International Journal of Information Science and Management
Vol. 21, No. 3, 2023, 187-208
DOI: https://doi.org/10.22034/ijism.2023.1977626.0 / DOR: https://dorl.net/dor/20.1001.1.20088302.2023.21.3.11.5

Original Research

FarsAcademic: A Standard Persian Test Collection for Information Retrieval in Scientific Texts

Davoud Haseli
Assistant Prof., Department of Knowledge and Information Science, Kharazmi University, Tehran, Iran.
dhaseli@khu.ac.ir
ORCID iD: https://orcid.org/0000-0002-1406-211X

Hashem Atapour
Associate Prof., Department of Knowledge and Information Science, University of Tabriz, Tabriz, Iran.
hashematapour@tabrizu.ac.ir
ORCID iD: https://orcid.org/0000-0003-0763-8413

Fatima FahimNia
Associate Prof., Department of Knowledge and Information Science, University of Tehran, Tehran, Iran.
Corresponding Author: fahimnia@ut.ac.ir
ORCID iD: https://orcid.org/0000-0002-8991-5391

Nader Naghshineh
Associate Prof., Department of Knowledge and Information Science, University of Tehran, Tehran, Iran.
nnaghsh@ut.ac.ir
ORCID iD: https://orcid.org/0000-0002-0509-5454

Molouk Sadat Hosseini Beheshti
Associate Prof., Information Science Research Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran, Iran.
beheshti@irandoc.ac.ir
ORCID iD: https://orcid.org/0000-0003-0059-6927

Mohammad Sadegh Zahedi
Ph.D. Student, Department of Electrical and Computer Engineering, University of Tehran, Tehran, Iran.
sadeghzahedi@ut.ac.ir
ORCID iD: https://orcid.org/0000-0001-5022-8423

Received: 11 December 2021
Accepted: 28 September 2022

Abstract

Many scientific texts are produced in Persian and available in scientific information databases through the Web. In this paper, FarsAcademic, a test collection of Persian scientific texts, has been built to implement information retrieval models among academic search comprising 102238 documents and 61 topics. While constructing FarsAcademic, we have tried to resolve the problems specific to information retrieval (IR) and natural language processing (NLP) in Persian scientific texts. Domain experts were employed to create queries within their research area, and user relevance and topical relevance were applied to improve the precision of relevance judgment of documents. Further, to improve retrieval performance in Persian scientific texts, automated query expansion was applied using one of the relevant feedback techniques, the Local Context Analysis algorithm. The result showed that query expansion techniques outperformed other information retrieval models in the Persian scientific texts retrieval task. Eventually, FarsAcademic became the only one that has been free of charge for Iranian information retrieval scholars to implement and evaluate different information retrieval models and algorithms on Persian scientific text and academic search.

Keywords: Information Search, Farsacademic, Information Retrieval, Persian Language, Scientific Texts, Test Collection, Retrieving Information Tool.

Introduction

Modeling the information-seeking behavior of academics and scientists (Ellis & Haugan, 1997; Nwone & Mutula, 2018) differs from modeling the information-seeking behaviors of the general public. The sheer number of academic fields further complicates this. These differences have impacted academic search (Li, Schijvenaars & de Rijke, 2017; Li & de Rijke, 2019). Academic search tasks are inherently complex, uncertain, and multifaceted (Du & Evans, 2011). It is unclear to what extent searchers use specified strategies when faced with different variety of information needs, ranging from simple tasks (such as fact-checking) to complex tasks (such as knowledge discovery) (Hoeber, Patel & Storie, 2019). Academic search has been widely examined across disciplines (Hsin, Cheng & Tsai, 2016).

In academic queries, scholars search for scientific texts. Academic search presents diverse challenges from the collections and information needs previously studied. The queries and their vocabulary can be highly technical and domain-specific. The type, volume, and structure of the documents searched, the content of the documents which contain technical terms or search for particular document components (such as author, title, or keywords), and the kind of expected relevance of the documents which is knowledge discovery level are but to name a few (Kluck, 2003; Mandl, 2008; Sanderson, 2010; Dietz & Petras, 2017). Scientific text differs from other types of text, such as news text (Heffernan & Teufel, 2018); the text is long (Pueyo & Redrado, 2003). The scientific paper is essentially the document outlining the process of solving problems (Popper, 1972), offering and testing hypotheses, and as a document, it becomes part of the objective public knowledge of specific science and it may be seen as continuing a long rhetorical tradition (Wang, Zhai, Lin & Wang, 2018). Therefore, the scientific text or research paper has specific characteristics which would be considered in designing searching, and retrieving information tools.

Information retrieval (IR) studies have largely focused on assessing the effectiveness of an information retrieval system. Evaluating information retrieval systems is crucial to progress in search technologies (Hutchins, 1977). Test collection construction is one of the prerequisites of information retrieval studies to evaluate information retrieval systems' performance. Constructing, using, evaluating, and sharing test collections have been major innovations in information retrieval (Carterette, 2007). While test collections can be expensive and complex to construct, and instantiate a particular abstraction of the information retrieval process, they have played an important role in providing a basis for the measurement and comparison of the effectiveness of different information retrieval algorithms and techniques (Scholer, Kelly & Carterette, 2016). Building a high-quality test collection is a time-consuming, costly process requiring human effort (Sanderson & Joho, 2004; Carterette & Bennett, 2008).

Scientific texts are of some consideration when constructing scientific test collection, the user’s experience in a real task and prior knowledge or motivation for searching (Voorhees, 2008) is important. Also knowledge discovery level of relevance judgments in scientific texts can be very costly. There are several scientific test collections in the English language, but there aren’t any in the Persian language. Academicians and scholars produce many Persian scientific texts available online through scientific information databases. However, no major study has been conducted on information retrieval of Persian scientific texts. One of the main reasons is the lack of a standard test collection in Persian scientific texts which is a prerequisite for studying theories or developing algorithms in this field.

Persian (also known to its native speakers in Iran as Farsi) is spoken in several countries

like Iran, Tajikistan, and Afghanistan. But historically, the language was widely understood throughout an area ranging from the Middle East to India. Persian is a Western Iranian language belonging to the Iranian branch of the Indo-Iranian subdivision of the Indo-European languages. Persian uses Arabic-derived script for writing and consists of 32 characters that are written continuously from right to left. During its long history, the language has been influenced by other languages such as Arabic, Turkish, and even European languages such as English and French (AleAhmad, Amiri, Darrudi, Rahgozar & Oroumchian, 2009). Persian has different prescribed forms of writing, differing in how words are written, using or eliminating spaces, and using various forms of characters (Sadeghi & Vegas, 2017). Persian text does not typically include diacritics and letters exhibit polymorphism in writing (Sadeghi & Vegas, 2014). There are more problems in tokenizing Persian texts than in other languages, such as English (Shafiee & Shamsfard, 2018). Persian words can be connected such as کمک (کـمک)(help) or a single word like درد (pain) or both of them as (سرما) سـرما (cold). Because of the Persian language's morphological specificities and structural characteristics, retrieval performance could be less influential on the Web (Sadeghi & Vegas, 2017). The Persian language has strategic importance for Iranians, the past two decades have seen a significant rise in the interest in Persian information retrieval and Persian natural language processing (NLP).

The studies seeking to identify the evaluations of information retrieval methods in Persian scientific databases shall not be applicable. As a result, Persian scientific information databases do not have a clear view of the methods of improving information retrieval performance. Accordingly, a standard test collection called "FarsAcademic" was generated from the Persian research articles in the present research. This test collection has been provided for the information retrieval researchers to progress in retrieving information from scientific texts by implementing their ideas and evaluating the consequences of their ideas. This standard test collection is specific to Persian scientific texts, which have not existed in the Persian language until now. The test collections of non-Persian cannot be applied to information retrieval in the Persian language.

