GUIDING STUDENTS TO ANSWERS: Query Recommendation

Ozgür YILMAZEL
Anadolu University
Department of Computer Engineering
2 EylulCampus, Eskisehir, TURKEY

ABSTRACT

This paper reports on a guided navigation system built on the textbook search engine developed at Anadolu University to support distance education students. The search engine uses Turkish Language specific language processing modules to enable searches over course material presented in Open Education Faculty textbooks. We implemented a guided navigation engine by using query log mining to our application. It makes use of previous users' sessions to help students find information they are looking for by using fewer queries. We used item-based similarity to do query recommendation.

This paper describes the search application, query expansion and evaluation of the system over existing query logs. We show that our system suggested relevant queries with a success rate of 85%.

Keywords: Question expansion, query recommendation, indexing, searching, metadata generation, information retrieval

INTRODUCTION

In the last decade online university education has become ever more popular, because it allows many people to use the new communication medium to get advanced degrees even if they can't attend to universities as full time students. Anadolu University is one of the largest universities that offer higher education opportunities through distance education. The learning materials are textbooks, TV programs, academic counseling services, videoconferences, computers and Internet-based applications (http://www.anadolu.edu.tr). One of the internet based applications provided by the university is the e-book service, which allows students to access course textbooks via internet. These textbooks are published in PDF format, and offered as links to full PDF files (~ 300 pages per book). In order to find the information, students either should look through these PDF documents or they should know exact place of the information, which means spending more time and effort.

The aim of this research is to build a search engine that is aware of Turkish Language properties and also to build mechanisms to help students find relevant information faster to improve distance education. By developing a search and guided navigation system, we will enable students to get clear answers to their questions within context. Fewer query attempts to reach answers within context will assist them to get to the relevant material faster, and ultimately improve their learning experience. When searching through a large volume of documents students have to communicate their information need in query terms.

For instance a student studying in Anadolu University Open Education Faculty Veterinary Laboratory Services program would like to retrieve information related to 'the dangers of anthrax vaccine on farm animals', the student needs to formulate a highly-qualified query with appropriate words to be able to find answer to his information need.

Not being able to formulate a qualified query student would have to deal with information overload resulting in excess time and effort on both the student and our university systems. Highly qualified query is essential for the student to be able to retrieve relevant and precise information.However, formulating highly-qualified queries is a difficult task for the distance education students.

Generally web users use short queries that have less than two words to search engines (Ji-Rong, Jian-Yun, & Hong-Jiang, 2001). This is also true for our textbook search system, that 40% of all queries submitted are single word queries.

To help web users to formulate better queries, researchers have focused on query expansion methods (Hang, Ji-Rong, Jian-Yun, & Wei-Ying, 2002). In query expansion, by suggesting additional query terms, users give additional input on query words or phrases. Commercial web search engines, like Yahoo and Google, suggest related queries or related searches in response to a query. The user has a chance to use one of these alternative queries to refine his search (Manning, Raghavan, & Schütze, 2008).

In this paper, we implemented an algorithm to recommend new queries to students based on analysis of user logs. Our goal is to help students to reformulate their queries by recommending new alternative queries.

Remainder of the paper is organized as follows: section 2 gives an overview of our textbook search system, section 3 provides information about our query recommendation system, section 4 reports on our evaluation, and finally section 5 presents our results and conclusions.

