2009

Hamshahri: A standard Persian Text Collection

Abolfazl Aleahmad
University of Tehran, Iran

Hadi Amiri
University of Tehran

Masoud Rahgozar
University of Tehran

Farhad Oroumchian
University of Wollongong in Dubai, farhado@uow.edu.au

Publication Details

Aleahmad, A., Amiri, H., Rahgozar, M. & Oroumchian, F. 2009, Hamshahri: A standard Persian Text Collection, Knowledge-Based Systems, vol. 22, no. 5, pp. 382-387.

Hamshahri: A Standard Persian Text Collection

Abolfazl AleAhmad [a], Hadi Amiri [a], Masoud Rahgozar [a], Farhad Oroumchian [a,b]

[a] Electrical and Computer Engineering Department, University of Tehran
[b] University of Wollongong in Dubai
{a.aleahmad, h.amiri}@ece.ut.ac.ir, rahgozar@ut.ac.ir, farhadoroumchian@uowdubai.ac.ae,

Abstract. The Persian language is one of the dominant languages in the Middle East, so there are significant amount of Persian documents available on the Web. Due to the special and different nature of the Persian language compared to other languages like English, the design of information retrieval systems in Persian requires special considerations. However, there are relatively few studies on retrieval of Persian documents in the literature and one of the main reasons is lack of a standard test collection. In this paper we introduce a standard Persian text collection, named Hamshahri, which is built from a large number of newspaper articles according to TREC specifications. Furthermore, statistical information about documents, queries and their relevance judgment are presented in this paper. We believe that this collection is the largest Persian text collection, so far.

Keywords: Persian test collection, Farsi text retrieval.

1 Introduction

The Persian language (also know as Farsi) is one of the dominant languages in the Middle East that is spoken in several countries like Iran, Tajikistan and Afghanistan. Persian uses Arabic like 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. Today's Persian contains many words from the above languages and in some cases these words still follow the grammar of their original languages in building plural, singular or different verb forms. Therefore, the morphological analyzers for this language need to deal with many forms of words that are not actually Persian [15]. So, because of the special and different nature of the Persian language compared to other languages like English, the design of information retrieval systems in Persian requires special considerations [14].

In spite of special characteristics of the Persian language, little efforts have been focused on retrieval of Persian text compared to other languages. As existence of a standard test collection is a prerequisite for research in information retrieval, creation of a standard Persian text collection is essential and gives more validity to Persian information retrieval researches. Also, there are some works in the literature that have been focused on creation of Persian collections. Authors of [20] used a 25 MB text

collection that contains laws and regulations passed by Iranian Parliament. As they cited, the results may be biased because the size of the collection is small and it covers just one subject category. Shiraz corpus [10] is a 10 MB bilingual tagged corpus developed from a Persian corpus of on-line material to test machine translation project at New Mexico State University. FLDB corpus [19] comprises a selection of contemporary Modern Farsi literature, formal and informal spoken varieties of the language, and a series of dictionary entries and word lists (about 3 millions). The comprehensiveness of FLDB presents it as a well-structured modern Farsi corpus. However, its size isn't good enough for extensive information retrieval researches. Another Persian corpus is Mahak [12] that is prepared for evaluation of information retrieval systems. Also, this corpus contains 3007 documents that make it unsuitable for large scale information retrieval systems.

In this paper we will investigate creation of Hamshahri collection. Hamshahri is one of the first online Persian newspapers in Iran that has been published for more than 20 years and it has presented its archive to the public through its website [9] since 1996. Creation of Hamshahri collection starts from [7] in which Oroumchian et al employed a crawler to download available online news from the web site of Hamshahri newspaper and presented some advantageous statistics of Hamshahri corpus based on characteristics of the Persian language. After that some researches have been conducted based on this collection.

Compression of Persian text for web was assessed on Hamshahri collection in [6]. The authors in [2] proposed the design and testing of a Fuzzy retrieval system for Persian (FuFaIR) with support of Fuzzy quantifiers in its query language. Also experiments in [8] on Hamshahri collection suggest the usefulness of language modeling techniques for Persian. Furthermore in [1] Aleahmad et al evaluated vector space model on Persian text with different weighting schemes and show that 4-gram vector space model using Lnu-ltu weighting with slope 0.25 produces good results.