Research objective

Thus, briefly summarizing the contributions of this article, the objectives of this article are:

- To Create a new standard test collection of Persian scientific texts to deal as much as possible with the evaluation of information retrieval in Persian scientific texts;
- To increase the accuracy of relevance judgment by using subject experts to provide queries in association with their research content and relevance judgment of documents;
- To improve the retrieval process of Persian scientific texts by solving common problems of NLP and information retrieval specific to Persian scientific texts;
- To determine the impact of different information retrieval models vs query expansion approach in improving Persian scientific texts retrieval.

The organization of this paper is as follows: Section 2 is a review of related works, and a brief description of some studies related to negation in the context of scientific search and scientific test collection is presented. Section 3 introduces the FarsAcademic and describes some attributes of the documents, topics, pool creation, and relevance judgment of the documents to the queries. Section 4 presents the performance of some retrieval algorithms on the FarsAcademic scientific texts and finally, we will discuss and conclude this paper in

sections 4 and 5.

Literature Review

Creating scientific texts collection is different from conventional texts in many aspects. Keeping up with rapidly growing research fields, especially with multiple interdisciplinary sources, requires substantial effort for researchers, program managers, or venture capital investors (Dunne, Shneiderman, Gove, Klavans & Dorr, 2012). Several test collections in different languages contain scientific texts and are domain-specific, the most significant of which is Cranfield 2 in English. The Cranfield 2 tests (Cleverdon, Mills & Keen, 1966) introduced an alternative method for creating a test collection, specifically for scientific texts. Furthermore, TREC (Text Retrieval Conference) added scientific text documents to its collections in 2005 (Hersh, Bhupatiraju, Ross, Roberts, Cohen & Kraemer, 2006).

Dietz and Petras (2017) created an academic search test collection with over 100 topics tested involving variations in queries, documents, and system components, resulting in 327,600 data points.

By using scientific Articles, Some test collections were created for specific domains, e.g. plagiarism detection (PAN[1] 2010–2014) (Gipp, Meuschke & Breitinger, 2014); pseudo test collections (Berendsen, Tsagkias, De Rijke & Meij, 2012); structured scientific data (Domain Specific 2000–2008) (Ferro, 2014); structured scientific bibliographic collections (Fautsch, Dolamic & Savoy, 2008); and Research-paper recommender systems (Raamkumar, Foo & Pang, 2017). Also, scientific articles are needed to create test collection for some fields, e.g. TREC–CHEM 2009/2010, which contains both patent documents and scientific articles, all chemistry-related (Piroi, Lupu & Hanbury, 2012). Also, for other languages, such as TREC created German, scientific text documents. In the German language, GIRT (German Indexing and Retrieval Test Database) collection to retrieve scientific text from structured social science documents was created (Kluck, 2003). CLEF 2005‘s domain-specific retrieval track is a scientific test collection, too (Piroi, Lupu & Hanbury, 2012).

In the Persian language, test collections have been created by the Database Research Group of the University of Tehran to advance the Persian textual information retrieval studies and other areas related to information retrieval, such as indexing methods, and etymology and linguistic issues. Among the most important collections, one can refer to Bijankhan Corpus, *Hamshahri* collection, and dotIR collection. The first collection, named after Dr. Bijankhan, has been collected from daily news and general texts and is appropriate for natural language processing studies in Persian. The Second collection is the *Hamshahri* collection which is a corpus containing 318000 documents related to the years between 1996 and 2007-8 (1375 SH-1386 SH) and created by crawling the website of *Hamshahri* newspaper followed by several stages of preprocessing and tagging. All the documents in the Hamshahri collection have the tag "Cat", which indicates each document's category. The second edition of the *Hamshahri* corpus has been produced by the Database Research Group of the University of Tehran with the support of the Iran Telecommunications Research Center (AleAhmad, Zahedi, Rahgozar & Moshiri, 2016). This test collection is suitable for various purposes, such as IR, computational linguistics, and NLP. *Hamshahri* collection consists of 166,774 documents in Persian with two sets of topics used in CLEF Persian–English bilingual track: *Ham'08* with 50 topics 551–600 used in CLEF 2008 (Agirre, Di Nunzio, Ferro, Mandl & Peters, 2008) and *Ham'09* with 50 topics 601–650 used for evaluation in CLEF 2009 (Ferro & Peters, 2009). Most previous work

on English–Persian cross-language information retrieval (CLIR) used the Hamshahri test collection at CLEF-2008 task: Retrieval of Persian documents from topics in English (Hashemi & Shakery, 2014). Rahimi, Shakery, and King (2016) used the Hamshari2 collection with two sets of topics used in the CLEF Persian–English bilingual track (*Ham'08 and Ham'09*) for collection CLIR.

The dotIR collection contains 1 million Persian webpages created to prepare a ground truth for information retrieval studies in Persian (Derhami, Khodadadian, Ghasemzadeh & Bidoki, 2013).

Another test collection in Persian is the Iranian Byelaws Test Collection which contains 90 years of Iranian legal texts with a wide range of lengths. For example, a short-length law with one or more paragraphs and a long rule of Iran's Annual Budget has been considered. To examine all the information retrieval models, the articles of the collections of laws were fragments consisting of sections or subsections of the laws and form several paragraphs. This test collection comprises 177089 sections (Saboori, Bashiri & Oroumchian, 2008). Furthermore, some foreign test collections such as CLEF, Fire, and 4BBC News contain standard Persian documents and queries used for the tasks of information retrieval and NLP in the Persian language.

University of Tehran Persian–English Comparable Corpus (UTPECC) consists of news articles in Persian (*Hamshahri* collection1) and English (BBC articles). This comparable corpus consists of more than 5,700 document pairs from 2002 to 2006 (Hashemi, Shakery & Faili, 2010). Hashemi, & Shakery (2014) used UTPECC as a Persian–English comparable corpus, for CLIR.

The Test Collection of FarsAcademic

The FarsAcademic, created in the present research, includes 102238 documents, each containing title, abstract, and keywords indexed in the Scientific Information Database (SID) in all scientific domains. The number of queries is 61 and the pool consists of 14987 documents from which 2768 are relevant, 1730 are partially relevant, and 10489 are irrelevant. Subject specialists have made the judgment of the relevance of the documents to the queries. The performance evaluation of the information retrieval models has been carried out using five basic information retrieval models for the main query and expansions of the queries.

The following stages of building the FarsAcademic will be described in six sections. These include gathering the documents for the test collection, query building, implementation of the information retrieval models, pooling, judgment of the relevance of the documents to the queries, and measuring the performance of information retrieval models in building the test collection. Furthermore, the validity of the test collection's main components, including the pool's documents and the relevance judgments, will be discussed.

Materials and Methods

The Documents of the FarsAcademic

There are different methods for providing the documents of the test collection, including crawling the target web resources, systematically downloading documents from databases, and directly acquiring documents from their producers. For building the present test collection, the bulk of the documents have been acquired from their producers, and the remaining part has been obtained by systematically downloading the documents from the journals' websites. 102238 documents containing the title, abstract, and keywords of scientific articles were downloaded from SID[2]. These documents include the articles indexes in SID during 1389 SH-1395 SH (approx. 2010-2017 CE), which cover all the scientific areas. Figure 1 shows a sample document of the test collection in XML format.