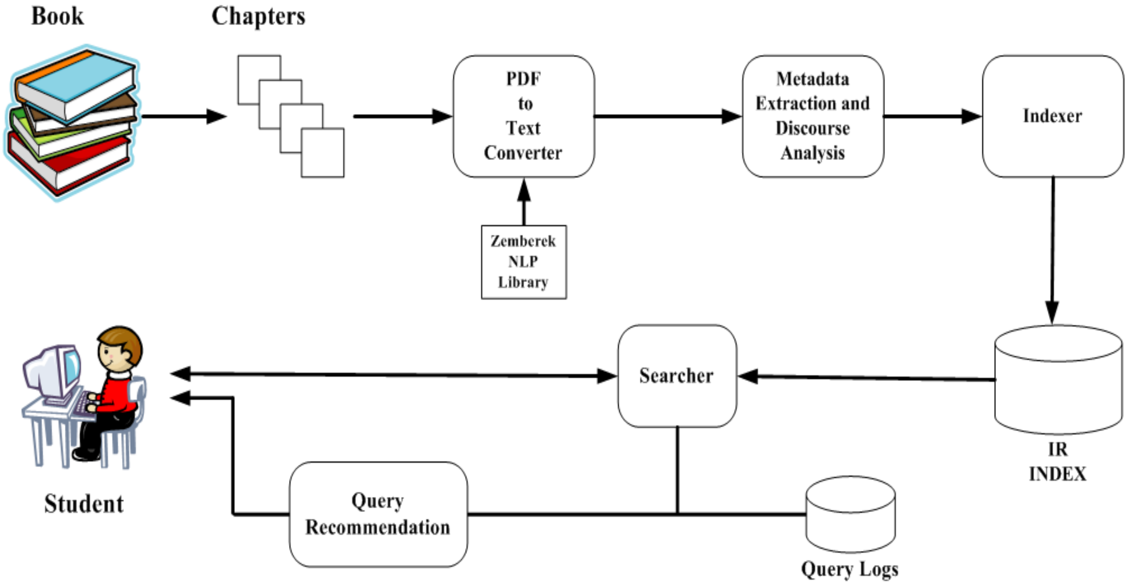


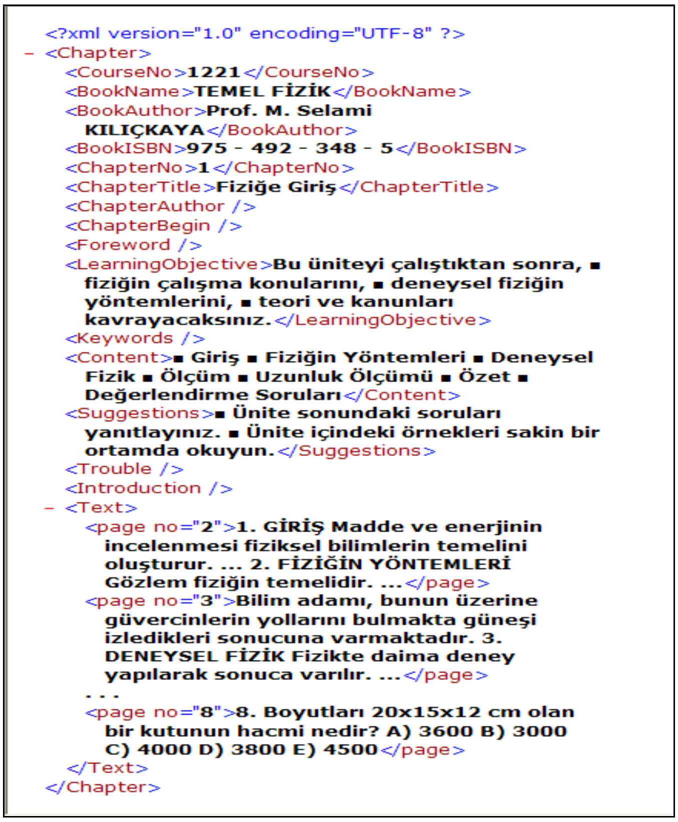
Figure: 1
System overview

TEXTBOOK SEARCH SYSTEM

Anadolu University Distance Education Textbook Search System consists of five different processes: preprocessing, metadata generation, indexer, searcher, and Query Recommendation module (see Figure 1). Yurekli et al. (Yurekli, Arslan, Senel, &Yilmazel, 2008) provides in depth explanation for each of the modules in the search system except the Query Recommendation Module. Below we provide a short overview of the modules.

Preprocessing

For automatic assignment of metadata, first the educational textbooks have been partitioned into chapters which resulted in 2654 chapters. Then, these chapters, which are in PDF format, have been converted to texts by using an open source java pdf library to work with PDF documents. While converting PDFs into texts, we have faced encoding and OCR problems. Especially Turkish language specific characters were corrupted. Once we corrected the erroneous characters we generated metadata information for each chapter. At this stage full-text of the chapters, which were obtained in the PDF to text conversion step, are converted into XML representations. Metadata such as author, summary, keywords and learning objectives are extracted so that we could display this information to the user in the result set. We followed similar research that has been done for automatic extraction of metadata (Yilmazel, Finneran, & Liddy, 2004) as it takes great amount of time and effort to create metadata of digital contents manually.



