Stop Words in Persian language Stem

Shapur Reza Berenjian

Faculty Member with Regional Information Center for Science and Technology (RICeST)
Sh_berenjian@yahoo.com

Abstract: In modern systems of information storage and retrieval, searching methods are of special interest, since how to search and use appropriate words have direct effect on speed of retrieving, volume of space occupied in PC memory and user's satisfaction. A group of words that are discussed by linguists are stop words, deleting of which will produce better results in indexing documents. Therefore preparing a list of stop words in each language requires knowing different criteria and methods. However, no standard stop word list has been extracted from Persian language texts up to now. This article aims at providing instructions and criteria for preparing stop word list in Persian language in order to develop information retrieval systems.

[S.H. Berenjian. **Stop Words in Persian language Stem.** *Stem Cell* 2015;6(3):1-5]. (ISSN 1545-4570). http://www.sciencepub.net. 1. doi:10.7537/marsscj050115.01

Key words: Stop words, Information retrieval, Semi- stop words, Stop word list, Persian language.

1. Introduction

By entering computer to the information field of information, information storage and retrieval systems have been changed. The changes are considered because the systems which were previously adapted with manual and traditional ones, should match themselves with low and high capacities of modern technologies and proceed in this direction (Horri, 1994). The main topic in this study is how to retrieve information from the computer.

Nowadays, widespread use of computer in libraries, increase of digital libraries and also increase in the users' usage of World Wide Web, caused the authorities and who are involved in these kinds of libraries and information networks and websites pay too much attention to the different theories of information storage and retrieval. So how to search in computers can be very effective factor in accessing users to their needed information. It is obvious that searching has also direct relationship to the kind of indexing. Indexing and retrieving information are always dependant to the key words users choose and users' satisfaction is completely dependent to the selection of the appropriate words in searching.

However, selecting suitable words for optimal search are hard for the most of users (Fattahi, 2006) and using unsuitable words increases the number of retrieved false documents. One of the problems in information storage and retrieval, is that most of the time the main key words with weight and meaning, are just entered by indexers and this causes retrieving the large number of irrelevant documents. While some other aspects of a subject may be desired by users for which these kinds of searching are not suitable (Fattahi, 2006). Because Natural Language processing (NLP) techniques depend on the length of the queries, the longer the queries, the more useful is Natural Language Processing (Mehrad & Naseri, 2008). In order to achieve these needs and increase the accuracy in retrieval results, some worlds called "stop words" can be added to the key words.

2. Definition

General categories are the complex issues which draws linguists attention to themselves and several analyses have been provided in this regard (Esmaeeli fard, 2005). In this part we are going to define "stop words". In Wikipedia, it is defined as: stop words are words which are filtered out prior to, or after, processing of natural language data (text). Hans Peter Luhn, one of the pioneers in information retrieval, is credited with coining the phrase and using the concept in his design.

It is controlled by human input and not automated. There is not one definite list of stop words which all tools use, if even used. All NLP tools do not use stop words list. Some tools specifically avoid using them to support phrase search. Making use of root computing may reduce some parts of logic or dependence on a list of stop words to filter words. Stop words can cause problems when searching for phrases that include them, particularly in names such as än Kæs (The Who), or än rä begir (Take That) (Wikipedia, 2010). Dr. Fattahi defines it as those words that are not searched alone because they have not any special meaning. They are always used along with subject terms to show a special aspect of that subject. For example: moghædæmei bær (an introduction to….), äshenai bä (familiarity with…), dærbäreye (about…) (Fattahi, 2006). A universal feature of general categories is that none of them have an informative meaning in documents. However they are used in sentences and phrases because of their grammatical role (zo, 2006), (Rogavan, 1986), (Ricardo, 1999).

As a general it can be said that stop words are those common words that have too little meaning and have only grammatical role with no topic and content (Abolkheir, 2006). Stop words are a list of words in the list of terms (preposition, pronouns, etc) which do have little meaning and have been ignored while being retrieved in a document(Mehrad & Naseri, 2008). Almost all information retrieval applications remove stop words before processing documents and queries. This usually increases system performance (Mehrad & Naseri, 2008). In addition to stop words, a number of specialists called some other words as semi- stop words. Dr. Fattahi defines semi-stop words as the words which are not usually searched alone but like general words, are used along with subject key words. For example, xætære (risk of ….), hädeseye (incident of …), pishgiri æz (prevention of…) (Fattahi, 2006). Although semi-stop words can apparently be used as key word, they are preferably used with other words as complementary.