All these researches were conducted on previous version of the collection and leads to completion and creation of the current version of Hamshahri collection. Previous version of the collection contains 58 queries that were not prepared based on TREC specifications. In this work we follow TREC specifications to create a standard version of Hamshahri collection. TREC uses a technique called pooling [11, 13] in which a pool of subset of documents is created for each topic and is judged for relevance by the topic author. For this purpose 65 topics are added to the collection according to TREC specifications [17]. In addition, some useful statistical information of the new queries and the evaluation results based on them are presented in this paper. The current standardized version of Hamshahri collection contains more that 160,000 documents, 65 queries and their relevance judgments which are publicly available on the web[1].

Remaining parts of this paper are organized as follows: section 2 describes some attributes of the collection and its documents, section 3 details collection topics and their categories and section 4 describes pool creation and relevance judgments. Section 5 presents performance of a number of retrieval engines on Hamshahri collection and finally we will conclude this paper in section 6.

[1] Hamshahri collection: http://ece.ut.ac.ir/DBRG/hamshahri/

2. Collection Documents

Documents of Hamshahri collection are actually news articles of Hamshahri newspaper from year 1996 to 2002 that is written by many authors from a variety of backgrounds and covers a range of different topics, each with a credible size of data. It also consists of real text in everyday use of Persians that implies it has the sampling and representativeness property. The preparation process of the collection's documents is discussed in [7] in which Darrudi et al employed a crawler to download available online news from the newspaper's website. The original size of the downloaded HTML files was 820 MB. Then they built a consistent corpus of 345 MB from those raw HTML files using conversions like: Merging HTML files, removing HTML tags, aligning Farsi words, aligning numbers and punctuations and finally removing extra spaces between adjacent words.
Furthermore, we merged all files to one big file and tagged each article with an ID that we assigned to it, its publication date and category. Figure 1 depicts a sample document from this collection; the first line in each document contains a DID tag that shows the document's ID, the second line shows publication date of the document, the third line is its category and document's content proceeds the third line.

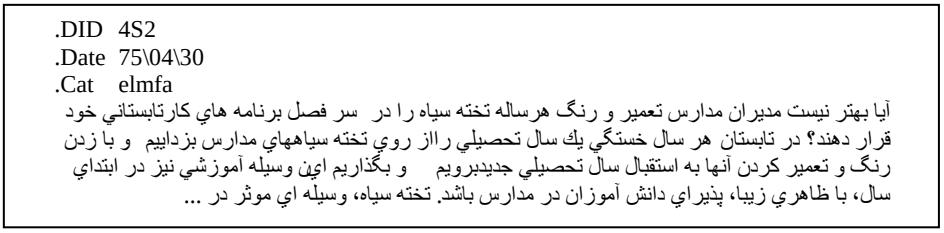
.DID 4S2
.Date 75\04\30
.Cat elmfa
آيا بهتر نيست مديران مدارس تعمير و رنگ هرساله تخته سياه را در سر فصل برنامه هاي كارتابستاني خود قرار دهند؟ در تابستان هر سال خستگي يك سال تحصيلي رااز روي تخته سياههاي مدارس بزداييم و با زدن رنگ و تعمير كردن آنها به استقبال سال تحصيلي جديدبرويم و بگذاريم اين وسيله آموزشي نيز در ابتداي سال، با ظاهري زيبا، پذيراي دانش آموزان در مدارس باشد. تخته سياه، وسيله اي موثر در ...

Figure 1. A sample document from Hamshahri collection

The size of the documents varies from short news (under 1 KB) to rather long articles (e.g. 140 KB) with the average of 1.8 KB. Figure 2 shows the size distribution of the documents and table 1 summarizes some attributes of Hamshahri collection.

Table 1. Attributes of Hamshahri collection

Attributes	Value
Collection size	564 MB
Documents Format	Text
No. Of documents	166,774
No. Of unique terms	417,339
Average length of documents	380 Terms
No. Of categories	82
No. Of Topics	65

There are just 2 documents in the collection whose length is more than 9000 terms (i.e. 30478, 11701) and are not shown in the figure to give it a better view. As it is mentioned in table 1 average length of documents is 380 terms.