<DOC>

<DOCNO>2</DOCNO>

<DOCHDR>2</DOCHDR>

<TITLE>آموزش شیمی سبز با طراحی و اجرای آزمایش‌های سبز در مبحث استوکیومتری شیمی متوسطه</TITLE>

<ABSTRACT> هدف از انجام این پژوهش آموزش شیمی سبز با طراحی و اجرای آزمایش‌های سبز در مبحث استوکیومتری شیمی متوسطه است. طرح آزمایشی از نوع پیش آزمون پس آزمون بوده است. انتخاب دبیرستان محل اجرای طرح و انتخاب کلاس ها و اختصاص آنها به گروه‌های آزمایش و کنترل تصادفی است و دانش‌آموزان دختر پایه سوم متوسطه شهر جعفرآباد می باشد. روش کار پژوهش در دو مرحله و به صورت عملی که در مرحله اول آزمایش چسب‌سازی جهت آموزش عملی مبتنی بر شیمی سبز توسط گروه آزمون و در مرحله دوم آزمایش صابونی شدن که با توجه به جنبه‌های نگرشی موضوع و مبحث استوکیومتری بسیاری از اهداف شیمی سبز را در بر میگرفت انجام شد. آزمایش صابونی شدن بدون در نظر گرفتن اهداف شیمی سبز برای گروه کنترل انجام شد. داده‌های مورد نیاز پژوهش برای حیطه شناختی به صورت سوالات چهار گزینه ای حیطه مهارتی به صورت چک لیست مشاهدات و حیطه نگرشی فراگیر ان ترسیم نقاشی باز در راستای شیمی سبز جمع آوری شده است. جامعه آماری پژوهش شامل گروه آزمون ۳۰نفر و گروه کنترل ۳۰ نفر فراگیران متوسطه دوره دوم میباشد. نتایج به دست آمده بیانگر آن است که از بین ۱۲ اصل شیمی سبز فراگیران اهداف مربوط به اصول ۱، ۲، ۳، ٤، ٦، ۷، ۱۱ و ۱۲ را یاد گرفته اند و مورد توجهشان قرارگرفته است. همچنین آموزش بر اساس اهداف شیمی سبز در تربیت دانش آموزانی مسئول پرورش خلاقیت تشویق و ایجاد انگیزه در آن ها برای مطالعه و پژوهش در زمینه شیمی سبز و جایگزین کردن روش‌های سبز به جای روش‌های سنتی در انجام واکنش ها نقش به سزایی دارد و فراگیران راه‌های کاهش مصرف انرژی به حداقل رساندن ضایعات و استفاده از کمترین مواد شیمیایی را بدون آفت کیفیت محصول یاد می گیرند. </ABSTRACT>

<KEYWORD>آموزش شیمی سبز طراحی آزمایش سبز واکنش سبز استوکیومتری ارزشیابی</KEYWORD>

</DOC>

Figure 1: A Sample Document of the Test Collection

Table 1 shows some of the significant statistics of the collection. The average number of words in the documents was 115, calculated by subtracting the total number of words from the unique words. The longest word length in these documents exists in a 20-letter word, which is most likely a multi-word phrase with semi-space or has been created by the adhesion of several words to each other.

Table 1
General Features of the Test Collection

No.	Feature	Count
1	The number of documents	102,238
2	The number of unique words	232,952
3	Total number of words	26,678,602
4	The longest word length	20 (letters)

Topic creation

Topic creation or building the queries is one of the important and difficult stages of the test

collection building process. In this research, a form was designed to collect the queries and then the students and graduates in information science were asked to raise three queries in the subjects they have sufficient knowledge. The reason for selecting this statistical population for building the queries was that Ph.D. students and graduates had conducted at least one serious research (master's or doctoral thesis) and more or less maintained their contact with the academic-research world. Employing domain experts (Clough & Sanderson, 2013) and gathering real user retrieval tasks, and providing interactive information retrieval (Borlund, 2003) have been indicated in certain research (Voorhees, 2008). These two reasons make these persons qualified to make relevance judgments. The reason for requesting three queries from the participants was that appropriate queries (in the sense that a sufficient number of documents exist in Persian) were examined and identified and the identification and selection of common queries to make their relevance judgments were performed by two persons. The participants were also asked to build queries that contain at least two words (queries for which at least 10 relevant documents be retrieved in SID) and avoid offering very general and eccentric queries which might lead to too many relevant or irrelevant results. TREC queries are of a special format, each one consists of an identifier, title, description, and narrative. Thus the individual participants were asked to build queries that have the following three sections:

- Title: In sum, this element represents the query itself. The query's title is similar to the text we submit to the retrieval system while searching for a particular item of information.
- Description: In this section, the intended information need is described, and is longer the title and shorter than the query's narrative.
- Narrative: This section expresses the user's information need and his/her purpose in developing the query. If the query includes a subsection, it should be mentioned in this section.

To further guide the respondents, the following three queries were designed and presented (see Table 2).

Table 2

Examples of the Queries Built

Example	Query's title	Query's description	Query's narrative
1	public policy-making	national policy-making in the public sector	information about national policy-making in the public sector, including problem-finding, agenda-setting, problem-solving, policy development, implementation, and evaluation
2	metadata standards	metadata standards in information organization	Information about metadata standards, their types, their applications, advantages, the number of their elements, characteristics of each standard, and about which information medium or system is specific to each standard
3	the performance of public libraries	The performance evaluation models of public libraries	Information about the necessity of performance measurement, the methods and models of performance evaluation,

Example	Query's title	Query's description	Query's narrative
			and the performance excellence models of public libraries

Queries on various topics/themes should be selected with varying characteristics to test information retrieval systems under multiple settings. Research has shown that some queries are more useful in evaluation than others (Mizzaro & Robertson, 2007; Clough & Sanderson, 2013).

Figure 2 presents two examples of queries of the test collection in XML format, stored in the TREC standard format.

<QUERY>

<ID>1</ID>

<TITLE>نظریه‌های علم اطلاعات و دانش‌شناسی</TITLE>

<DESCRIPTION>تأثیر نظریه‌های علم اطلاعات و دانش‌شناسی بر توسعه سیستم‌های بازیابی اطلاعات</DESCRIPTION>

<NARRATIVE> نظریه‌های مرتبط به علم اطلاعات و دانش‌شناسی و نحوه تأثیر آنها بر توسعه سیستم‌های بازیابی اطلاعات</NARRATIVE>

</QUERY>

<QUERY>

<ID>2</ID>

<TITLE>گسترش پرس‌وجو</TITLE>

<DESCRIPTION>روش‌های مختلف گسترش پرس‌وجو در زبان فارسی و انگلیسی</DESCRIPTION>

<NARRATIVE> روش‌های گسترش پرس‌وجو، نقش گسترش پرس‌وجو در بازیابی اطلاعات، معرفی رویکردهای موثر در گسترش پرس‌وجو به زبان انگلیسی و فارسی</NARRATIVE>

</QUERY>

Figure 2: Examples of the Queries of the Test Collection "Knowledge And Information Science Theories"

Different test collections might have different numbers of queries. TREC's and CLEF's tendency in recent years has been to use 50 queries in each test collection (Sakai, 2014). In practice, a minimum of 50 topics should be included in the test collection to ensure reliability and to control for user variability (Voorhees, 2008; Clough & Sanderson, 2013). While the number of queries of Cranfield 2 tests collection is 221 (Cleverdon, 1967) and the number of queries of MEDLARS's test collection equals 29 (Sanderson, 2010), for the present research 61 queries were selected from among 204 queries which 70 persons created. Some of these queries which lacked the characteristics aforementioned for selection, were rejected. In selecting the queries, judging the documents' relevance was also considered. At the first stage of relevance judgment, each query was selected for building the test collection. Furthermore, by analysing the queries and the participants' backgrounds, we tried to select queries that no participant other than their developer can make a judgment about their relevance.