3. Background

Fox made the list of stop words for the first time in English language. He prepared a list containing 421English words (Abolkheir, 2006). Up to now, some lists of stop words in some languages such as English, French, Germany, Arabic, and Russian have been published. In Persian language, for the first time, (Savi, 2008) had made a list of stop words. However, as the list was a translation of the stop words of English and Arabic languages, it has several basic problems. First, no criteria were provided to prepare the list so it is not in accordance with the extraction criteria of Persian language and some of the words in this list does not have any meaning in Persian. Second, the list has not shown thousands of words with high frequency or it shows some words in different forms such as bale, bali (yes)- are, ari (yes),… This list lacks constant stop words exist in Persian language and as it is prepared based on an specific subject area, it cannot be extended to other areas. Third, some foreign words entered in Persian language such as merci (thanks), and connected pronouns such as shan, im, ash are part of this list too. Also inarticulate (h) which is either an identifier of subjective adjective or objective adjective are included in this list of stop words. Although based on written form of Persian language, these suffixes can be written separately, it should be considered that they are part of a word and their deletion will result in changing their meaning.

So preparing an standard and complete list of stop words in Persian language is very important. Moreover, these kind of lists, either or not they will be applied in information processing, have different functions (Mehrad & Naseri, 2008).

4- Kinds of stop word lists

Since the stop word lists have very important role in language processing, storing and retrieving information and electronic documents, it is very crucial to find suitable criteria for preparing a standard list.

So far, several lists of stop words have been prepared which are extracted traditionally by analyzing all words of a great corpus of language, for example English, French, Germany, etc. However the results of analyzing various corpuses are often similar and are generally used as standard lists.

As a general, the list of stop words should include adverbs, prepositions and conjunctions. Nevertheless the list should be revised in the data analysis of a specific field of study. It should be noted that most of researchers have emphasized on using fixed terms such as those available in the list of stop words especially empty phrases (Mehrad & Naseri, 2008).

High frequency of common words in different areas or overlap of the stop words in several fields show that these lists are completely general (absolutely general) and do not depend on any particular subject area. For example tärixe (history of …), moghædæmei bær (an introduction to…), dærbäreye (about…). On the other hand, there are other kinds of stop words that are applied in a specific subject area. For example æxbäre (news of …), æshkäle (forms of …) which are called stop words of special fields (Fattahi, 2006).

As a whole, to produce a list of stop words, the corpus of intended language should be determined. For example 66406121 words of English language and 34841204 words of French language are analyzed in their corpus to produce their stop word lists and about 571 and 463 stop words were extracted in English and French languages respectively. These extracted words are those called absolutely general. It is obvious that to produce a list of stop words of a special field, its specific corpus should be studied. Sometimes we encounter problems using this kind of stop word list. For example some words which are part of subject words, are used as stop words in this kind of corpuses such as the word "physics" in the documents related to Physics area.

Generally, the words are extracted as stop words based on their high frequency or repetition in the corpus. Because, although these words may be used as subject words, they occupy additional space and increase the number of retrieved documents. Moreover, low frequency words also should be regarded as stop words since they have not any effect on the information content of the retrieved documents even if they are used as subject words. Anyway, unauthorized terms are identified by frequency analysis and expert judgment (Mahrad & Naseri, 2008). In addition to the mentioned points, the following items should also be considered in producing a standard list of stop words in Persian language.

1-4: adverb

1-1-4: specific adverbs end with "æn" such as mondærejæn (inserted), Mashrohæn (comprehensive), tadrijæn (gradually)

2-1-4: common adverbs: xänä (readable), zibä (beautiful)

3-1-4: adverbial group: dær æväsete in doreh (in the middle of this period)

4-1-4: compound adverb: hær ru:z (every day), hær jä (anywhere)

5-1-4: adverb of time: Mordäd (July), æsr (evening)

6-1-4: adverb of place: bälä(up), päi:n (down), dær hæyät (in the yard)

7-1-4: adverb of quantity: kæm (little), ziad (many)

8-1-4: adverb of quality: xu:b (good), bæd (bad), kæj (crooked), ähesteh (slowly)

