampiyonası (10)	78	297	01/15 - 11/07
16	Yüzüncü Yıl Üniversitinde ihale yolsuzlugu iddiası (2)	326	79	10/14 - 12/31
17	Kral Fahd hastaneye kaldırıldı (12)	51	77	05/27 - 08/11
18	Memurlarının bir üst dereceye terfisi (9)	52	110	01/06 - 04/25
19	Bill Gates Türkiye'ye geldi (12)	17	8	01/30 - 02/06
20	Mısır'da patlamalarda çok sayıda kişi öldü (6)	120	43	07/23 - 09/03
21	Atillâ İlhan vefat etti (12)	40	70	10/11 - 12/19
22	Ata Türk'ün ölümü (12)	43	47	09/18 - 11/03
23	DT Genel Müdürü Lemi Bilgin görevden alındı (12)	63	109	08/19 - 12/05
24	Universiade 2005 Yaz Spor Oyunları (10)	248	289	03/04 - 12/17
25	Yahya Murat Demirel Bulgaristan'da yakalandı (3)	192	345	01/03 - 12/13
26	Bağdat El Ayma köprüsünde izdiham (6)	29	9	08/31 - 09/08
27	Prof. Dr. Sadettin Güner'e saldırı (3)	41	291	01/08 - 10/25
28	Nestle'de mürekkepli süt (2)	11	2	11/22 - 11/23
29	Nermin Erbakan tedavi altında (12)	45	46	10/20 - 12/04
30	Ulubey'de çocukla annenin peş peşe ölümü (13)	6	31	05/19 - 06/18
31	15. Akdeniz Oyunları (10)	193	86	05/02 - 07/26
32	Kemal Derviş'in UNDP başkanı seçilmesi (8)	118	181	03/11 - 09/07
33	Irak başbakanı Caferi Tahran'ı ziyaret etti (11)	22	94	07/05 - 10/06
34	Gediz'de grizu patlaması (5)	39	36	04/21 - 05/26
35	Sarıgül'ün CHP'de kendini savunması (11)	110	352	01/02 - 12/19
36	Paris'de polisle göçmenler arasındaki çatışma (6)	245	51	10/29 - 12/18
37	Rock'n Coke açık hava müzik etkinliği (13)	11	5	09/02 - 09/06
38	Ankara Garı'nda tren kazası (5)	13	5	01/13 - 01/17
39	2005 Nobel tıp ödülü (7)	19	75	10/03 - 12/16
40	Kayseri Erciyes Üniversindeki bebek ölümleri (2)	39	60	08/03 - 10/01
41	Marburg virüsünden ölenler (13)	25	65	03/16 - 05/19
42	Gamze Özçelik'in görüntüleri (2,12)	43	116	08/29 - 12/22

Table A.2: (Cont.) Summary information for annotated events.

