IFLA WLIC 2018
KUALA LUMPUR
Submitted on:19.07.2018

Lessons learned from Automatic Indexing Projects regarding to Persian Language Specifications

Mahboubeh Ghorbani
PhD in LIS & Deputy General manager of Research and Education, National Library and Archives of I.R.of Iran
E-mail: mahghorbani1353@gmail.com

Fattaneh Torkashvand
PhD student in LIS, expert responsible for indexing and thesaurus, National Library and Archives of I.R.of Iran
E-mail: f.torkashvand94@gmail.com

Abstract:

Persian reading and writing are associated with some difficulties due to specific features of this language. this paper attempts to examine automated indexing experiences, lessons, and outcomes of Persian language documents to provide effective solutions for improvement of indexing and retrieval of them. The most important problems in Persian language and script in automatic indexing include selection of an appropriate keyword, building a vocabulary, Semantic, Verb and word sense ambiguities in the sentences, Spaces and Pseudo-spaces in Persian script, isolated and cursive writing, morphology of Persian language, typographical and spelling errors. Removing the stop words, pre-processing of characters and script, identifying the boundaries of words, equalizing different spellings, the automatic stemming, Weighting and scoring of words, Detection of phrasal verbs and compound phrases, Spellchecking through creating morphological or even syntactic spellcheckers design of a corrector and proposer system , developing an infrastructure database for Persian language and script usage are solutions proposed to facilitate the automatic indexing of Persian texts.

Keywords: Persian language; automated indexing; storage & retrieval

Introduction

The Iranians and peoples of several neighboring countries of Iran speak beautiful Persian language; and use this language to exchange and share their knowledge and information in the electronic environments.

Persian language (Farsi) is one of the used languages in the Middle East and as is a member of the Indo-European family written using Arabic letters. There are 32 letters in Persian language. Words are written from right to left and numbers from left to right, both horizontally. Most letters join the following letters. Depending on the position of each letter in a word, it can come after joining or non-joining letters and this has caused complexities in this language (Danesh, Minaei, and Kashefi, 2013).Having more than 100 million native speakers, Persian is to some extends related to the majority of European languages (Dolamic and Savoy, 2009; Bijankhan*et et al*., 2011; quoted in Akasereh and Savoy**,** 2012)**.**The underlying morphology is slightly more complex than the English one, but we cannot qualify it as hard compared to some languages (Dolamic and Savoy, 2009).

In some cases, Persian reading and writing is associated with some difficulties due to specific features of this language. With the sudden entry of computers into the wide range of different social, cultural, economic and technical activities, experts did not find the opportunity to think and apply a fundamental and comprehensive strategy to confront the challenges of writing style (Horri, 1993; quoted by Sotoudeh and Honarjouyan, 2012).The lack of a comprehensive and popular writing standard has led to the lack of coordination and inconsistency in the way information is embodied in databases, websites, blogs and other digital sources, which in turn provided several problems for the searches in Persian. On the other hand, unsolicited use of these commands makes the adoption and institutionalization of this style to be a very long-term, if not impossible, process (Sotoudeh and Honarjouyan, 2012).

Due to increasing number of e-publications and users' desire for full text searches, the designers of automatic indexing software seek to increase the capabilities of their programs; and also seek to add the special methods of manual indexing to them (Dolani and Farhadpour, 2009). The most important issues of automatic indexing are linguistic problems, and human social activities, not the technology- and computer-related issues. (Moradi Moghadam, 2009).

Design or methodology:
The study is a survey research and data were collected through reviewing and studying research reports and projects in the field of automatic indexing, as well as computerized storage and retrieval of Persian sources. The problems of Persian texts in automatic indexing and the suggested solutions were then analyzed. In the report of each research or project, the goal, identified problems and tried or proposed solutions are presented structurally.

Findings and implications:
In a project entitled "Making an Automatic Indexer for Persian Texts", Tashakkori and Meybodi (2003) addressed the problem of different methods of composition, selection and use of vocabulary in the Persian texts and the automatic indexing of Persian texts.