9-1-4: adverb of manner: xændän (smiling), dæliräneh (bravely), geryän (weepy)

10-1-4: adverb of wish: käshki (I wish)

11-1-4: adverb of surprise: ei æjæb (surprisingly), shegefta (wonderfully)

12-1-4: adverb of frequency: dobäeh (again), digær (else), bäz hæm (anymore)

13-1-4: adverb of explanation: be ebäræte digær (in other words, that is)

14-1-4: adverb of sequence: pei dær pei (successively, serially)

15-1-4: adverb of question: koja (where), chera (why)

16-1-4: negative adverb: hærgez (never), be hich væjh (no way)

17-1-4: adverb of emphasis: be rästi(indeed), be dorosti (verily, truly)

18-1-4: adverb of uncertainty: shäyæd (maybe), ehtemälæn (probably)

19-1-4: adverb of simile: gu:i: (as if, like)

20-1-4: adverb of reason: bænäbærin (so, therefore), xira (because)

21-1-4: adverb of polite: xoda nækærdeh (god forbid)

22-1-4: adverb of abbreviation: xoläse (in sum)

23-1-4: there are other kind of adverbs such as:

Adverb of count: do bär (twice), adverb of sequence: ævælin (first), dovomin (second), adverb of doubt: shäyæd (perhaps), adverb of means: dæsti (manually), telefoni (by telephone), adverb of approximation: hodu:de (about), adverb of value: pænj tomæn (five tomans), exclusivity adverb: fæghæt (only), … however there are more adverb complements in Persian language (Farshidvar, 2003).

2-4: verbs

1-2-4: successive verbs; begir beshin (sit down), bezan berim (lets go)

2-2-4: verbs with no complete meaning; gashtan (be), gardidan (become)

3-2-4: Modal verbs; budan(be), shodan (become)

4-2-4: infinitive verbs; goftan (saying), raftan (going)

5-2-4: semi- modal verbs; bayestan (should), tavanestan (could), shayestan (may)

6-2-4: conjugated verbs

3-4: adjectives

1-3-4: interrogative adjective: chegu:neh, kodäm, chænd tä (how, which, how many)

2-3-4: exclamatory adjective: che zibäst, che bozorg æst (what nice it is, what big it is)

3-3-4: objective adjective: shenideh, gerefteh (heard, taken)

4-3-4: attributive adjective ; næmækin(salty), doru:ghin (deceptive)

5-3-4: competency adjective: xändæni, xästæni(readable, desirable)

6-3-4: negative adjective: näshäyest, bixod, bidærdesær (improper, unduly, effortlessly)

7-3-4: demonstrative adjective: in, än, inha, änha (this, that, these, those)

8-3-4: simple adjective: xu:b, bæd, særd, sabz (good, bad, cold, green)

9-3-4: compound adjective: sorx xu:n, nime væght (red-blooded, part-time)

10-3-4: main numerical adjective: yek, do, seh (one, two, three)

1-10-3-4: ordinal numerical adjective: ævælin, hæftomin (first, seventh)

2-10-3-4: fractional numerical adjective: seh dæhom, yek pæjom (three tenth, one fifth)

11-3-4: indefinite adjective: hær, digær (any, another)

4-4: count unit (portion-mass): 'Qors' meaning 'loaf', eg. One loaf of bread. (Safavi,2008)

5-4: pronouns

1-5-4: separable personal pronouns: mæn, to, u: (I, You, He)(bateni,1977)

2-5-4: common pronouns: xodæh, hæmeh, digæri (himself, all, other)

3-5-4: interrogative pronouns: che, kodäm, chegu:neh (what, which, how)

4-5-4: indefinite pronouns: goru:hi æz, ghesmæti æz (group of, part of)

5-5-4: exclamatory pronoun: che ziba (what nice)

Note: usually adjective and pronouns (demonstrative, interrogative, exclamatory, indefinite) are too similar to be recognized. It can be said that if the word is used along with noun, it is called adjective and if it is used alone, it is called pronoun (Anvari, 2006).