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <Chapter>
    <CourseNo>1221</CourseNo>
    <BookName>TEMEL FİZİK</BookName>
    <BookAuthor>Prof. M. Selami
      KILIÇKAYA</BookAuthor>
    <BookISBN>975 - 492 - 348 - 5</BookISBN>
    <ChapterNo>1</ChapterNo>
    <ChapterTitle>Fiziğe Giriş</ChapterTitle>
    <ChapterAuthor />
    <ChapterBegin />
    <Foreword />
    <LearningObjective>Bu üniteyi çalıştıktan sonra, ■
      fiziğin çalışma konularını, ■ deneysel fiziğin
      yöntemlerini, ■ teori ve kanunları
      kavrayacaksınız.</LearningObjective>
    <Keywords />
    <Content>■ Giriş ■ Fiziğin Yöntemleri ■ Deneysel
      Fizik ■ Ölçüm ■ Uzunluk Ölçümü ■ Özet ■
      Değerlendirme Soruları</Content>
    <Suggestions>■ Ünite sonundaki soruları
      yanıtlayınız. ■ Ünite içindeki örnekleri sakin bir
      ortamda okuyun.</Suggestions>
    <Trouble />
    <Introduction />
  - <Text>
      <page no="2">1. GİRİŞ Madde ve enerjinin
        incelenmesi fiziksel bilimlerin temelini
        oluşturur. ... 2. FİZİĞİN YÖNTEMLERİ
        Gözlem fiziğin temelidir. ...</page>
      <page no="3">Bilim adamı, bunun üzerine
        güvercinlerin yollarını bulmakta güneşi
        izledikleri sonucuna varmaktadır. 3.
        DENEYSEL FİZİK Fizikte daima deney
        yapılarak sonuca varılır. ...</page>
      . . .
      <page no="8">8. Boyutları 20x15x12 cm olan
        bir kutunun hacmi nedir? A) 3600 B) 3000
        C) 4000 D) 3800 E) 4500</page>
    </Text>
  </Chapter>
```

Figure: 2 XML document.

Indexer

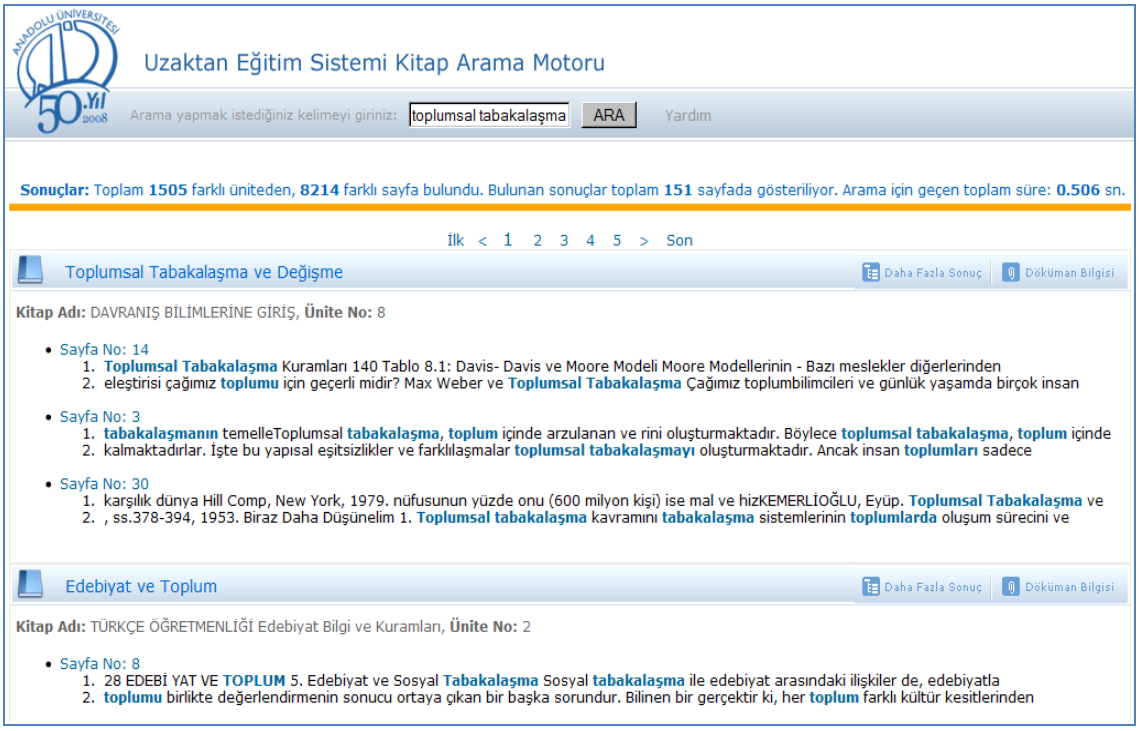
We identified a single page to be the unit of retrieval in our application and used Lucene, an open source information retrieval library, to create retrieval indexes for our textbooks. Since retrieval quality is highly dependent on the representation of the documents with correct terms, we used a Turkish Language Analyzer developed by Arslan and Yilmazel (Arslan & Yilmazel, 2008). Turkish Analyzer processes raw text and produces tokens to be used by the Lucene indexer. Turkish Analyzer removes any inflectional suffixes from the terms and leaves the derivational suffixes intact. Applies synonym filters to allow matching of similar terms when a synonym is used. It also produces ASCII versions of Turkish words that contain Turkish specific characters, which will allow matches over Turkish terms written in ASCII, which is a common unhandled case in many Turkish Search applications. Example: In our system if a user queries the word 'ogrenci' (not a valid Turkish word), he gets the results containing the word 'öğrenci' (student in English).

Table: 1
The equivalence of Turkish specific characters in English.