Implementation of the Information Retrieval Models

At this stage, text analysis of documents is carried out. Then the queries are submitted to the test collection based on different information retrieval systems, and the results of each query

are merged and undergo a relevance judgment. Queries with too many or too few results are rejected (Sanderson, 2010) and the queries which have passed successfully the evaluation filter constitute the test collection's queries.

At this stage, before analyzing the text of documents, given the special writing issues of the Persian language, a preprocessing operation was performed on the queries and documents. This preprocessing was done using Microsoft Excel software as string replacement. In preprocessing the documents, the Persian letters "ک" and "ی" with two Persian and Arabic forms were uniform in the queries and documents. The spaces and semi-spaces were also made uniform in the two-word phrases of the queries and documents.

After preprocessing the documents, their text was analyzed. To define the index fields, the documents were converted into XML format. The query number, title, abstract, and keywords index fields were generated using the XML format. An open-source search engine was needed to be used to index and retrieve the documents of the corpus. The Terrier search engine was selected for implementing the information retrieval models in this research. This search engine, developed at the University of Glasgow, is suitable for performing retrieval experiments on large-scale collections of documents. In addition, Terrier has performed well in the TREC test collections [51]. To access the content of XML documents, it was necessary to process them. All the XML tags of the corpus's documents were omitted to do so. Then the information retrieval models were applied to the documents.

Usually, to ensure the maximum retrieval of the documents relevant to queries and their incorporation into the pool, different information retrieval models are used. In some studies, not only queries but also their descriptions are utilized. For example, in building the irBlog test collection, the query's title and description have been used with different information retrieval models (AleAhmad, Zahedi, Rahgozar & Moshiri, 2016).

In this research, the original queries and their expansions were used to retrieve the articles relevant to the queries from the collection of documents. Query expansion solves the problem of brevity and ambiguity of a query by adding relevant terms to it (Keyvanpour, Karimi Zandian & Abdolhosseini, 2018).

Searching for relevant documents that satisfy a user's information need is a critical activity. Query expansion has been a motivation for a long time for improving the retrieval efficiency of information retrieval systems (Shaker, Farhadpoor & Nazari, 2017). The query expansion may be done in three: manual, interactive, and automatic ways (Gavel & Andersson, 2014). The main Query Expansion techniques are knowledge-based, corpus-based, and relevance feedback (Atwan, Mohd, Rashaideh & Kanaan, 2016). Pseudo-relevance feedback (PRF) is a very effective query expansion approach, which reformulates queries by selecting expansion terms from top k pseudo-relevant documents (El Mahdaouy, El Alaoui & Gaussier, 2019). In PRF-based query expansion, the first 10 or 20 retrieved documents are considered relevant and used as a source for selecting the expansion terms (Bhatnagar & Pareek, 2014). Xu and Croft (2000) performed a series of studies comparing the Local context analysis (LCA) query expansion method with the standard PRF approach using TREC collections. The performance improvement by LCA over local feedback is due to a better metric for selecting expansion terms.

In this project query expansion is performed (Table 3), and ten terms were added from 20 pseudo-relevance feedback documents with the LCA approach.

Table 3
An Example of the Query Expansion Terms[3]

No.	Query	Query expansion terms
1	semantic search engine	pages, information, Web, ranking, users, algorithms, improvement, results, ontology
2	success factors of knowledge management	organization, implementation, system, manager, communication, customer, advantage, competitive, firm
3	scientific collaboration networks of researchers	network, knowledge, social, partnership, sciences, co-authorship[4], university, domain, centrality

Five basic information retrieval models, including PL2, Lemur TF-IDF, BM25 probabilistic model, Dirichlet language model (DLM), and Hiemstra language model (HLM) were applied to the two groups of original queries. They expanded queries in the test collection by using the Terrier search engine.

Pool creation

To construct a test collection, it is essential to judge the relevance of topics to every document. In large-scale environments, exhaustive judging of relevance becomes infeasible. Instead, only a pool of documents is judged for relevance (Losada, Parapar & Barreiro, 2017). Different methods for building this sub-collection exist, such as pooling, interactive searching and judging (ISJ), and relevance feedback (Sanderson & Joho, 2004). The pooling method suggested within the framework of the ideal test collection by Jones and Van Rijsbergen (1976) is the most commonly used method in which the documents relevant to each query are identified from among the document pool. Here, the queries built using several basic information retrieval models are submitted to the document pool, and the documents relevant to each topic are retrieved. Then the top-ranked results for each model (usually the first 50 or 100) are selected and placed in a pool. After omitting the repeated documents (which might be retrieved by several models), the relevance judgment is performed on them (Sanderson & Joho, 2004). Pooling is based on the assumption that using several basic information retrieval models leads to the retrieval of a significant percentage of the sources relevant to a topic so that it can be considered equivalent to all the pertinent documents of the pool for the topic. If documents that do not exist in the pool are retrieved in the new efforts to test the new algorithms, one can consider them irrelevant (Carterette, Gabrilovich, Josifovski & Metzler, 2010). The pooling method has become the basis for establishing well-known test collections such as TREC, CLEF, and LETOR. The implicit goal of pooling is to achieve a good test reliability and validity level while limiting the number and expense of manual relevance judgments (Bodoff, 2008). The success of the pooling method is necessary for building up reusable relevance judgments (Bailey, Craswell & Hawking, 2003).

After applying the five information retrieval models to the original and expanded queries, ten lists of information retrieval results were recalled for each query. The top hundred results were selected from each list, and after the exclusion of recurring items, the remaining results were incorporated into the pool. The resulting pool contains 14987 articles. To supplement the pool, there is a need to judge the relevance of documents to the queries.

Relevance Judgement

For each topic/query in the test collection, a set of relevance judgments must be constructed ascertaining which one of the pool's documents is relevant to each topic/query. The relevance judgment assigns the value of the relevance by the judge (assessor) over a while. The relevance judgment is based on the document title, abstract, and keywords (Wildemuth, Marchionini, Fu, Oh & Yang, 2019). The relevance judgment has different grades. The relevance grades are usually used in the form of binary measures (relevant/irrelevant) or the form of a continuum (scale) (Carterette & Bennett, 2008). In building the present test collection, a three-level scale was used for the relevance judgment. Ideally, the topic generator will also perform relevance judgments. (Bailey, Craswell, Soboroff, Thomas, de Vries & Yilmaz, 2008; Clough & Sanderson, 2013). We used user relevance and topical relevance that improved the accuracy of relevance judgment of documents. The people who created the queries also made a judgment regarding the relevance of the documents. Although human-based relevance judgment could be time-consuming, expensive, and resource intensive, it is highly accurate (Tabrizi, Shakery, Zamani & Tavallaei, 2018).

To judge the relevance of the pool's documents, a database was designed using the SQL Server 2016 and the documents were stored in the database. The database was then put on the Web by designing a user interface page layout. A personal profile with the user name and password was created for each of the 70 individuals who participated in the research. The pool's documents for each person's query(ies) were incorporated in his/her profile. Then the participants were requested by a letter containing the instructions for making relevance judgment, to judge the relevance of the documents retrieved. Then the users made their relevance judgment about the documents retrieved by logging into their profile.

