2006

FuFaIR: a Fuzzy Farsi Information Retrieval System

A. Nayyeri
University of Tehran, Iran

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

Publication Details

Nayyeri, A and Oroumchian, F, FuFaIR: a Fuzzy Farsi Information Retrieval System, Proceedings of the 4th ACS/IEEE International Conference on Computer Systems and Applications, Dubai/Sharjah, UAE, 8-11 March, 2006, 1126-1130. Copyright Institute of Electrical and Electronics Engineers 2006. Original conference information available here

FuFaIR: a Fuzzy Farsi Information Retrieval System

Amir Nayyeri

Department of Electrical & Computer Engineering, Univ. of Tehran
a nayyeri@ece ut ac ir

Farhad Oroumchian

University of Wollongong in Dubai;
Control and Intelligent Processing Center of Excellence, Faculty of Eng., Univ. of Tehran
FarhadOroumchian@uowdubai ac ae

Abstract

Persian (Farsi) is one of the languages of Middle East. There are significant amount of Persian documents available in digital form and even more are created every day. Therefore, there is a necessity to implement Information Retrieval System with high precision for this language. This paper discusses the design, implementation and testing of a Fuzzy retrieval system for Persian called FuFaIR. This system also supports Fuzzy quantifiers in its query language. Tests have been conducted using a standard Persian test corpus called Hamshari. The performance results obtained from FuFaIR are positive and they indicate that the FuFaIR could notably outperform well known industry systems such as the vector space model.

1. Introduction

Spoken in several countries like Iran Tajikistan and parts of Afghanistan Farsi language is one of the dominant languages in the Middle East During its long history this language is influenced by other languages such as Arabic Turkish Kurdish and even European languages such as English and French Today s Persian contains many different words from above languages and in some cases these words still follow the grammar of their original languages in building plural or singular or different verb forms Therefore morphological analyzers for this language need to deal with many forms of words that are not actually Persian Arabic script has been adapted for writing Persian language Persian alphabet contains 3 characters 4 more than Arabic and is written from Right to left and continuously

The early attempts in applying Fuzzy logic to Persian language have not produced acceptable results [ ] In this paper we introduce FuFaIR (Fuzzy Farsi Information Retrieval which utilizes a different Fuzzy logic approach than those tested before The fuzzy language used for querying in FuFaIR can support logical operations such as AND OR NOT and even fuzzy quantifiers that can help the user to express himself herself more effectively FuFaIR performance is measured by precision at first 5 5 and documents That is the ratio of relevant items at those document cutoffs FuFaIR outcomes benchmark systems such as the vector space model [ ]

Section two provides an overview of the related works in Persian IR or Fuzzy IR Section three gives an overview of Fuzzy Logic with emphasis on the techniques used in this paper The proposed method is detailed in section 4 Section five describes the experimental results and section six concludes the paper

2. Related Works

There are two different types of prior work that relate to this work These are the Fuzzy IR and the Persian IR approaches This section briefly provides an overview of both approaches For a very good resource of classical approaches to IR we refer you to [ ]

The classical Boolean approaches for information retrieval lacked flexibility Therefore some approaches were suggested to alleviate this shortcoming by using extended Boolean techniques probabilistic approaches or fuzzy logic Authors in [ ] suggested an extended Boolean

approach Then a fuzzy information retrieval system was proposed in [3] That approach also supported fuzzy quantifiers and other connectors Two of the most famous classical fuzzy approaches are presented in [4][5] The main problem with all these approaches is none of them were able to show an acceptable performance comparable to classical retrieval models such as ector Space [ ] Later [ ] has tried to embed fuzzy quantifiers in the [4][5] methods Our work is mostly inspired by [ ]; we applied a similar technique to Farsi IR

Due to the different nature of the Persian language the design of an IR system for it requires special considerations Unfortunately little efforts have been focused on this problem in comparison to other languages Authors in [7] report the result of a series of experiments conducted by applying the existing information retrieval techniques to the Persian text In [ ] the author has described more than combinations of retrieval models term weights and methods that have been tested on Persian text and their results Experiments in [8] suggested the usefulness of language modeling techniques for Farsi Furthermore the design and implementation of a Farsi stemmer is reported in [8]

3. Fuzzy Logic Overview

The fuzzy set theory defines sets whose boundaries are not well defined Considering U as a set of discourse the membership function of the fuzzy set A can be defined as: $\mu_A : U \rightarrow [\quad]$ For every element $u \in U$ μ_A (u stands for its membership degree to the fuzzy set A Ordinary sets can be considered as a special case of the fuzzy sets in which the membership function is either or