TurkishLetter	ç	ğ	I	ö	ş	ü
English Letter	c	g	i	o	s	u

Searcher

We developed a Web application that can return a ranked list of documents relevant to the given user query. Our results displayed book name, chapter title, page number and highlighted text snippets from the document to the student. Document metadata is also made available in the results screen by screen overlay. The first result page from our web application for the query "toplumsaltabakalaşma" (social stratification) is shown in Figure: 3.



Sonuçlar: Toplam **1505** farklı üniteden, **8214** farklı sayfa bulundu. Bulunan sonuçlar toplam **151** sayfada gösteriliyor. Arama için geçen toplam süre: **0.506** sn.

Toplumsal Tabakalaşma ve Değişme

Kitap Adı: DAVRANIŞ BİLİMLERİNE GİRİŞ, **Ünite No:** 8

- Sayfa No: 14
  1. **Toplumsal Tabakalaşma** Kuramları 140 Tablo 8.1: Davis- Davis ve Moore Modeli Moore Modellerinin - Bazı meslekler diğerlerinden
  2. eleştirisi çağımız **toplumu** için geçerli midir? Max Weber ve **Toplumsal Tabakalaşma** Çağımız toplumbilimcileri ve günlük yaşamda birçok insan
- Sayfa No: 3
  1. **tabakalaşmanın** temelleToplumsal **tabakalaşma**, **toplum** içinde arzulanan ve rini oluşturmaktadır. Böylece **toplumsal tabakalaşma**, **toplum** içinde
  2. kalmaktadırlar. İşte bu yapısal eşitsizlikler ve farklılaşmalar **toplumsal tabakalaşmayı** oluşturmaktadır. Ancak insan **toplumları** sadece
- Sayfa No: 30
  1. karşılık dünya Hill Comp, New York, 1979. nüfusunun yüzde onu (600 milyon kişi) ise mal ve hizKEMERLİOĞLU, Eyüp. **Toplumsal Tabakalaşma** ve
  2. , ss.378-394, 1953. Biraz Daha Düşünelim 1. **Toplumsal tabakalaşma** kavramını **tabakalaşma** sistemlerinin **toplumlarda** oluşum sürecini ve

Edebiyat ve Toplum

Kitap Adı: TÜRKÇE ÖĞRETMENLİĞİ Edebiyat Bilgi ve Kuramları, **Ünite No:** 2

- Sayfa No: 8
  1. 28 EDEBİ YAT VE **TOPLUM** 5. Edebiyat ve Sosyal **Tabakalaşma** Sosyal **tabakalaşma** ile edebiyat arasındaki ilişkiler de, edebiyatla
  2. **toplumu** birlikte değerlendirmenin sonucu ortaya çıkan bir başka sorundur. Bilinen bir gerçektir ki, her **toplum** farklı kültür kesitlerinden

Figure: 3 Web application.

There also exists links for More Results ("DahaFazlaSonuç") and Document Info ("DokümanBilgileri"), which shows the rest of the results for the document and gives summary information about the document, respectively.In addition to the highlighted snippets, a direct link to the page of the original PDF is also provided. Giving opportunity for accessing the original documents and to a specific page has many advantages like seeing figures, tables and visual components that cannot be converted to text.