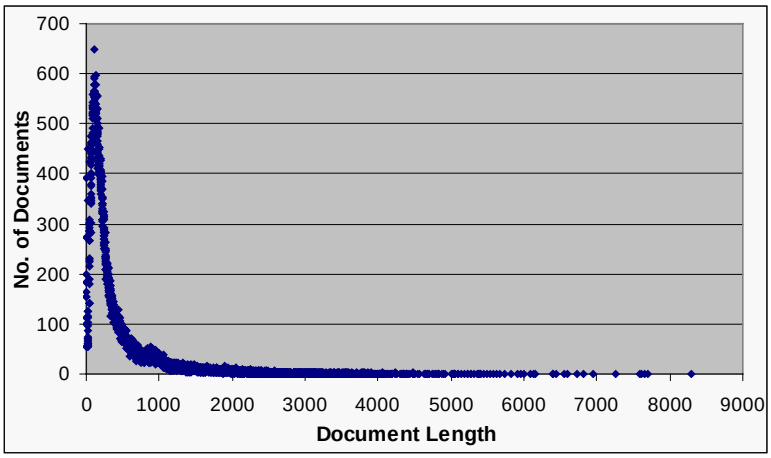

Figure 2. Document length distribution of Hamshahri collection

The authors of Hamshahri newspaper have manually categorized their articles in different categories and made them available at their site. Consequently, all documents in our collection are categorized in 82 different subjects by the crawler in [7] based on news categories that were available at the Hamshahri newspaper's site. So, each document in the collection is tagged with a word that indicates its category. As an instance, the tag "siasi" means "سیاسی" in Persian or "political" in English. Appendix B contains a list of main category tags used in the collection and their actual name in Persian and English.

Although Hamshahri collection has 82 categories but only 12 categories contain more than 1000 documents that cover nearly 72 percent of the collection. These categories and the number of documents in each category are depicted in figure 3. It should be mentioned that some of the categories are highly related to each other, so it's a good idea to merge some of them together and reduce the number of categories. But we leave this task on final users of this collection and remain the collection's categories intact.

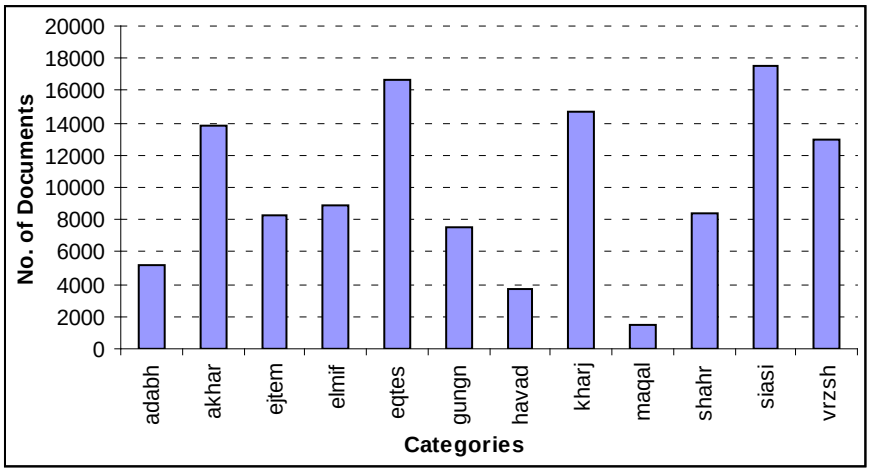

Figure 3. Distribution of documents in 12 main categories

3. Topics

There exist 65 different topics in this version of Hamshahri collection. In order to create the topics, we briefly discussed the collections content and asked 17 different users to issue 5 queries each and write a narrative and description for them. Then we omitted overlapped queries and chose 4 out of 5 queries that yielded 68 queries. Moreover, as described in section 4, 3 further queries were omitted that yields 65 queries. A sample topic of the collection is depicted in figure 4.

```
<QID>
        2
</QID>
<title>
        انرژي هسته قوانين
</title>
<description>
        قوانين جهاني حاكم بر استفاده از انرژي هسته اي كدامند.
</description>
<narrative>
        چه قوانيني براي استفاده از انرژي هسته اي وضع شده است و كدام كشور ها ناقض آن بوده اند و چه
        مجازات هايي براي آنها در نظر گرفته شده است.
</narrative>
```

Figure 4. Sample Topic (Topic No. 2 about "Nuclear energy regulations")

We assigned a unique identification number to each query that is characterized by QID tag. As it is shown in figure 4 each topic is presented by use of 4 tags: QID, title, description and narrative according to [5].