To judge the relevance of the pool's documents, Saracevic's three-point scale including the levels "relevant", "partially relevant" and "non-relevant" was provided for the judges. One option (i.e. non-judgeability) was also offered for the cases in which making a judgement was

Figure 3: Screenshots of the Login Page of a Profile, the Results Display, and the Relevance judgment

impossible or difficult. Figure 3 shows a screenshot of a page relevant to a document on the abovementioned website.

Sixty-one queries underwent the relevance judgment. In the first stage, the relevance judgment of the documents for each query was carried out simultaneously by two persons. For

arranging the composition of queries and individuals, the common or partially common queries were selected (user relevance). For other queries, an attempt was made to select the persons with experience in scientific and research activities in a field as the second judges for the documents relevant to those in that field (topical relevance). Finally, each query was judged by two persons, considering the participants' expertise and their likeliness to respond to one or two. In some limited cases, three queries were allocated for the relevance judgment by the persons. After receiving the first stage relevance judgments from the participants, the documents with similar, contradictory, and un-reviewed relevance judgment results were identified. Upon examination of the participants' backgrounds, the final relevance judgment of the contradictory and un-reviewed results was delegated to a third person. By repeating the process of personal profile building, these documents were sent and underwent a relevance judgment. 14987 documents experienced a relevance judgment by two persons simultaneously (29974 relevance judgments).

Of these documents, the relevance judgment of 2085 documents was contradictory or un-reviewed and hence underwent a relevance judgment by the third person again. An arrangement was made to select queries for the third judges not present in their previous-stage queries. The website's personal profile, user name, and password above were redefined and resent. In total, 32059 relevance judgments were made for pooling.

After the relevance judgment stage finished, the number of relevant, partially relevant, and irrelevant documents for each query was made clear. Figure 4 shows the number of relevant, partially relevant, and irrelevant documents from the pool for 61 queries of the test collection. Among the 14987 documents retrieved, 2768 were relevant, 1730 were partially relevant, and 10489 were irrelevant. In the two-level scales, the partially relevant documents were considered relevant; in total, 4498 relevant documents exist in the pool.

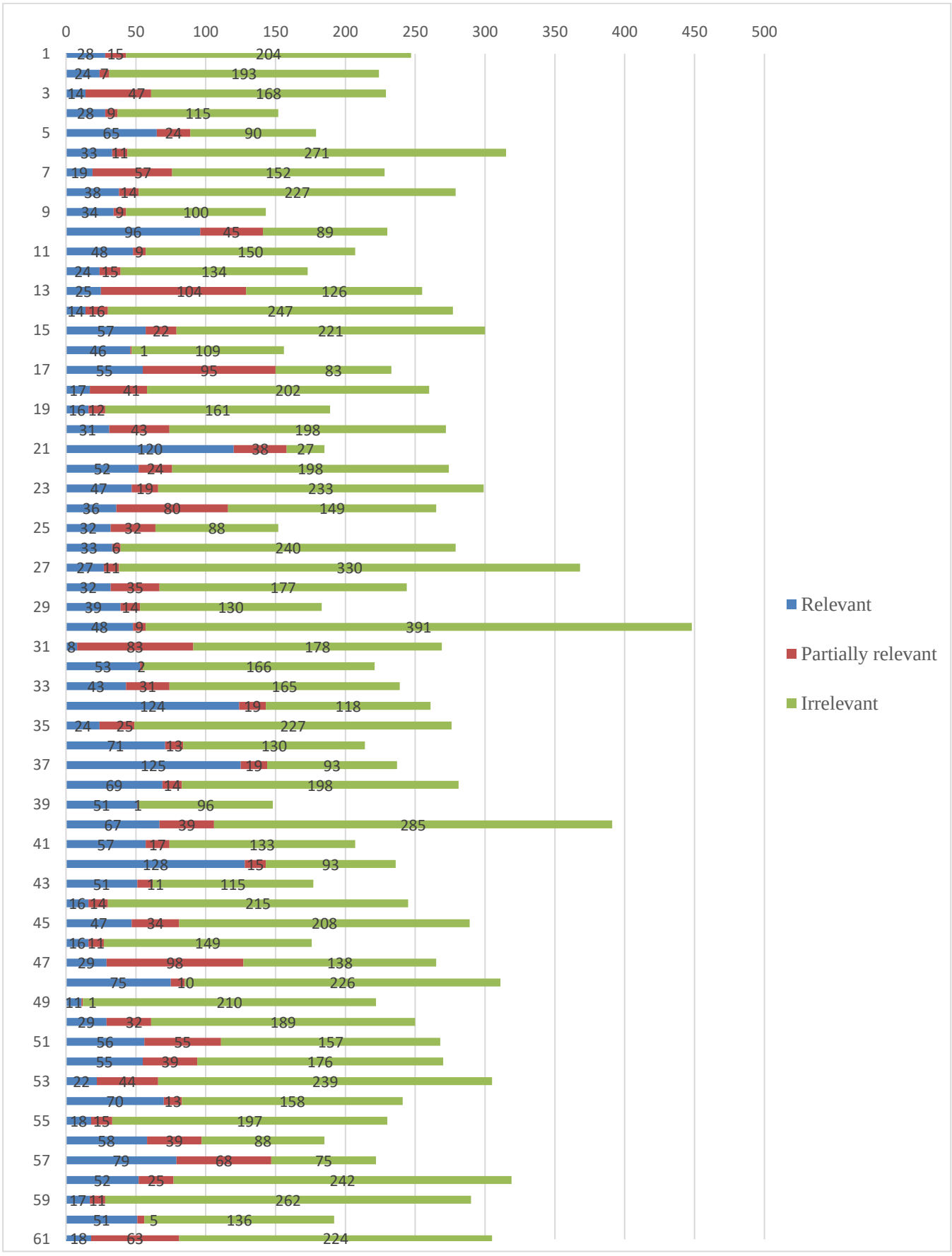

Figure 4: The Pool of the FarsAcademic

Results

Performance measurement results

After the relevance judgment of the documents was accomplished, the performance of each information retrieval model was measured by using the evaluation metrics of information retrieval systems. The evaluation metric used in this research are as follows: the relevant documents retrieved, precision at 5 (P@5), precision at 10 (P@10), mean average precision (MAP), mean reciprocal rank, and R-precision, which have binary (two-level) scales; and the normalized discounted cumulative gain (nDCG), which has a three-level scale.

To this end, the participants' relevance judgments were converted into text outputs and entered into the TREC-EVAL software. The evaluation metrics of the information retrieval systems were calculated using the TREC-EVAL software. Using EVAL requires two files: 1) a file containing the relevance judgments of the documents existing in the pool; and 2) a file containing the ranking of the information retrieval system under experimentation (National Institute of Science and Technology, 2017). Tables 4 and 5 represent examples of the information retrieval models' results and the relevance judgments carried out for documents. In the file of information retrieval models' results (see Table 4), the first and fourth columns are query id, the document's rank among the results retrieved, and the weight assigned to the document by the R system.

Table 4
An Example of the Information Retrieval Model's Results

Query id	Doc id	Rank	Sim
1	48729	15	10.6626
1	99980	10	11.0218
1	1642	11	10.9671
1	67754	93	8.3059

The first and third columns in the relevance judgment file (see Table 5) include the query id, doc id, and relevance judgment. In the relevance judgment column, digit 0 represents a partially relevant document, and digit 2 represents a relevant document. After comparing the two files mentioned earlier, the TREC-EVAL software calculates the metrics for the information retrieval system and produces its output.