Logging

We implemented a logging mechanism to capture user activity in our search system. The logs included session id, search text, ip address of the user and time stamp for the given query. Table 2 shows an example log entry.

Table: 2
Example of a record in the search logs data.

Field	Value
Date	2010-11-26
Time	09:34:10
Client IP address	88.231.144.112
Session ID	38FA336B32CED64F14AD338F3719C8AF
Query	islamdüşüncetarihi

QUERY RECOMMENDATION

As mentioned in the introduction section users need to formulate highly complex queries to find the specific information they are looking for. There have been various studies on using query recommendation and expansion techniques to ease the life of the user. In this section we will review the existing research and how it applies to our textbook search application and ultimately how it could improve the learning experience of our distance education students. We model query recommendation as a collaborative filtering task, in which past user queries will drive our recommendations.

Collaborative filtering (CF) is a complement to the content based filtering approaches that has been available for a long time. It recommends items to an active user based on correlations between the active user and other similar users (Jonathan, Joseph, Al, & John, 1999). Collaborative filtering applications can be grouped under two groups:

- **user based and**
- **item based collaborative filtering.**

User based collaborative filtering weights all users with respect to similarity to active user, selects a subset of users as predictors and uses normalized ratings to compute prediction from the selected neighbour's preferences. Similar to user based collaborative filtering item based collaborative filtering computes item to item similarities, selects most similar items for each item with respect to item-item similarity. The outcome is computed for the target item by observing target users outcomes for similar items. Item based collaborative filtering provides fast and effective results than user based collaborative filtering (Badrul, George, Joseph, & John, 2001).

In our application recommending new queries to students will require a scalable recommendation algorithm. By using query-query similarity modelling after item-based CF, we can isolate the steps of similarity computation phase from the prediction phase, and can perform scalable recommendations. Since query-query similarities are pretty static, item based recommendation is more applicable. We treat each search session as a single user and queries as items. We construct a matrix of similar queries offline (Yurekli, Capan, Yilmazel, & Yilmazel, 2009).