Event No.	Brief Description of Event (Event Type)	No. of Tracking Stories	Time Span (days)	Event Days Month/Day
43	Türkiye'nin ilk yediz bebekleri geliyor (7)	56	301	02/17 - 12/14
44	Yeni Türk Ceza Kanunu (9)	53	193	06/01 - 12/10
45	Saddam Hüseyin'in yargılanmasına başlandı (3)	182	72	10/19 - 12/29
46	Beylikdüzü çöpte patlama (6)	17	5	11/18 - 11/22
47	Endonezya'nın Bali Adası'nda 4 bomba patladı (6)	15	4	10/01 - 10/04
48	Sahte rakı (3)	323	182	03/01 - 08/29
49	Hindistan'da üç saldırıda 66 kişi öldü (6)	21	5	10/29 - 11/02
50	Bülent Ersoy ve Deniz Baykal polemiği (12)	52	132	08/19 - 12/28
51	Tahran'da askeri uçak düştü (5)	9	2	12/06 - 12/07
52	Sochi seferinde Ufuk-1 gemisi yanmaya başladı (5)	20	3	08/25 - 08/27
53	İstanbul'da kanalizasyonda işçiler zehirlendi (5)	9	2	12/05 - 12/06
54	İKadınlara copla müdahele eden polisler (3)	104	297	03/06 - 12/27
55	Kuşadası'nda minibüsdeki patlama (3, 6)	50	4	07/16 - 07/19
56	Esenboğa Havalimanı iç hatlarda yangın (5)	18	36	11/14 - 12/19
57	Zeytinburnu'nda bir evde patlama (3, 6)	28	4	08/08 - 08/11
58	Malatya çocuk yuvasında işkence (2, 3)	192	67	10/26 - 12/31
59	ABD denizaltısı ile Türk gemisi çarpıştı (5)	7	1	09/05 - 09/05
60	Prof. Dr. Kalaycı suikast sonucu öldürüldü (3)	44	23	11/11 - 12/03
61	İlk yüz nakli (7)	14	17	12/01 - 12/17
62	15 yeni üniversite kurulmasina ilişkin kanun (9)	59	50	11/12 - 12/31
63	Gaziantep tanker patlaması (5)	33	7	08/06 - 08/12
64	Hakkari'de bomba patladı (3, 6)	10	4	07/29 - 08/01
65	Erzurum çocuk yuvasında bebek ölümlü (2)	9	3	11/04 - 11/06
66	Kâzım Koyuncu'nun ölümü (12)	30	129	06/25 - 10/31
67	Melih Kibar'ın ölümü (12)	16	120	04/07 - 08/04
68	Sarıkamış şehitleri anıldı (13)	5	3	12/23 - 12/25
69	Endonezya'da yolcu uçağı düştü (5)	15	2	09/05 - 09/06
70	Şanlıurfa'da köprü inşaatı çöktü (5)	7	2	04/13 - 04/14
71	Japonya Osaka'da tren kazası (5)	29	4	04/25 - 04/28
72	Manken Tuğçe Kazaz hıristiyan oldu (12)	11	76	09/22 - 12/06
73	Fotoğraf sanatçısı Mehmet Gülbiz öldürüldü (3)	14	127	02/04 - 06/10
74	Atina'daki Kara Harp Okulu'nda Türk bayrağı olayı (2)	55	71	04/16 - 06/25
75	Maslak'ta patlama (3, 6)	30	18	10/15 - 11/01
76	Didim'de denize uçak düştü (5)	13	2	07/19 - 07/20
77	Rum yolcu uçağı düştü (5)	106	115	08/14 - 12/06
78	İstiklal Caddesindeki ağaçlar kaldırıldı (13)	8	16	11/02 - 11/17
79	Zeytinburnu'nda gemi battı (5)	38	3	03/13 - 03/15
80	İngiltere'de Osmanlı kültürü hakkında sergi açıldı (13)	22	103	01/01 - 04/13
Avg.	-	73	92	-
Min.	-	5	1	-
Max.	-	454	357	-

Appendix B

Tables for Information Retrieval

Table B.1: Stopword List with 147 words

ama
ancak
arada
ayrıca
bana
bazı
belki
ben
beni
benim
beri
bile
bir
birçok
biri
birkaç
biz
bize
bizi
bizim
böyle
böylece
bu
buna
bundan
bunlar
bunları
bunların
bunu
bunun
burada
çok
çünkü
da
daha
de
değil
diğer
diye
dolayı
dolayısıyla
edecek
eden
ederek
edilecek
ediliyor
edilmesi
ediyor
eğer
etmesi
etti
ettiği
ettiğini
gibi
göre
halen
hangi
hatta
hem
henüz
her
herhangi
herkesin
hiç
hiçbir
için
ile
ilgili
ise
işte
itibaren
itibariyle
kadar
karşın
kendi
kendilerine
kendini
kendisi
kendisine
kendisini
ki
kim
kimse
mı
mi
mu
mü
nasıl
ne
neden
nedenle
o*
olan
olarak
oldu
olduğu
olduğunu
olduklarını
olmadı
olmadığı
olmak
olması
olmayan
olmaz
olsa
olsun
olup
olur
olursa
oluyor
ona
onlar
onları
onların
onu
onun
öyle
oysa
pek
rağmen
sadece
şey
siz
şöyle
şu
şunları
tarafından
üzere
var
vardı
ve
veya
ya
yani
yapacak
yapılan
yapılması
yapıyor
yapmak
yaptı
yaptığı
yaptığını
yaptıkları
yerine
yine
yoksa
zaten

* Among letters only “o” is listed as a word (since it is a meaning word in Turkish).