Table 5
An Example of Relevance Judgment for Documents

Query id	Doc id	Relevance
1	48729	0
1	99980	2
1	1642	1
1	67754	0

Table 6 shows the evaluation metric for the information retrieval models' results. As we pointed out earlier in the "implementation of information retrieval models" section, five information retrieval models were applied to the original and expanded queries. Usually, to ensure a pool's validity and the presence of all the documents relevant to a query in that pool,

different information retrieval models are tried to be used. Here, we have used only five basic information retrieval modes. However, the significant difference found between the performance metrics of the original queries and the expanded queries indicates that query expansion has been more effective in the high performance of the documents' retrieval and thus the pool's validity than the basic information retrieval models themselves. For example, while the PL2 model retrieved 91 relevant documents more than the BM25 probabilistic model for the original query, the Dirichlet language model retrieved, 1020 relevant documents, in the expanded query, more than the original query. This comparison can also be seen in all other performance metrics.

Table 6

Evaluation Metrics of Information Retrieval Systems

Query	Model	The relevant documents retrieved	MAP	Rprec	MRR	P@5	P@10	nDCG
Original	BM25 probabilistic model	2456	0.2481	0.2901	0.6477	0.5115	0.4852	0.2881
	PL2 model	2547	0.2529	0.2962	0.7035	0.5541	0.5115	0.2992
	Dirichlet-LM	2516	0.2398	0.2829	0.6788	0.5344	0.4885	0.2992
	Hiemstra-LM	2529	0.2547	0.2999	0.7042	0.5738	0.5098	0.2973
	Lemur TF-IDF	2490	0.2416	0.2837	0.6463	0.5213	0.4951	0.2936
Expanded	BM25 probabilistic model	3493	0.407	0.4481	0.9342	0.7541	0.7066	0.3798
	PL2 model	3529	0.4263*	0.4656*	0.9404*	0.7902*	0.7164*	03839
	Dirichlet-LM	3536*	0.3887	0.4249	0.9068	0.7377	0.6902	0.3846*
	Hiemstra-LM	3525	0.4201	0.4582	0.9254	0.777	0.7066	0.3845
	Lemur TF-IDF	3477	0.3932	0.4283	0.9033	0.7279	0.6902	0.3797

* P<.001 for the paired t-test analysis for comparing the mean difference between two sets of data

Discussion

Shared test collections such as TREC are used in many international information retrieval research activities. These test collections are the main methodology for validating new retrieval approaches (Sanderson, 2010). Although there are several test collections in Persian for information retrieval, no test collection has been created for the Persian scientific texts thus far. Hence the present research began to build and validate a standard test collection based on the articles gathered from SID to accomplish as much as possible in the evaluation information retrieval tasks in the Persian scientific texts. To do this, six stages went through, including gathering the documents for the test collection, query building, implementation of the information retrieval models, pooling, judgment of the relevance of the documents to the queries, and the measurement of the performance of information retrieval models, and then the validity of the pool's documents and the relevance judgments were discussed.

The FarsAcademic includes 102238 documents, each containing title, abstract, and keywords indexed in the Scientific Information Database (SID) in all scientific domains. The number of queries is 61, and the pool consists of 14987 documents, from which 2768 are relevant, 1730 are partially relevant, and 10489 are irrelevant. Subject specialists have made the judgment of the relevance of the documents to the queries. The performance evaluation of the information retrieval models has been carried out using five basic information retrieval models for the main query and expansions of the queries.

To build this test collection, an attempt was made to resolve the problems specific to information retrieval and natural language processing (NLP) in Persian. Considering the special writing issues of the Persian language, a preprocessing operation was performed at the text analysis stage. This operation included resolving the special writing issues of the Persian language in both the query terms and documents.

One of the strengths of the FarsAcademic is receiving queries from real users. The queries in this test collection represent a set of topics that show the users' information needs and intentions, and each one includes a title, a description, and an adequate narrative for revealing the users' intentions. Furthermore, this collection contains 61 queries. Given that information retrieval studies need 30 to 50 queries, researchers can choose appropriate queries for retrieving information.

Another strong point of this test collection is that the relevance judgment has been made by the specialists who have generated the queries themselves (user relevance), those with experience in research activities in a field. The validity of the relevance judgment is one of the most important factors in the quality of a test collection. The relevance judgment of each query was made simultaneously by two specialists. When the first two persons' relevance judgment contradicted each other, it was delegated to a third person. The simultaneous relevance judgment of each query by its relevant specialists has provided a reliable test collection of the specialists' relevance judgments in the present collection.

One of the most important strong points of this test collection is the authority of the documents in its pool. At the document retrieval stage, using expanded queries led to an extraordinary improvement in information retrieval, enhancing the document retrieval performance, particularly the relevant retrieved documents metric, much more than the basic models. This ensured the presence of the maximum number of relevant articles for the queries in the pool and the validity of the collection. However, to use this collection in information retrieval studies, the results of the performance metrics of original queries (see Figure 4) will be the basis of comparison with future research results.

Compared to scientific text collections in English, the scientific collection developed in this study had better performance regarding information retrieval measures. The Cranfield 2 tests, despite having a high recall performance (60-80%), had 8-20 percent precision (Cleverdon, Mills & Keen, 1966), while the FarsAcademic showed a significant increase in precision percentage (Rprec=0.4656, MRR=0.9404, P@5=0.7902, andP@10=0.7164). The most important metric to evaluate information retrieval systems is the MAP, a benchmark for comparison of test collection performance. This measure was reported as 0.4075 for TREC Scientific text documents in Hersh, Bhupatiraju, Ross, Roberts, Cohen and Kraemer (2006). For test collection, GIRT4 in Dietz and Petras (2017) reported 0.1961 for the MAP. This metric was 0.4 for TREC–CHEM 2009/2010 (Piroi, Lupu & Hanbury, 2012). the GIRT section of TREC reported 0.3765 (Leveling, 2009), and CLEF 2005’s reported 0.2 for MAP, respectively

(Piroi, Lupu & Hanbury, 2012). The high value of the MAP metric and its significant growth in the FarsAcademic test collection (0.4656) can be attributed to the use of various information retrieval models in the test collection, especially the use of query expansion techniques.

As discussed in the literature review, no previous Persian scientific text collection exists. However, several Persian test collections can be labeled general test collections, as they do not focus on specific topics or bodies of documents. Performance measures of these test collections are as follows: Precision metrics were 0.3 in Hamshahri Corpus (AleAhmad, Amiri, Darrudi, Rahgozar, & Oroumchian, 2009). This metric was below 0.3 in irBlog (AleAhmad, Zahedi, Rahgozar & Moshiri, 2016). University of Tehran Persian–English Comparable Corpus (UTPECC) reported 0.42 for MAP, 0.62 for P@5, and 0.596 for P@10 (Hashemi, Shakery & Faili, 2010). In the FarsAcademic test collection, MAP was 0.4263, P@5 was 0.7902, and P@10 was 0.7164 (table 6). These all show the outperformance of the FarsAcademic test collection compared to previously developed Persian test collections.

Conclusion

The FarsAcademic was built to fill the information retrieval gap in the Persian scientific text collections. This test collection is a tool for implementing and testing the new models and algorithms of information retrieval and measuring their performance in the Persian scientific texts, and using it eliminates the information retrieval researchers' need to build queries as well as the user's need to make relevance judgment about the queries. Moreover, this test collection provides information retrieval researchers with an example of the performance of different metrics of information retrieval models for comparison.

Acknowledgments

We would like to thank the Scientific Information Database (SID) for providing the documents.

Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Endnotes

1. plagiarism, authorship attribution, and social software misuse
2. Scientific Information Database (or SID) is the Iranian database for the calculation of Persian and English articles citation.
3. Although all the terms in the collection are in Persian, for this research which is in English, they have been translated into English too.
4. It should be noted there are two equivalents in Persian for this term, which both exist in the collection and are as follows: 1. هم‌نویسندگی 2. هم‌تألیفی

References

Agirre, E., Di Nunzio, G. M., Ferro, N., Mandl, T. & Peters, C. (2008). CLEF 2008: Ad hoc track overview. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, 15-37. Springer, Berlin, Heidelberg. http://doi.org/10.1007/978-3-642-04447-2_2

AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M. & Oroumchian, F. (2009). Hamshahri: A standard Persian text collection. *Knowledge-Based Systems*, 22(5), 382-387. https://doi.org/10.1016/j.knosys.2009.05.002

AleAhmad, A., Zahedi, M., Rahgozar, M. & Moshiri, B. (2016). irBlogs: A standard collection for studying Persian bloggers. *Computers in Human Behavior*, 57, 195-207. https://doi.org/10.1016/j.chb.2015.11.038

Atwan, J., Mohd, M., Rashaideh, H. & Kanaan, G. (2016). Semantically enhanced pseudo relevance feedback for Arabic information retrieval. *Journal of Information Science*, 42(2), 246-260. https://doi.org/10.1177/0165551515594722

Bailey, P., Craswell, N. & Hawking, D. (2003). Engineering a multi-purpose test collection for web retrieval experiments. *Information Processing & Management*, 39(6), 853-871. http://doi.org/10.1016/S0306-4573(02)00084-5

Bailey, P., Craswell, N., Soboroff, I., Thomas, P., de Vries, A. P. & Yilmaz, E. (2008). Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 667-674. https://doi.org/10.1145/1390334.1390447

Berendsen, R., Tsagkias, M., De Rijke, M. & Meij, E. (2012). Generating pseudo test collections for learning to rank scientific articles. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 42-53. Springer, Berlin, Heidelberg. Retrieved from https://staff.fnwi.uva.nl/m.derijke/wp-content/papercite-data/pdf/berendsen-generating-2012.pdf

Bhatnagar, P. & Pareek, N. (2014). Improving pseudo relevance feedback based query expansion using genetic fuzzy approach and semantic similarity notion. *Journal of Information Science*, 40(4), 523-537. https://doi.org/10.1177/0165551514533771

Bodoff, D. (2008). Test theory for evaluating reliability of information retrieval test collections. *Information Processing & Management*, 44(3), 1117-1145. https://doi.org/10.1016/j.ipm.2007.11.006

Borlund, P. (2003). The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3), 8-23. Retrieved from http://informationr.net/ir/8-3/paper152.html

Carterette, B. (2007). Robust test collections for retrieval evaluation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 55-62. https://doi.org/10.1145/1277741.1277754

Carterette, B. & Bennett, P. N. (2008). Evaluation measures for preference judgments. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 685-686. https://doi.org/10.1145/1390334.1390451

Carterette, B., Gabrilovich, E., Josifovski, V. & Metzler, D. (2010). Measuring the reusability of test collections. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM)*, 231-240. http://doi.org/10.1145/1718487.1718516

Cleverdon, C. (1967). The Cranfield tests on index language devices. *Aslib Proceedings*. 19(6), 173-194. https://doi.org/10.1108/eb050097

Cleverdon, C. W., Mills, J. & Keen, M. (1966). Factors determining the performance of indexing systems. *Aslib Cranfield Research Project*, 1, 9-18. Retrieved from https://dspace.lib.cranfield.ac.uk/bitstream/handle/1826/861/1966%20ASLIB%20part1.pdf?sequence=2

Clough, P. & Sanderson, M. (2013). Evaluating the performance of information retrieval systems using test collections. *Information Research*. 18(2), 18-32. Retrieved from https://www.informationr.net/ir/18-2/paper582.html#.U2unLPldXTp

Derhami, V., Khodadadian, E., Ghasemzadeh, M. & Zareh Bidoki, A. M. (2013). Applying reinforcement learning for web pages ranking algorithms. *Applied Soft Computing*, 13(4), 1686-1692. https://doi.org/10.1016/j.asoc.2012.12.023

Dietz, F. & Petras, V. (2017). A component-level analysis of an academic search test collection. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 16-28. Springer International Publishing AG. https://doi.org/10.1007/978-3-319-65813-1_3

Du, J. T. & Evans, N. (2011). Academic users' information searching on research topics: Characteristics of research tasks and search strategies. *The Journal of Academic Librarianship*, 37(4), 299-306. https://doi.org/10.1016/j.acalib.2011.04.003

Dunne, C., Shneiderman, B., Gove, R., Klavans, J. L. & Dorr, B. (2012). Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12), 2351-2369. Retrieved from http://researchgate.net/publication/216017171

El Mahdaouy, A., El Alaoui, S. O. & Gaussier, E. (2019). Word-embedding-based pseudo-relevance feedback for Arabic information retrieval. *Journal of Information Science*, 45(4), 429-442. https://doi.org/10.1177/0165551518792210

Ellis, D. & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of Documentation*, 53(4), 384-403. https://doi.org/10.1108/EUM0000000007204

Fautsch, C., Dolamic, L. & Savoy, J. (2008). UniNE at Domain-Specific IR-CLEF 2008: Scientific Data Retrieval: Various query expansion approaches. In *Proceedings of the 9th Cross-Language Evaluation Forum Conference on Evaluating Systems for Multilingual and Multimodal Information Access*, 199-202. Springer, Berlin, Heidelberg. Retrieved from https://www.researchgate.net/publication/237697000_UniNE_at_Domain-Specific_IR_-_CLEF_2008_Scientific_Data_Retrieval_Various_Query_Expansion_Approaches

Ferro, N. (2014). CLEF 15th birthday: Past, present, and future. In *ACM SIGIR Forum*, 48(2), 31-55. Publication History. https://doi.org/10.1145/2701583.2701587

Ferro, N. & Peters, C. (2009). CLEF 2009 ad hoc track overview: TEL and Persian tasks. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, 13-35. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15754-7_2

Gavel, Y. & Andersson, P. O. (2014). Multilingual query expansion in the SveMed+ bibliographic database: A case study. *Journal of Information Science*, 40(3), 269-280. https://doi.org/10.1177/0165551514524685

Gipp, B., Meuschke, N. & Breitinger, C. (2014). Citation-based plagiarism detection: Practicability on a large-scale scientific corpus. *Journal of the Association for Information Science and Technology*, 65(8), 1527-1540. https://doi.org/10.1002/asi.23228

Hashemi, H. B. & Shakery, A. (2014). Mining a Persian–English comparable corpus for cross-language information retrieval. *Information Processing & Management*, 50(2), 384-398. https://doi.org/10.1016/j.ipm.2013.10.002

Hashemi, H. B., Shakery, A. & Faili, H. (2010). Creating a Persian-English comparable corpus. In *Multilingual and Multimodal Information Access Evaluation. CLEF 2010. Lecture Notes in Computer Science*, 6360. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-15998-5_5

Heffernan, K. & Teufel, S. (2018). Identifying problems and solutions in scientific text. *Scientometrics*, 116(2), 1367-1382. https://doi.org/10.1007/s11192-018-2718-6

Hersh, W. R., Bhupatiraju, R. T., Ross, L., Roberts, P., Cohen, A. M. & Kraemer, D. F. (2006). Enhancing access to the Bibliome: the TREC 2004 Genomics Track. *Journal of Biomedical Discovery and Collaboration*, 1, 3. http://doi.org/ 10.1186/1747-5333-1-3