Operators similar to the Boolean operators defined for crisp sets are defined and implemented on fuzzy sets also For example a possible definition of Complement intersection and union operators for the fuzzy sets could be as below:

$$\mu_{\bar{A}} = \quad -\mu_A$$
$$\mu_{A\cup B} = \max(\mu_A \quad \mu_B$$
$$\mu_{A\cap B} = \min(\mu_A \quad \mu_B$$

In addition in this paper P(U refers to the crisp power set of U while $\tilde{P}(U$ represents the fuzzy power set of U i e a set which contains all the fuzzy sets that can be defined over U Given as U = {u u u_n} a fuzzy set A constructed over U can be shown as:

$$A = \{\mu_A(u \quad /u \ , \mu A(u \quad /u \ ,\ldots, \mu_A(u_n /u_n\}$$

By applying an α cut operation on a fuzzy set a crisp set is produced that contains major elements of that fuzzy set In a formal manner:

DEFINITION : (α CUT Given a fuzzy set $X \in \tilde{P}(U$ and $\alpha \in [\quad]$ the α cut of level α of X is the crisp set $X_{\geq\alpha}$ defined as:

$$X_{\geq\alpha} = \{u \in U : \mu_X(u \ \geq \alpha\}$$

In the fuzzy set theory fuzzy quantifiers are introduced to provide more flexibility and expressiveness to the operations than crisp quantifiers The quantifiers act as functions that modify the interpretation and value of a quantified statement For example in the following statements "approximately 8 % of smart people are rich" or "most modern cars have air conditioner" the fuzzy quantifiers "approximately 8 %" or "most" influences and modify the sets referred to by "smart people" or "modern cars" In order to define the fuzzy quantifier concept first we need to introduce the semi fuzzy quantifier concept The semi quantifiers works on crisp sets

DEFINITION : (SEMI FUZZY QUANTIFIER A unary semi fuzzy quantifier Q on a base set $U \neq \varnothing$ is a mapping *Q: P(U) →[0,1]* which maps each crisp set $X \in P(U$ into a gradual result $Q(X \ \in [\quad]$

To define the quantifiers different mathematical functions have been proposed For example "Approximately 8 %" as a semi fuzzy quantifier in the evaluation of the sentence "approximately 8 % of X is X " could be defined as below:

$$aprox\_8\ \%(X \quad X \quad = \begin{cases} \frac{|X \ \cap X \ |}{|X \ |} + & \frac{|X \ \cap X \ |}{|X \ |} < \ 8 \\ 8 - \frac{|X \ \cap X \ |}{|X \ |} & \frac{|X \ \cap X \ |}{|X \ |} \geq \ 8 \end{cases}$$

Now the fuzzy quantifier concept can be defined in terms of the semi fuzzy quantifier

DEFINITION 3 (FUZZY QUANTIFIER A unary fuzzy quantifier $\tilde{Q}$ on a base set $U \neq \varnothing$ is a mapping $\tilde{Q} : \tilde{P}(U \ \rightarrow [\quad]$ which maps each fuzzy set $X \in \tilde{P}(U$ into a gradual result $\tilde{Q}(X \ \in [\quad]$

Fuzzy quantifiers as a tool for handling linguistic expressions have been utilized widely in the literature However extra attention should be paid to such properties of these functions as generalization or monotonicity as noted in [3] Some good functions with solid behavior

have been proposed in [ ] [9] [ ] proposes constructing fuzzy quantifiers from semi fuzzy quantifiers through the fuzzification mechanisms In that process the domain of the function is in the universe of semi fuzzy quantifiers and the range is in the universe of fuzzy quantifiers:

$$F : (Q : P(U \rightarrow [\quad] \rightarrow (\tilde{Q} : \tilde{P}(U \rightarrow [\quad]$$