Solution: through studying the methods of evaluating the text retrieval systems, an automatic indexing system of Persian texts was provided, and was evaluated by using two parameters of "recall" and "precision". The qualitative changes of the system during stemming were also studied using the automatic stemmer of Persian words. Then, in a test environment consisting of 450 abstracts and 32 different Persian queries in the field of computer, it was found, by verifying repetitive words, that 20% of the words used in the Persian texts were distinguished; and undistinguished words accounted for more than 85% of the total words used in the texts. That is, a large part of the words of the texts were composed of general words. After identifying each of the words, a list of 150 words from the general words of this language was prepared. In general, when high level of “precision” is more important than the level of “recall”, vocabulary stemming is not recommended, but if the level of recall is more important, the vocabulary stemming is a great help for this.

Bashiri, Karbalaei, and Mousavi (2005) Since the selection of keywords that indicate the content of the document and the difference in the construction and usage of the words in Persian language have been mentioned as the most difficult activities in the indexing process in this project, designed and implemented "Sina" automatic indexer for Persian texts.

Solution: Sina automatic indexer for Persian texts was designed and implemented with an automatic stemmer to reduce the size of index and use the roots of Persian words at the place where the indexes are kept instead of keeping all the words. Four different weighting methods were used to evaluate the indexes created by Sina. At the end, the indexer and the stemmer evaluated themselves. In the first step, a 180-word list was created for the general vocabulary of the Persian language. The use of the indexer's stemming algorithm, provided in this project, also resulted in the improvement of the precision of indexer and 53% decrease in the volume of the index. The implementation of different weighting patterns in this indexer led to the identification of weighting patterns suitable for the Persian texts.

Rasouli and Bidgoli (2008) conducted a New Method for Spellchecking in Persian Language. Some of the most important problems presented in this research are: the polysemy and multi-role status of words ; the use of phrasal verbs, expressions (slangs) and proverbs that usually have a completely different meaning to their apparent meaning; having no arrangement that makes it difficult to compose grammar and do syntax analysis of sentences; misalignment of the components of the sentence, such as the numerical correspondence of the subject and the verb, and the semantic correspondence of the components of sentence and the components of the nominal phrases ;Computer processing issues include lack of vocabulary, grammar, patterns, and codified grammar rules ; the absence of a definite criterion in the Persian script, the mixing of the Persian and Arabic computer script, the issue of spaces and semi-spaces and the lack of a reliable lexical treasure in the Persian language.

Solution: different methods of spellchecking through the use of intelligent machine, the problems have been resolved in the case of letters that have several types of characters in computerized script. After spellchecking, the spellchecker provides correct suggestions to users. To find the correct words for the wrong words, the program uses the previous and the next words of the considered word, and analyzes the wrong word itself so that it can extract a combination of two correct words from the wrong word. Stemming of nouns, adjectives, adverbs and verbs of Persian language has been implemented in the spellchecker. The infinitive forms of the verbs are separated based on the tenses; the stemming is also evaluated based on the tenses, and the issue of connected pronouns has also been resolved in it. The spellchecker uses a separate method to retrieve the verbs. As far the nouns are concerned, being singular or plural, definite or indefinite and having affixes have also been evaluated in it.

In 2009, in the comprehensive plan of the Persian Language Corpus, executed with the cooperation of the Supreme Council of Information and Communication Technology as well as the University of Science and Technology, the issues of automatic indexing of Persian texts in different phases were investigated and some solutions were suggested.

In this project, the challenges of creating the Persian language and script infrastructure database in the natural language processing of Persian text corpus include text segmentation (Splitting of text into semantic units such as words, sentences, phrases), semantic ambiguity in phrasal words (Pseudo-space challenge), part-of-speech-tagging (Preparation of a list of types and roles of words in the Persian language that can be selected from that list), word sense disambiguation, the recognition of the considered sense of any word used in several senses,

syntactic disambiguation that results from the relationship between the words and phrases in the sentence, normalization (the existence of multiple isomorphic or similar characters that are different in computational linguistics), the presence of additional spaces in the text that makes it difficult to extract words and phrases, detection of speech acts, multiplicity of scientific equivalents, diversity of recording names, determination of the boundaries of words, cursive writing, isolated writing, and break-less writing of words, types of plurals, various forms of writing and the use of colloquial language in the text, especially in the blogs, differences such as the connection of suffixes, spacing, transformation of words during linking, adding diacritical marks to specific characters, coding of Persian characters, complicated and ambiguous morphology of Persian script with the approach of spelling problems such as the verb-related rules, affix-related rules, and spacing rules, the existence of many identical letters, homophones, homonyms, and different distribution of typographical errors.