Average length of the 65 queries is 2.84, the minimum length is 2 and maximum length is 4. Length distribution of the queries is depicted in figure 5. It shows that, like other languages, in Persian most users use queries with 2 or 3 terms to express their information need.

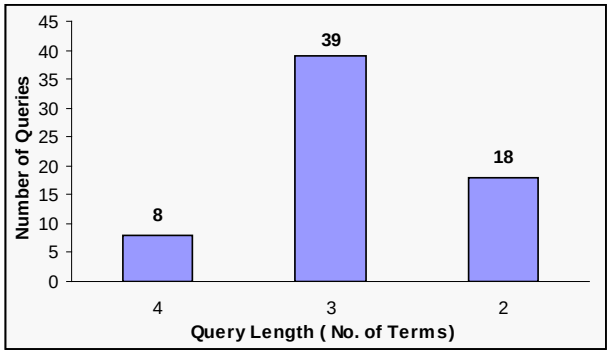

Figure 5. Length distribution of queries

In order to categorize the 65 queries into the 82 existing categories of the collection, we used maximum likelihood estimation. In other words, number of relevant retrieved documents in each category was calculated for each query and the category with the most relevant documents was considered as the corresponding query's category. All queries and their categories are presented in appendix A. Also, figure 6 shows distribution of queries over different categories (note that queries no. 5, 34, 57 and 58 are categorized into more than one category because of equal probability of being in two or three categories). As one can speculate, "eqtes" and "elmif" categories ("Economy" and "Since & Culture" respectively) encompass more queries than others.

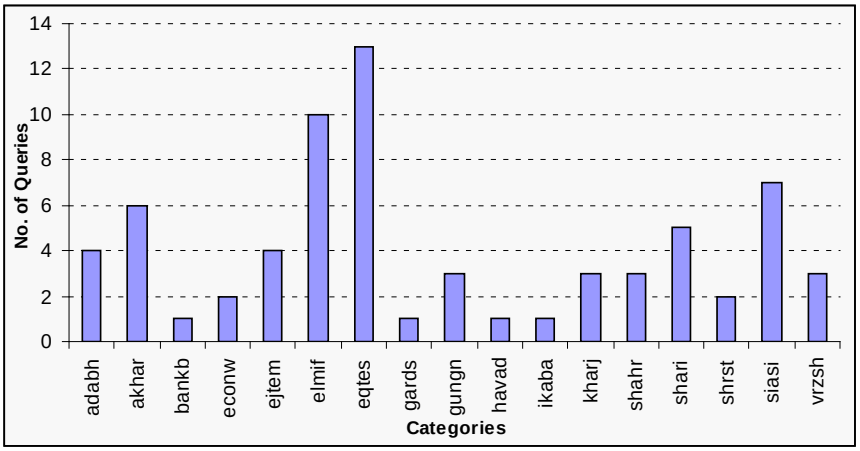

Figure 6. Distribution of queries over categories

4. Pool Creation and Relevance Judgment

The relevance judgments are what turn a set of documents and topics into a test collection. For this task, all the 68 initial queries were given to 7 different retrieval engines, namely LM1,LM2, LM3, LM4, VS2, VS4, VS5 that work based on vector space and language modeling [18] and are described in table 2. Then their results were combined to create a pool based on TREC specifications [17]. After that we created a simple interface and asked 17 different users that were IT students in our faculty to judge 100 top documents as either relevant or not relevant for each of their own query. Also, the assessors were told to assume that they are writing a report on the subject of the topic statement. If they would use any information contained in the document in the report, then the (entire) document should be marked relevant, otherwise it should be marked irrelevant. The assessors were instructed to judge a document as relevant regardless of the number of other documents that contain the same information as stated in [5].

After the judgment step, we omitted 3 queries from the collection that had less that 10 or more than 90 relevant documents. Totally 6500 relevance judgments exist in Hamshahri collection that includes 6183 unique documents. The number of relevant documents is 2352 that is 36.2 percent of relevance judgments. Figure 7 depicts number of relevant documents for each of the 65 queries.