Hoeber, O., Patel, D. & Storie, D. (2019). A study of academic search scenarios and information seeking behaviour. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, (pp. 231-235). https://doi.org/10.1145/3295750.3298943

Hsin, C. T., Cheng, Y. H. & Tsai, C. C. (2016). Searching and sourcing online academic literature: Comparisons of doctoral students and junior faculty in education. *Online Information Review*, 40(7): 979-997. https://doi.org/10.1108/OIR-11-2015-0354

Hutchins, J. (1977). On the structure of scientific texts. *UEA Papers in Linguistics*, 5(3), 18-39. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.578.1532&rep=rep1&type=pdf

Jones, K. S. & Van Rijsbergen, C. J. (1976). Information retrieval test collections. *Journal of Documentation*, 32(1), 59-75. https://doi.org/10.1108/eb026616

Keyvanpour, M. R., Karimi Zandian, Z. & Abdolhosseini, Z. (2018). A useful framework for identification and analysis of different query expansion approaches based on the candidate expansion terms extraction methods. *International Journal of Information Science and Management (IJISM)*, 16(2), 19-42. Retrieved from https://ijism.ricest.ac.ir/article_698277.html

Kluck, M. (2003). The GIRT Data in the Evaluation of CLIR Systems–from 1997 until 2003. *In Peters, C., Gonzalo, J., Braschler, M., Kluck, M. (eds) Comparative Evaluation of Multilingual Information Access Systems. CLEF 2003. Lecture Notes in Computer Science*, 3237. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-30222-3_37

Leveling, J. (2009). A comparison of sub-word indexing methods for information retrieval. In *Proceedings of the Lernen-Wissen Workshop week (LWA'09).* Retrieved from https://www.researchgate.net/publication/228722635_A_comparison_of_sub-word_indexing_methods_for_information_retrieval

Li, X. & de Rijke, M. (2019). Characterizing and predicting downloads in academic search. *Information Processing & Management*, 56(3), 394-407. https://doi.org/10.1016/j.ipm.2018.10.019

Li, X., Schijvenaars, B. J. A. & de Rijke, M. (2017). Investigating queries and search failures in academic search. *Information Processing & Management*, 53(3), 666-683. https://doi.org/10.1016/j.ipm.2017.01.005

Losada, D. E., Parapar, J. & Barreiro, A. (2017). Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management*, 53(5), 1005-1025. https://doi.org/10.1016/j.ipm.2017.04.005

Mandl, T. (2008). Recent developments in the evaluation of information retrieval systems: Moving towards diversity and practical relevance. *Informatica*, 32(1), 27-38. Retrieved from https://www.researchgate.net/publication/220166136_Recent_Developments_in_the_Evaluation_of_Information_Retrieval_Systems_Moving_Towards_Diversity_and_Practical_Relevance

Mizzaro, S. & Robertson, S. (2007). Hits hits trec: exploring information retrieval evaluation results with network analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 479-486. https://doi.org/10.1145/1277741.1277824

National Institute of Science and Technology. (2017). Retrieved from https://www.nature.com/nature-index/annual-tables/2017

Nwone, S. A. & Mutula, S. (2018). Information seeking behaviour of the professoriate in selected federal universities in southwest Nigeria. *South African Journal of Libraries and Information Science*, 84(1), 20-34. https://doi.org/10.7553/84-1-1733

Piroi, F., Lupu, M. & Hanbury, A. (2012). Effects of language and topic size in patent IR: an empirical study. In *Information Access Evaluation. Multilinguality, Multimodality, and Visual Analytics. CLEF 2012. Lecture Notes in Computer Science*, 7488(54-66). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33247-0_7

Popper, K. R. (1972). *Objective knowledge: An evolutionary approach*. New York: Oxford University Press. Retrieved from http://www.math.chalmers.se/~ulfp/Review/objective.pdf

Pueyo, I. G. & Redrado, A. (2003). A functional-pragmatic approach to the analysis of internet scientific articles. *LSP and Professional Communication*, 3(1), 43-59. Retrieved from https://rauli.cbs.dk/index.php/LSP/article/view/1982

Raamkumar, A. S., Foo, S. & Pang, N. (2017). Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. *Information Processing & Management*, 53(3), 577-594. https://doi.org/10.1016/j.ipm.2016.12.006

Rahimi, R., Shakery, A. & King, I. (2016). Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework. *Information Processing & Management*, 52(2), 299-318. https://doi.org/10.1016/j.ipm.2015.08.001

Saboori, F., Bashiri, H. & Oroumchian, F. (2008). Assessment of query reweighing, by Rocchio method in Farsi information retrieval. *Journal of Information Science and Technology*, 6(1), 9-16. https://ro.uow.edu.au/dubaipapers/58/

Sadeghi, M. & Vegas, J. (2014). Automatic identification of light stop words for Persian information retrieval systems. *Journal of Information Science*, 40(4), 476-487. https://doi.org/10.1177/0165551514530655

Sadeghi, M. & Vegas, J. (2017). How well does Google work with Persian documents?. *Journal of Information Science*, 43(3), 316-327. https://doi.org/10.1177/0165551516640437

Sakai, T. (2014). Designing test collections for comparing many systems. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, (pp. 61-70). https://doi.org/10.1145/2661829.2661893

Sanderson, M. (2010). *Test collection based evaluation of information retrieval systems*. Now Publishers Inc. http://dx.doi.org/10.1561/1500000009

Sanderson, M. & Joho, H. (2004). Forming test collections with no system pooling. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 33-40). https://doi.org/10.1145/1008992.1009001

Scholer, F., Kelly, D. & Carterette, B. (2016). Information retrieval evaluation using test collections. *Information retrieval evaluation using test collections*, 19(3), 225-229. https://doi.org/10.1007/s10791-016-9281-7

Shafiee, F. & Shamsfard, M. (2018). Similarity versus relatedness: A novel approach in extractive Persian document summarisation. *Journal of Information Science*, 44(3), 314-330. https://doi.org/10.1177/0165551517693537

Shaker, H., Farhadpoor, M. R. & Nazari, F. (2017). Effect of Expansion and Reformulation of Query on Improved Precision of Retrieval Results. *International Journal of Information Science and Management*, 15(2), 123-134. Retrieved from https://www.researchgate.net/publication/318262524_Effect_of_Expansion_and_Reformulation_of_Query_on_Improved_Precision_of_Retrieval_Results

Tabrizi, S. A., Shakery, A., Zamani, H. & Tavallaei, M. A. (2018). PERSON: Personalized information retrieval evaluation based on citation networks. *Information Processing & Management*, 54(4), 630-656. https://doi.org/10.1016/j.ipm.2018.04.004

Voorhees, E. M. (2008). On test collections for adaptive information retrieval. *Information Processing & Management*, 44(6), 1879-1885. https://doi.org/10.1016/j.ipm.2007.12.011

Wang, X., Zhai, Y., Lin, Y. & Wang, F. (2019). Mining layered technological information in scientific papers: A semi-supervised method. *Journal of Information Science*, 45(6), 779-793. https://doi.org/10.1177/0165551518816941

Wildemuth, B. M., Marchionini, G., Fu, X., Oh, J. S. & Yang, M. (2019). The usefulness of multimedia surrogates for making relevance judgments about digital video objects. *Information Processing & Management*, 56(6), 102-109. https://doi.org/10.1016/j.ipm.2019.102091

Xu, J. & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems (TOIS)*, 18(1), 79-112. https://doi.org/10.1145/333135.333138