Solution: One of the solutions is to create an infrastructure database that covers features of Persian language using the text segmentation tools, part of speech tagging, part of speech disambiguation, syntactic disambiguation, and normalization of the text corpus consisted of language vocabulary sections, thesaurus, language templates, common language syntagmatic relation, data tagged corpus, and specialized corpus. A finite-state morphology method was proposed to diagnose compound verbs, in which verb detection is done in two steps: firstly, the verbs are only detected based on their appearance. Secondly, it is tried to find the most likely verbs through higher level processing. At the beginning of the process, in order to organize the verb pattern in the database and define the list of verbs in the database by the users, a simple language is provided to the users, by which they must identify the fixed and inflectional parts of verbs and the type of verb structure. Solving the problems that are influenced by the structure of the language requires a pre-processing step, including the equalization of characters coding, equalization of script, detection of the boundary of words, elimination or equalization of diacritics, equalization of different spelling words, removal of stop words, matching the list of these words (such as pronouns, adverbs, propositions and conjunction which are repetitive), identification and correction of keywords that the user has entered incorrectly (due to mismatch of putting space and semi-space between the Persian words), stemming of words that requires the recognition of the grammar and provision of a machine compatible with the structure of the Persian language in which the indexer matches the words with the prefixes and suffixes in several steps in order to extract the root of the word, and the use of different weighting patterns in which the weights indicate the effect of the word on the subject matter compared to other words. Designing an appropriate spellchecker is another solution that is presented for problems arising from the existence of different and sometimes conflicting approaches and metrics for how words are spelled out or written. An appropriate spellchecker for the Persian language should be equipped with a lexical stemmer to recognize the different derivations of each word, and have the conjugation of all verbs in its glossary. Therefore, the use of a combination of robust and reliable stemmers for the majority of verbs and the insertion of particular verbs in the glossary is now considered to be the best solution. The use of automatic affix-separating tools has also been recommended. The wrong spacing in the Persian language is a challenge in the design and development of morphological and even syntactic spellcheckers. Paying attention to spelling correction of homonyms should be considered separately in the design of Persian spellcheckers.

Moradi Moghadam (2009) explored different aspects of extraction indexing with a Glance at Indexing Problems in Persian Language. Issues such as the disagreement over the recognition of the boundaries of Persian words, the failure to follow a certain rule in distinguishing

between compound and simple words, as well as singularity- and plurality-related issues, articles and other Arabic words entered in Persian language, cursive or isolated writing, and compound words are of the most important difficulties in Persian language indexing. Other problems with Persian scripts are the lack of separate signs for each single phoneme or the presence of several signs for a phoneme.

Solutions: Assignment indexing, as a complementary for the extraction indexing, can be effective in addressing the problems associated with extraction indexing, because indexing seeks linguistic analysis and the discovery of hidden meanings of documents. Weighting words is another method that reduces false drop and increases the accuracy. In this method, the words are scored according to their location in the text (such as title, abstract, etc. Usually, the presence of a word in the title of a document will have the highest score. In order to resolve the problem of frequency indexing, it is necessary to set a threshold for the repetition of word frequency. The threshold of frequency can vary depending on the type of text in which the word is located (the subject area). The application of an intelligent Persian interface in the search engines can reduce false drops caused by the indexing methods of Persian language websites. The use of a powerful national exploration tool, which is compatible with the information needs of Persian-speaking users, while taking into account the challenges of Persian script, and the design of a Persian language spellchecker seem to be promising.

Sotoudeh and Honarjouyan (2012) reviewed the literature and existing backgrounds and identified the challenges of Persian writing and their impact on the effectiveness of information retrieval and made suggestions to resolve these difficulties. The most important challenges identified in this research were cursive writing or isolated writing; the diversity of plural morphemes (indicators of plurality; the difference in phonemes/diacritics; the diversity of other writings; the gap between words; and writing from right to left. The challenges presented in the texts are divided into three semantic, syntactic, and morphological levels. Semantic challenges can be attributed to the diversity of vocabulary and richness of language as well as the dependence on foreign languages. Some of these challenges are the dependence of Persian language to the foreign languages, the lack of a standard for the phonetic transcription of foreign vocabulary resulted in the difficulty of Persian language retrieval, and morphological diversity of Persian writing that leads to writing, non-writing, or variation in the writing of letters, signs or diacritics. Each of these challenges, depending on the frequency of their occurrence in the text or in the search phrase, affects the recovery results in varying degrees.