6-4: letters

1-6-4: conjunction

1-1-6-4: simple conjunction: væ, ægær, æma (and, if, but)

2-1-6-4: compound conjunction: Anja ke (there), Angah ke (then)

2-6-4: preposition

1-2-6-4: simple preposition: æz, ba (from, with)

2-2-6-4: compound preposition: bedu:ne (without)

3-6-4: interjection; hey (calling attention)(lazard, 2005)

7-4: quasi sentences: äfærin (excellent, bravo)

8-4: repeated words ; pareh pareh (torn), tekeh tekeh (in pieces)

Conclusion

Generally, there are three different ways to produce stop word lists:

a) A list of words consists of conjunctions, adverbs and some adjectives are considered as stop words in all languages (other than when they are used as key words) and we call them "fixed stop word list".

b) Words obtained from language corpuses (a language corpus is a basic source which contains strategic collection of natural language elements that is readable for machine. The corpuses' data usually cover speech and written modes and encompass a vast range of non-technical language which is used normally -not in poem, fiction or local accent- by adults). This kind of words which are obtained from the collections of documents and based on language corpus of special fields are called "corpus stop word list".

c) List of words which is produced by the combination of the two previous lists "fixed stop word list" and "corpus stop word list" is called "joint stop word list".

Therefore, it seems that producing a standard list of stop words for each language specially Persian language is necessary. Stop words can have two different effects on information processing and retrieval. On one hand, as they have high frequency, their deletion will increase efficiency of retrieving and indexing data. Because this will result in reducing the volume of saved information in computer memory. On the other hand, by adding stop words and phrases and expanding search, users can receive more precise results (Fattahi, 2006). This contradiction shows that stop words are of two kinds: independent words and dependant words. Independent stop words are considered as general terms in any cases such as dær, æz, væ (in, from, and,…) but dependant words are those which are used as auxiliary words while applied along with a subject term such as moghædæmei bær (an introduction to..).

To apply stop words appropriately, it is better to use each group of words in suitable places, and prepare a list combined of the two mentioned lists and use it as standard one in information processing and retrieval.

Correspondence to:
Shapur Reza Berenjian
Faculty Member with Regional Information Center for Science and Technology (RICeST)
Sh_berenjian@yahoo.com

References

1. Anvari, H. & Ahmadi Givi, H. (2006). Persian grammar 2. Tehran: Fatemi Cultural Institute Publication.
2. Abu El-Kahir, I. (2006). Effects of stop words eliminating for Arabic information retrieval: A comparative study. International Journal of Computing and Infromation Sciences, Vol. 4 (3).
3. Bateni, M. R. (1977). Describing the grammatical structure of Persian language. Tehran: Amir Kabir Publication
4. Esmaeeli fard, Z. (2005). Comparing empty pronouns within nominal groups in Persian and Italian languages. Foreign Language Research

Quarterly. Vol. 28(winter), pp. 5-8.

5. Fattahi, R. (2006). Identification and analysis of stop words in web resources: A new approach to the development of search phrases in search engines using natural language. Journal of Education and Psychology of Ferdowsi University, Vol. 7 (1).
6. Farshidvar, Kh. (2003). Detailed grammar of today. Tehran: Sokhan Publication.
7. Horri, A. (1994). Problems in storing pre co-ordinate and retrieving post co- ordinate in computer system. Book Quarterly (fall & winter).
8. Lazard, G. (2005). Grammaire du persan contemporain (M. Bahraini Trans.). Tehran: Hermes Publications (Original work published 1957).
9. Meshkat-aldini, M. (2000). Persian grammar based on transformational theory. Meshad: Meshad Ferdowsi University Publication.
10. Mehrad, J. & Naseri, M. (2008). Natural language processing and information retrieval. Regional Information Center for Science and Technology. Shiraz: Chapar.
11. Ricardo, B. Y & Berthier, R. N. (1999). Moderne information retrieval. Addison Wesley Longman Publishing Boston.
12. Reghavan, V. V. & Wong S. K. M. (1986). A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science.
13. Savoy, Jacques. http:// members.unine.ch/Jacques, savoy/Clef Persian ST.text/.
14. Safavi, K. (2008). An introduction to semantics. Tehran: Soore Publication.
15. Zou, F., Wang, F. & Deng, X. (2006). Automatic identification of Chinese stop word. A special issue on Advances in Natural Language Processing of the journal Research on Computing Science, pp. 151-162.
16. Zou, F., Wang, F. & Deng, X. (2006). The 5th edition of the International Conference on Language Resources and Evaluation (LREC), Genoa, Italy.

5/12/2015