In this study we used query logs collected from our textbook search system that was in operation since April 2008. Query logs included a session id and search text. We grouped the queries submitted from a user within one session. Our goal is to find similar queries so that we could help a user to expand their current query, this requires us to capture as much similar queries as possible. We assume that each session focuses on one information need; hence queries within one session are related to each other. From the session data we chose sessions with at least 3 queries. There were a maximum of 41 queries for one session.

TABLE: 3
The experimental query set.

Popular	Moderate	Rare
iktisat	hayvanfizyolojisi	türkdili
sosyoloji	havataşımacılığı	korelasyon
işletme	kamuyönetimi	tahminleme
arapça	katmadeğervergisi	anayasahukuku
istatistik	aöf	istatistikserileri
ingilizce	toplumsaltabakalaşma	öfke
hukuk	türkmedeniyeti	kablosuz
matematik	fonksiyonlar	regresyonanaliz
hukukagiriş	normaldağılım	yahudi
Genelmuhasebe	televizyon	Davranışbilimleri

Out of 148560 unique sessions, on average each session had 2.7 queries, with a median of 2 queries. 22356 sessions had at least 3 queries or more with an average 4.56 queries. Session data is converted to a session-query matrix which had 37442 different log queries that was searched in 22356 sessions. The session query matrix obtained is 37442 x 22356, but there are only 101944 entries in the matrix, so the session-query matrix is very sparse.

We chose 10 query terms from most popular, moderate and rare terms randomly. Popular queries are searched over 1000 times where as moderate between 200-500, and rare less than 50 times. Table: 3 lists the 30 queries used in our experiments.

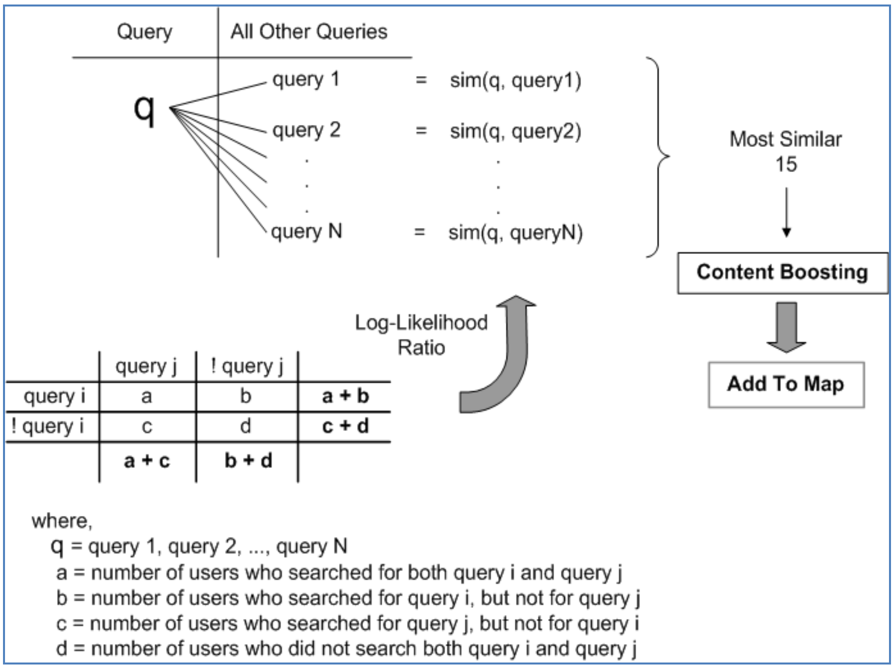
Query Based Collaborative Filtering

We used Apache Mahout's Taste module, which is a flexible, fast collaborative filtering engine for Java (Apache Mahout-Mahout Taste Documentation, http://lucene.apache.org/mahout/taste.html). Item based collaborative filtering uses query similarity. The idea is to recognize relations between items by analyzing the user-item (session-query) matrix and for a given pair predict related items based on these relations.

In other words, this method first computes similarity between items and then selects the most similar ones (Badrul, et al., 2001). In determination of similarities, we used Log-likelihood ratio (Jonathan, et al., 1999; Ted, 1993) which is a statistical analysis for computing similarity. We estimate that if two items are similar to each other, than they should be searched together.

Algorithm:
Query q:

```
        For each query {Compute similarity between each q and query}
```



	query j	! query j	
query i	a	b	a + b
! query i	c	d	c + d
	a + c	b + d	

where,
q = query 1, query 2, ..., query N
a = number of users who searched for both query i and query j
b = number of users who searched for query i, but not for query j
c = number of users who searched for query j, but not for query i
d = number of users who did not search both query i and query j

Figure: 4
Construction of thesimilaritymapfor item-basedcollaborativefilteringmethod.

For each query we computed similarity between that query and all other queries with log-likelihood ratio. Based on these similarities 10 queries that are most similar to that query have chosen. Among the set of most similar queries the ones which had more than or equal to 0.7 rank assigned as similar and added to candidate similar query set.

To populate 10 similar queries to the query q, we first consider the queries whose content matches with q an dtry to include up to 7 terms from this set. To divers if your expansion remaining 3 candidates are chosen from a set without term matches. Figure: 4 shows the construction of the similarity map for item-based collaborative filtering.

When a new query is submitted, by using this similarity map we can get the related queries recommended for this query. By this way for each of the 30 query in the experimental query set we obtained query recommendations.

Figure: 5 illustrate the top 10 expansion terms recommended for the query"matematik-math" by item-based CF method.

genelmatematik
lineercebir
analiz
finansalanaliz
soyutmatematik
matematikkitabı
elektronik
haberleşme
matematiksoruları
matrisler

Figure: 5
Expansion termsforquery "matematik-Math".

EVALUATION

The top 10 expansion terms for all the 30 queries found were judged manually by 5 human assessors. Each assessor judged the result either as 1 or 0, meaning relevant or not relevant respectively. To assess how consistent the ratings of the judges, we calculated the average pairwise kappa (Gwet, 2001), 0.8353. This kappa values show that there is an agreement between these judges. For all 30 queries in the experimental query set we analyzed the quality of the expansion terms found by the system. For each of the 10 expansion terms of a query, if at least three assessors agree that the expansion term is related to the query then that term is assigned as relevant. Out of 235 recommendations that our system made 200 of them was judged relevant and 35 of them judged as non-relevant to the current topic. The results show that item based collaborative filtering for query expansion produces relevant queries to the user with a success rate of 85%.

CONCLUSIONS AND FUTURE WORK

In this paper, we reported back on the textbook search engine that was developed for Anadolu University Distance Education students. The system supported full text faceted search over textbooks, and searches were aware of the Turkish language properties. Our analysis over the system logs showed that students have used it with various degrees of success for the last two years. To improve students' learning experience by suggesting query terms to search we implemented a query recommendation module. Query recommendation serves as a guided navigation in search applications. Experiments show that the system recommends correct queries with 85% success rate, which means the guided navigation system can successfully help students find information they are looking for in fewer queries.

As a future work, we plan on integrating the recommendations to the textbook search engine application and collect usage data. After empirical experiments to be made directly with system users (students), we plan to deploy the query recommendation engine on the live search engine.

A common problem with recommendation systems are with rare items. To improve our recommendation performance on rare terms we will work on smoothing methods to solve cold start problems. We believe our work on improving students' ease of access to relevant material faster will make a positive impact on students' academic success.

Acknowledgements

This research has been supported by Anadolu University as it gives permission to work on Distance Education System's course materials.

BIODATA and CONTACT ADDRESSES of AUTHORS

Dr. Özgür Yılmazel has published extensively in the field of natural language processing in Turkish language, text mining, text classification using natural language features. His present position is Associate Professor of the Faculty of Engineering and Architecture at Anadolu University and also the IT director of Anadolu University. His undergraduate and graduate studies are in Electrical Engineering field and he has earned a Ph.D. from Syracuse University specializing in the area of empirical selection of NLP-driven document representation for text categorization.

Associate Professor Özgür Yılmazel, PhD
Director, Center for Information Technology Research and Development,
Anadolu University, Yunus Emre Kampüsü, 26470 Eskişehir, TURKEY
Phone: (222) 335 05 80 or (222) 335 18 04
Email: ozgur@anadolu.edu.tr

REFERENCES

Apache Mahout-Mahout Taste Documentation, http://lucene.apache.org/mahout/taste.html

Arslan, A., & Yilmazel, O. (2008). A Comparison of Relational Databases And Information Retrieval Libraries On Turkish Text Retrieval. Paper presented at the IEEE NLP-KE'08.

Badrul, S., George, K., Joseph, K., & John, R. (2001). Item-Based Collaborative Filtering Recommendation Algorithms. Paper presented at the Proceedings of the 10th International Conference on World Wide Web.

Gwet, K. (2001). Statistical Tables for Inter-Rater Agreement. Gaithersburg: StatAxis Publishing.

Hang, C., Ji-Rong, W., Jian-Yun, N., & Wei-Ying, M. (2002). Probabilistic Query Expansion Using Query Logs. Paper presented at the Proceedings of the 11th International Conference on World Wide Web.

Ji-Rong, W., Jian-Yun, N., & Hong-Jiang, Z. (2001). Clustering User Queries of a Search Engine. Paper presented at the Proceedings of the 10th International Conference on World Wide Web.

Jonathan, L. H., Joseph, A. K., Al, B., & John, R. (1999). An Algorithmic Framework for Performing Collaborative Filtering. Paper presented at the Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval Cambridge, England: Cambridge University Press.

Ted, D. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1), 61-74.

Yilmazel, O., Finneran, C. M., & Liddy, E. D. (2004). Metaextract: An NLP System to Automatically Assign Metadata. Paper presented at the Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries.

Yurekli, B., Arslan, A., Senel, H. G., & Yilmazel, O. (2008). TURKISH QUESTION ANSWERING: Question Answering for Distance Education Students. Paper presented at the ICSOFT 2008.

Yurekli, B., Capan, G., Yilmazel, B., & Yilmazel, O. (2009). Guided Navigation Using Query Log Mining through Query Expansion. Paper presented at the In Proceedings of 2009 Third International Conference on Network and System Security, Gold Coast, Australia.