Solutions**:** Each solution can be attributed to one or more specific stages of the life cycle of the document, *i.e.,* prior to retrieval, and during retrieval. The first category consists of some solutions for writers and typists when producing documents or storing them. Also, these solutions can be used during indexing to produce document representations. The second-class solutions involve all those seeking information in digital environments. These strategies consider two general ways, i.e. development of tools and rules for standardizing document writing, index terminology, and query terminology. In the “script coordination” solution, the emphasis is on the need for a credible reference to adopt and provide a standard for writing, and to enforce it, so that the boundary and method of writing a word can be subordinated to a single rule. Another solutions are to use a ready-made checklist including various possibilities of writing are linked through references to each other; to compile a specialized thesaurus for the Persian language containing standard terms in any discipline and an acceptable writing style; to suggest rules for the uniformity of Persian writing; to use both singular and plural forms in indexing; to provide tools and standardizing text output that can

be reduce the difficulty of Persian language retrieval. Development and compilation of the tools such as thesaurus, spelling cultures, and standard writing rules is an effective step in increasing the effectiveness of retrieval.

Akasereh and Savoy (2012) conducted an experiment to evaluate the retrieval effectiveness of different IR models while using a light and a plural stemmer as well as n-grams and trunc-n indexing strategies. Problems included Word segmentation: Defining the word's boundary in Persian is, sometimes, a challenging task. this variety of writing creates challenges in the process of tokenization as well as stemming phase; Inflectional morphology: Persian language has affixitive morphology. Adding a prefix or suffix to a stem may cause a change in the word's orthography which makes the process of stemming and lemmatization more complex; grammatical issues: In Farsi proper names are written in the same way as other words so it is not easy to prevent them from being stemmed.

Solution: were proposed a light and a plural (very light) stemmer along with a stopword list; and finally, after applying these with different IR models on test collection, were made a query-by-query analysis on the results to discover the weaknesses of these different methods. the test-collection used for this experiment was the collection is made up of 166,774 newspaper articles, with approximately 202 terms per document (after stopword removal), extracted from a national Iranian newspaper ("Hamshahri") between the years 1996 to 2002. There are 100 topics in the collection which have a total number of 9,625 relevant documents. only the "title" section is used. The applied stop words list contains 881 terms covering the frequent terms. indexing approaches that were used were n-grams and trunc-n. For stemming a light suffix-stripping algorithm was used which removed the morphological suffixes (mostly inflections) such as possessive, plural, relative, etc. The procedure was mostly focused on nouns and adjectives. Another stemmer is also tested on the collection that removes only the plural suffixes. To evaluate the retrieval performance the mean average precision (MAP) was used based on the 100 queries. Also, a query-by-query analysis was also applied for some of the models and strategies. The results revealed that for Farsi stoplist removal plays an important role in improving the retrieval performance. A query-by query analysis on the results showed that avoiding extreme results would be possible by adding extra controls and rules, according to Farsi morphology, to the stemming algorithms.

Danesh, Minaei, and Kashefi (2013) present a method for Persian documents retrieving using N-gram[1] indexing and distribution technique. As the style of writing in Persian language is not firmly defined, The most common writing challenges in Persian language that affects the IR are : Writing in informal or colloquial writing also majorly affects the syntax of the sentences; Use of foreign words, especially in scientific web pages; Complex Inflection includes more than 2800 declensional suffixes, Close-fitting, non-close-fitting or fitting with pseudo-space of affixes; Multiple types of writing for a word; Homographs: The same word may have different meanings; Word spacing: white space as inter-words space, an intra-word space called pseudo-space delimits word's parts; Entrance of Arabic letters in Farsi, two writing forms of some letters, and create new words in Farsi that isn't in the dictionary are some problems causing changes in the results of internet searches and should be resolved in the normalization.