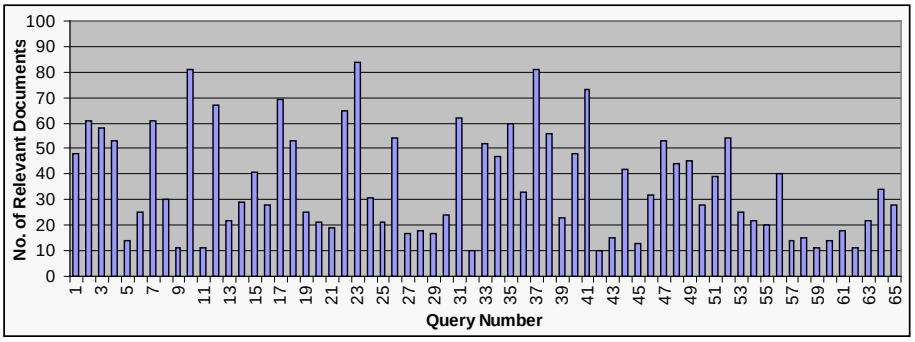

Figure 7. Number of relevant documents for each query

5. Experimental Results

In this part we will assess performance of different retrieval systems on Hamshahri collection. For this purpose we used all 65 available queries to assess performance of 8 retrieval systems that are listed in table 2.
In table 2 LM1 to 4 are equivalents of Score 1 to 4 presented in [4] that we ran on Persian text. For these language modeling experiments the λ parameter is considered 0.0485, the value that Taghva et al. determined as the optimal value of λ for Persian documents with no stemming and no stop word removal [14]. Besides, VS1 to 5 are vector space models and in VS4 and VS5 the Slope parameter in considered as 0.25 that was shown to have better performance over TREC collection [3] and Hamshahri collection [1].

Table 2. Retrieval systems that were assessed on Hamshahri collection

System	Description
LM1	Language model score 1 of [4] , Lambda=0.0485
LM2	Language model score 1 of [4] , Lambda=0.0485
LM3	Language model score 1 of [4] , Lambda=0.0485
LM4	Language model score 1 of [4] , Lambda=0.0485
VS1	4gram-based vector space model with atc.atc weighting scheme
VS2	Term-based vector space model with atc.atc weighting scheme
VS4	4-gram-based vector space model with Lnu.ltu weighting scheme (Slope = 0.25)
VS5	Term-based vector space model with Lnu.ltu weighting scheme (Slope = 0.25)

Moreover, the results of the retrieval systems are calculated by use of the trec_eval program [16] and are presented in table 3 and figure 8. It shows that our results are consistent with previous works. In [1] the authors mentioned that compared to other retrieval models, 4-gram-based vector space model have good performance in Persian as one can conclude from table 3 and figure 8. Furthermore, in [8] the authors show

that LM2 and 4 have higher precision than LM1 and LM3 and also LM3 has the worst precision in LM methods which is confirmed by our results.