[1] In the fields of computational linguistics and probability, an ***n*-gram** is a contiguous sequence of *n* items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The *n*-grams typically are collected from a text or speech corpus. When the items are words, *n*-grams may also be called *shingles*[clarification needed] [1]

Solution: were studied the behaviour of lemmatized N-grams in retrieving Persian information, as N-gram indexing method reacts well against spelling errors and changes and overlooks lingual morphemic change. Then were used document weighting to index and determine the rank of found documents. to evaluate the proposed method, were used a standard collection for Persian texts called Hamshahri, and were selected 20 queries of it, and be almost 4-grams and a few numbers of 3-grams. These N-Grams were processed simultaneity and execute in a few milliseconds. were reached a high performance in extracting Persian documents by presenting a formula for weighting these N-grams and finding documents with the highest weight. Also, we used distribution method to increase the performance speed of information retrieval. Results showed that using a combination of unigrams and 3-grams improves the retrieval of Persian document.

Parseh & Baraani (2014) presented automated document classification system that uses a novel term weighting method based on semantic relations between terms. Most automatic text classification systems designed for English texts. Because of the complex nature of Persian language, such as words with separate parts and combined verbs, the most of text classification systems are not applicable to Farsi texts. Most of the existing weighting methods exploit the statistical information of terms in documents and do not consider semantic relations between words.

Solution: The proposed method acquires semantic relations between terms using a thesaurus to obtain a semantic weight for each term in a document and using *tf-idf* method, obtains a statistical weight for each term, finally uses the sum of these weights for weighting terms. In this method, to rich documents, feature vector of each document is extended using thesaurus. The obtained results indicated that the proposed method is able to increase the performance of average recall, precision.

Tavakolizadeh-Ravari (2015) has identified the need to design a two-step break-cull model for the automatic indexing according to the characteristics of Persian language; so that these models would be designed in such a way that the interest in indexing is considered.

Solution: Persian automatic indexing model was evaluated through the Inclusion Index, which was used to determine the percentage of consistency among indexers. the degree of consistency of the terms created by this model was examined using the keywords of the authors of articles. The findings showed that in 90% of cases, the term that the model identified in an article as the most prominent term is similar to the first keyword of the author of that article. In sum, there were 76% consistency between the results of this model and the keywords of the authors. To implement the proposed model, the assumption is to use the language of the phrases with a template which is supported by many programming languages and reduces the need to install and use database tables for text processing. It also solves the problem of determining the upper threshold of the main terms. In addition, with a special algorithm, it also defines the lower limit (threshold), so that the number of selected terms would not depend on the length of the text. This capability ensures the integrity and accuracy of indexing.

Conclusion:

By reviewing the researches and projects of automatic indexing of Persian resources, the problems of this language in automatic indexing are shown in the Table 1 according to its characteristics in Table 1:

Table1. The most important problems in Persian language and script in automatic indexing

Problems			
1	Selection of an appropriate keyword	6	Spaces and Pseudo-spaces
2	How to use and build vocabulary in texts	7	Writing verbs and compound words including isolated writing and cursive writing
3	Semantic ambiguities of words in texts	8	The existence of a lot of homonyms and homophones
4	Verb and word sense ambiguities in the sentences	9	Morphology of Persian language
5	The issues of Persian script	10	Typographical and spelling errors

Each of the studies and projects evaluated in this research have tried to provide solutions for the problems identified in the field of automatic indexing of Persian texts, whether theoretically or through implementing the indexing systems. In Table 2, the most important solutions presented to facilitate the automatic indexing process of Persian texts are shown in terms of the features of this language.

Table 2. The solutions proposed to facilitate the automatic indexing of Persian texts

Solutions			
1	Removing the stop words	6	Detection of phrasal verbs and compound phrases using automatic methods
2	Pre-processing involves uniform coding of characters and script, identifying the boundaries of words, removing or equalizing different spellings.	7	Spellchecking of Persian words through creating morphological or even syntactic spellcheckers
3	The automatic stemming of Persian words	8	The design of a corrector and proposer system after spellchecking
4	Weighting and scoring of words based on their importance in the sentence	9	Standardizing of document writing, index terminology and query terms in the system
5	Developing an infrastructure database for Persian language and script usage, i.e. for the text segmentation, part of speech tagging, semantic and syntactic disambiguation and, and normalization	10	Compilation of specialized thesaurus for the uniformity and coherence of Persian writing

The increasing Persian resources on the Web have led to more efforts of experts in linguists, indexing and computer technology to resolve the challenges of automatic storage and retrieval of these resources. In line with this, the findings of this research show this concern in the field of automatic indexing of Persian language. Experts and researchers of the information retrieval in Iran and those interested in Persian language provide a clear vision for improving automated indexing and representing the Persian scientific products around the world.