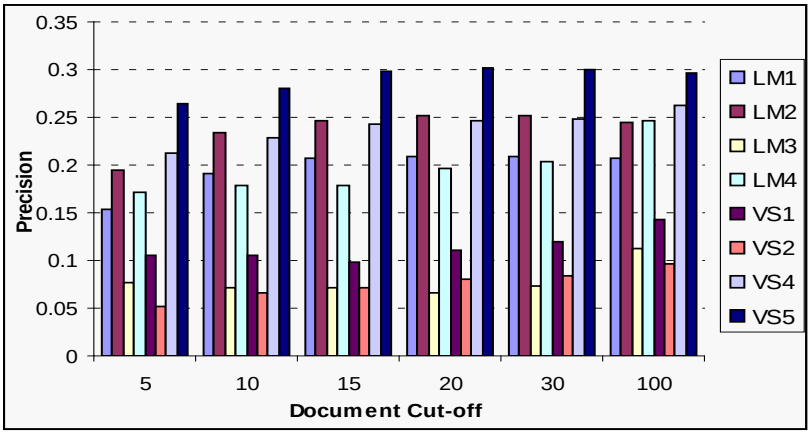

Figure 8. Precision at document cut-off of the 8 retrieval systems

Table 3. Precision-Recall of the 8 retrieval systems

Recall	LM1	LM2	LM3	LM4	VS1	VS2	VS4	VS5
0	0.3727	0.4078	0.2861	0.449	0.2898	0.1708	0.4658	0.5037
0.1	0.3108	0.3505	0.1597	0.3276	0.1917	0.1306	0.3477	0.4103
0.2	0.2818	0.3293	0.1282	0.2984	0.1616	0.1217	0.3302	0.3817
0.3	0.2505	0.3036	0.1106	0.2712	0.1433	0.1105	0.3067	0.3653
0.4	0.224	0.2844	0.0918	0.2646	0.1357	0.1004	0.2802	0.3489
0.5	0.1866	0.2606	0.0742	0.2437	0.1012	0.0934	0.2739	0.3314
0.6	0.1535	0.2169	0.0607	0.2194	0.0648	0.0767	0.2512	0.3139
0.7	0.1083	0.1866	0.0543	0.179	0.0572	0.0456	0.1935	0.2886
0.8	0.0695	0.1425	0.0485	0.1347	0.0377	0.0416	0.161	0.2533
0.9	0.018	0.0741	0.0212	0.0528	0.0218	0.0225	0.0604	0.1612
1	0.0073	0.0214	0.0162	0.0286	0.0159	0.0203	0.0191	0.0342

6. Conclusion

The lack of common large collections has been a major draw back in Farsi text retrieval experiments. It is very hard to generalize and compare the results of experiments conducted by different researchers in different universities and research centers unless there is a common collection that is widely adopted. This paper has described creation of a large scale text collection with known statistics for text retrieval and natural language processing purposes.

On the other hand, research projects need more topics to have a more accurate measurement so in future we are to add more queries to the collection. However there exist 58 additional topics with relevance judgment of 20 top retrieved documents for Hamshahri collection that we have used them in some of our previous experiments. But as they were not created according to TREC specifications we did not included them in this paper.

Hamshahri collection is downloadable as a package from its web site and can be used freely for noncommercial purposes. The package contains all relevance judgments for the 65 standard topics, some descriptions about previous researches conducted based on the collection and some source codes that we have implemented for indexing and retrieval on the collection. Also, the 58 older topics with their relevance judgment of top 20 retrieved documents are accessible through the site.

Moreover, as our experimental results showed, Lnu.ltu weighting scheme in vector space model produced the best results. So, we want to tune the slope parameter of Lnu.ltu weighting scheme on Hamshahri collection to improve retrieval results.

References

1. A. Aleahmad, P. Hakimian, F. Mahdikhani, F. Oroumchian: “N-Gram and Local Context Analysis for Persian Text Retrieval”, ISSPA 2007, Sharjah, UAE, 2007.
2. A. Nayyeri, F. Oroumchian: “FuFaIR: a Fuzzy Persian Information Retrieval System”. IEEE International Conference on Computer Systems and Applications, pp. 1126-1130, 2006.
3. A. Singhal, C. Buckley, M. Mitra: “Pivoted Document Length Normalization”, Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp21-29, 1996.
4. D. Hiemstra. Using Language Models for Information Retrieval. PhD thesis, University of Twente, 2001.
5. E.M. Voorhees, “Overview of TREC 2004”, Voorhees E. M. (2005). Overview of TREC 2004. IN: Voorhees, E., Buckland, L. (Eds.) Proceedings of the 13th Text Retrieval Conference, November 16-19, 2004, (TREC 2004). Gaithersburg, MD.
6. F. Oroumchian, E. Darrudi: Experiments with Persian Text Compression for Web, World Wide Web 2004, pp. 478-479, New York, New York, USA, May 2004.
7. F. Oroumchian, E. Darrudi, M.R. Hejazi: “Assessment of a modern Persian corpus”, Proceedings of The 2nd Workshop on Information Technology & its Disciplines (WITID), ITRC, Iran, 2004.
8. H. Amiri, A. AleAhmad, F. Oroumchian: C. Lucas, M. Rahgozar, "Using OWA Fuzzy Operator to Merge Retrieval System Results" in Computational Approaches to Arabic Script-based Languages (CAASL 2007), Stanford University, USA, 2007.
9. Hamshahri newspaper, http://www.hamshahri.net/

10. J.W. Amtrup, H. Mansouri Rad, K. Megerdoomian,R. Zajac, "Persian-English Machine Translation: An Overview of the Shiraz Project", NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-319), 2000.
11. J. Zobel: How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 307–314, Melbourne, Australia. ACM Press, New York, August 1998.
12. K. Sheykh Esmaili, H. Abolhassani, M. Neshati, E. Behrangi, A. Rostami, M. Mohammadi, Mahak: A Test Collection for Evaluation of Farsi Information Retrieval Systems, IEEE/ACS International Conference on Computer Systems and Applications, 2007.
13. K. Sparck Jones and C. van Rijsbergen: Report on the need for and provision of an "ideal" information retrieval test collection. British Library Research and Development Report 5266, Computer Laboratory, University of Cambridge, 1975.
14. K. Taghva, J. Coombs, R. Pareda, T. Nartker: "Language Model-Based Retrieval for Persian Documents". Internationa Conference on Information Technology: Coding and Computing (ITCC'04), 2004.
15. K. Taghva, R. Beckley, M. Sadeh: "A Stemming Algorithm for the Persian Language". International Conference on Information Technology: Coding and Computing (ITCC 2005), 2005.
16. National Institution of Standards and Technology, http://trec.nist.gov/trec_eval/
17. N. Craswell, D. Hawking, R. Wilkinson, and M.Wu: Overview of the TREC 2004 web track. In Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004), 2004.
18. Ponte, J.M., Croft, W.B.: A Language Modeling Approach to Information Retrieval. In ACM SIGIR (1998) 275-281, 1998.
19. S.M. Assi, "Farsi Linguistic Database (FLDB)", International Journal of Lexicography, Vol. 10, No. 3, EURALEX Newsletter p. 5, 1997.
20. Taghiyareh F., Darrudi E., Oroumchian F., Angoshtari N., "Compression of Persian Text for Web-Based Applications, Without Explicit Decompression", WSEAS Transactions on Computers, Issue 4, Vol. 2, pp.961-966, October 2003.

Appendix

A. Query categories

Table 4 contains the 65 queries and their categories. In queries column the first number is query ID and the number in parentheses is number of relevant retrieved documents in the corresponding category (e.g. 14(25) in the first row means query 14 has 25 relevant documents that categorized it into "adabh" category).

Table 4. Queries and their category

Categories	Queries

adabh	14(25), 43(7), 57(4), 59(7)
akhar	5(4), 28(8), 32(4), 55(5), 56(20), 58(3)
bankb	52(23)
econw	13(5), 18(22)
ejtem	8(9), 9(3), 48(21), 54(7)
elmif	19(11), 27(9), 29(13), 36(24), 42(6), 50(18), 57(4), 58(3), 61(13), 64(8)
eqtes	4(32), 12(54), 15(35), 17(34), 26(29), 33(18), 34(12), 35(33), 37(39), 41(59) ,46(21), 47(34), 62(3)
gards	49(11)
gungn	5(4), 51(23), 63(5)
havad	10 (32)
ikaba	11(3)
kharj	2(23), 16(17), 40(27)
shahr	5(4), 20(10), 21(10)
shari	1(11), 22(59), 25(5), 38(30), 45(4)
shrst	7(24), 65(13)
siasi	23(46), 24(16), 30(17), 31(25), 34(12), 44(34),60(11)

B. Main categories of Hamshahri collection

There are totally 82 categories in Hamshahri collection. Table 5 contains 16 main categories in Hamshahri collection that contain more documents than other categories.

Table 5. 16 main categories of Hamshahri collection

Category Tag	Category Name In Persian	Category Name In English
adabh	هنری-ادبی	Art-Literature
akhar	اخبار کوتاه	Short news
bankb	بورس و بانک	Stock market & Banking
econw	اقتصاد جهانی	World's Economy
ejtem	اجتماعی	Society
elmif	علمی و فرهنگی	Since & Culture
eqtes	اقتصادی	Economy (in Iran)
gards	گردش گری	Tourism
gungn	گوناگون	Miscellaneous
havad	حوادث	Social events
ikaba	فناوری اطلاعات	Information Technology
kharj	اخبار خارجی	Foreign news
shahr	شهر تهران	Tehran & Municipal affairs
shrst	شهرستانها	Iran cities (except Tehran)
siasi	سیاسی	Politic
vrzsh	ورزشی	Sport