References:

Akasereh, M.; Savoy, J. (2012) *Retrieval effectiveness study with Farsi language presented* at Actes 9e Conférence en Recherche d'Information et Applications CORIA'12. Retrieved on May 1,2018: https://pdfs.semanticscholar.org/63bf/36734f480ad184823703a40a1486d312bf39.pdf

Bashiri, H; Karbalaei, fatemeh and Mousavi, Shirin (2005) *Design and evaluation of Persian indexers automatically*, A Poster presented at 11th Annual Conference of Iranian Computer Society. Retrieved on May 9, 2018: https://www.civilica.com/Paper-ACCSI11-ACCSI11_100.html

Danesh, M.; Minaei, B., and Kashefi, O. (2013) A Distributed N-Gram Indexing System to Optimizing Persian Information Retrieval, *International Journal of Computer Theory and Engineering*, 5(2), 214-22, DOI: 10.7763/IJCTE.2013.V5.681

Dolamic, L.; Savoy, J. Ad (2009) *Hoc Retrieval with the Persian Language*. Retrieved on May 1,2018: https://pdfs.semanticscholar.org/a1a2/43b19aa31c48c8380acffa15a828df8ccb2b.pdf

Doulani, A.; Farhadpour, M.(2009) Automatic indexing and common softwares: a review , *National Studies on Librarianship and Information Organization (NASTINFO)*,20(3),291-310. Retrieved on May 1, 2018: http://nastinfo.nlai.ir/article_259_9922d1d7321ceafaf0ef78afed481a14.pdf

Farsi Concept Master Plan (2009) Research Project, Supreme Council of Information and Iran University of Science and technology.

Moradi Moqadam, H. (2009) Indexing auto-extraction with a glance over indexing problems in Persian language, *Journal of studies in library and Information Science*,3(2), 135-168.

Parseh, S. & Baraani, A. (2014) *Improving Persian Document Classification Using Semantic Relations between Words*. Retrieved on May 9,2018: https://arxiv.org/ftp/arxiv/papers/1412/1412.8147.pdf

Rasouli, M.S.; Minaei Bidgoli, B. (2008) A *new way to spell mistakes in Farsi*, presented at 2nd Iranian Data Mining Conference. Retrieved on May 9,2018: https://www.civilica.com/Paper-IDMC02-IDMC02_026.htm1

Sediqi, M.; Zamani far, K. and Sahidi, M. (2005) A way to resolve the language challenges of Persian-language web sites, *Journal of Information Science and Technology*, 21(2),47-69. Retrieved on May 1,2018: http://jipm.irandoc.ac.ir/article-1-97-fa.pdf

Sotudeh, H., Honarjooyan, Z. (2012) An overview of the difficulties of Persian language in the digital environment and their effects on the effectiveness of automatic text processing and data retrieval, *Library and Information Science*, 15(4),92-59. Retrieved on May 1,2018: http://lis.aqr-libjournal.ir/article_42651_05011294ef5c5cbaa8986071ca229087.pdf

Tashakori, M. and Meybodi, M. (2003) *Build an auto indexer for Persian texts*, Presented at 11th International Electrical Conference. Retrieved on May 9, 2018: https://www.civilica.com/Paper-ICEE11-ICEE11_018.html

Tavakolizadeh-Ravari, M. (2015) Two Steps Break-Cull Model for Automatic Indexing of Persian Texts, *Research on Information Science & Public Libraries*,21(1),13-40. Retrieved on May 9,2018: http://publij.ir/article-1-966-en.html

Varedi, M. (2010) *Problems and Disadvantages*, Retrieved on May 9,2018: http://islamicdoc.org/article-section/articles-namayeh/602-namayeh-mashin.html