In [9] a quantifier fuzzification mechanism that uses the notion of α cut (definition is introduced Formally the Choquet integral [ ] is applied as follows:

$$(F(Q \quad (X \quad = \int Q((X_{\geq\alpha} \, d\alpha$$

where Q: P(U →[ ] is a unary semi fuzzy quantifier $X \in \tilde{P}(U$ is a fuzzy set and $(X_{\geq\alpha}$ is the α cut of level α of X Note that as required by the semi fuzzy quantifier Q $(X_{\geq\alpha}$ is a crisp set In this method $(X_{\geq\alpha}$ are crisp representatives for the fuzzy set X and roughly speaking the integral over α cuts in the interval [ ] averages out the values obtained after applying the semi fuzzy quantifier to all the crisp representatives of X In this work we will use the restriction to the unary case of this framework because this type of quantifiers is expressive enough for the objectives pursued here Nevertheless we plan to apply quantifiers of higher orders in the near future If the universe of discourse U is finite the previous equation can be made discrete as follows:

$$(F(Q \quad (X \quad = \sum Q((X_{\geq\alpha_i} \quad (\alpha_i - \alpha_{i+}$$

For all discrete values of α_i where $\alpha =$ $\alpha_m =$ and $\alpha >$ $> \alpha_m$ represents a decreasing array of membership values in U to the fuzzy set X The voting model interpretation of fuzzy sets [ ] suggests value α_i α_i as the probability that $(X_{\geq\alpha_i}$ is selected as the crisp representative for the fuzzy set X Thus the semi fuzzy quantifier can be applied for every crisp representative of X These values are then weighted by the probability of their crisp representatives

4. The Proposed Method

Different classical models have been proposed for fuzzy Information Retrieval till now In almost all of such methods the query is considered as a fuzzy set of relevant documents in the database Using this scheme the documents will be sent to the client sorted based on their degree of membership in the query's fuzzy set More formally the membership function can be defined from D set of all documents to [ ] as μ_i: D→[ ]

The larger the value of μ_i the more relevant is the document to the query In order to compute this function a fuzzy set is defined for each term in the indexing process In this process each term of each document is assigned a membership degree based on the importance that term for representing the document s content This membership degree can even be computed easily with classical IR parameters such as tf idf (term frequency index document frequency In this paper we have used the following formula for calculating the membership degrees:

$$\mu_t(d \quad = \frac{f_{t\,d}}{\max_{t_k}(f_{t_k\,d}} \times \frac{idf(t}{\max_{t_k}(idf(t_k}$$

Where $\mu_t(d$ is the degree of membership of document d to the fuzzy set of the term t $\mathbf{f}_{t\,d}$ represents the frequency of the term t in the document d $\max_{t_k}(f_{t_k\,d}$ is the maximum frequency for any term in the document d $idf(t$ or Inverse Document Frequency represents the portion of the collection that contains the term t It is calculated as log (N n where N is the number of documents in the collection and n is the number of documents with the term t

The input query is considered as an algebraic sentence whose elements are fuzzy sets and fuzzy operators such as AND OR and NOT There are different forms of interpretations for the fuzzy operators Here we used the followings:

$$\mu_{A\ AND\ B} = \min(\mu_A \ \mu_B$$
$$\mu_{A\ OR\ B} = \max(\mu_A \ \mu_B$$
$$\mu_{NOT\ A} = \quad \mu_A$$

Still the primary operators lack sufficient flexibility For example the query: At least three sports such as soccer basketball olleyball and Hockey (at_least_3 (sport soccer basketball olleyball and Hockey may not be easily constructed by primary operators In order to overcome this shortcoming quantifiers are introduced that are explained formally in the previous chapter

The language for querying in FuFaIR supports both basic operators and fuzzy quantifiers

The following example demonstrates how the FuFaIR system processes its queries Assuming that a user is interested in any document with at least three out of four given terms and another fifth term The query will be written as at_least_3(t t t3 t4 AND t5

The α value	The α cut set	at_least_3 of the α cut
3	∅	
5	{t4}	
5	{t4 t3}	
	{t4 t3 t }	× =
	{t4 t3 t t }	

Table Calculation of α cut for the sentence at_least_3

Furthermore let s assume that document d is selected with the following membership values:

$\mu_t =$

$\mu_t = 5$

$\mu_{t_3} = 5$

$\mu_{t_4} = 3$

$\mu_{t_5} = 4$

First for computing the left part of the sentence at_least_3(t t t3 t4 we calculate the fuzzy set induced by the corresponding document d which will be: Cd = { 5 5 3 3 4} Table lists the calculation of α cut for the sentence at_least_3

Using the proposed method in the previous chapter for computing fuzzy quantifiers and with the consideration that the semi fuzzy quantifier at_least_3 is defined as follows:

$$at\_least\_3: P(U) \rightarrow [0,1]$$

$$at\_least\_3(X = \begin{cases} & if \quad X < 3 \\ & otherwise \end{cases}$$

Thus the overall value of the fuzzy quantifier can be computed as follows:

at_least_3(t t t3 t4 = × 3 × 5 × 5 × × =

Now the membership of the document d to the whole algebraic sentence will be as:

AND 4 = min (4 =

5. Experimental Results

The performance of the FuFaIR has been evaluated using Hamshahri corpora This corpus consists of 5 newspaper articles from Hamshari newspaper The total size of the collection is around 3 MB All the documents have been processed by first eliminating all the Persian stop words and then indexing the remaining terms No stemming has been applied to the documents An inverted index file has been created which contains the word frequencies For retrieving each query the fuzzy sets were constructed from the inverted file Totally 3 queries have been used for these experiments Since the aim of the project was to build a high precision Persian retrieval system the precision only on the first page of the results or top documents have been considered

Figure illustrates a comparison between FuFaIR and the ector Space using Lnu ltu weighting scheme [9] The Lnu ltu weighting scheme is one of the most effective weighting schemes for the ector Space model In previous experiments on Persian language also the same weighting had outperformed other types of the ector Space model and other models of retrieval The Lnu ltu version of ector Space model has two parameters to set Those are the pivot and the slope These parameters have been set to 3 3 and 75 respectively Those values are reported as the most effective values for pivot and slope parameters in [ ] To calculate the performance of each run the precision at 5 5 and document cut offs have been calculated and averaged over all 3 queries As it is seen in the figure FuFaIR outperforms the ector Space model significantly

The current result is far better than previously reported performances of other Fuzzy models for Persian text [ ] This result is also better than expected results from English text Since in English text the Lnu ltu model will outperform any Fuzzy retrieval system Therefore the current result is interesting in the sense that it differs from similar results reported for English This shows how different languages behave with respect to similar retrieval models Another interesting aspect of this work is that the obtained results for not stemmed terms are also better than what is expected from non stemmed terms in English retrieval This is also consistent with previous results obtained from experiments with Persian Language Somehow it seems stemming does not have a significant effect on the performance of retrieval models

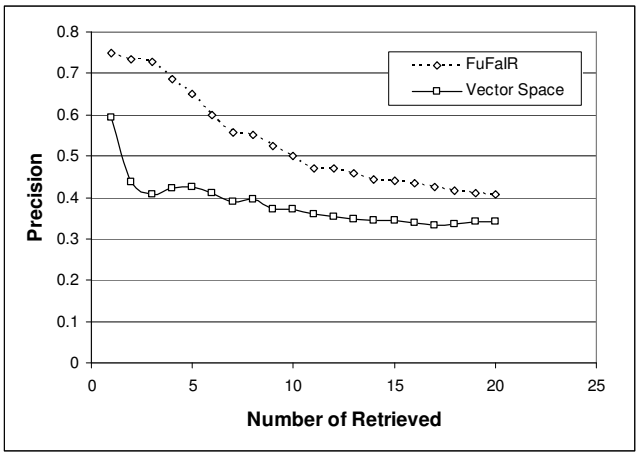

Figure 1 Precision of the FuFaIR against ector Space

6. Conclusion and Future Work

The contribution of this paper is the design implementation and testing of FuFaIR a Fuzzy retrieval system for Persian language In this work fuzzy quantifiers are also added to the original model to provide more flexibility for the system Finally the performance of the system has been compared with a high performance vector space model This comparison shows that the performance of FuFaIR is significantly better than the vector space model For the future work we will be testing different interpretation of the Fuzzy operators on the Persian corpora Another direction would be to examine the true value and contribution of a Persian stemmer in retrieval

7. References

[1] Firooz Mazhar Garamaleki, An Evaluation of Farsi Retrieval Mehtod, Msc Thesis, Department of Computer and Electrical Engineering, University of Tehran, 2003.

[2] G. Salton, E. A. Fox, and H. Wu. Extended Boolean information retrieval. Communications of the ACM, 26(12):1022–1036, 1983.

[3] G. Bordogna and G. Pasi. Linguistic aggregation operators of selection criteria in fuzzy information retrieval. International Journal of Intelligent Systems, 10(2):233–248, 1995.

[4] D.H. Kraft and Buell D.A. Fuzzy sets and generalized Boolean retrieval systems. International journal of man-machine studies, 19:45–56, 1983.

[5] Y. Ogawa, T. Morita, and K. Kobayashi. A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy sets and systems, 39:163–179, 1991.

[6] D.E. Losada, F.D. Hermida, A. Bugarin, S. Barro. Experiments on using fuzzy quantified sentences in adhoc retrieval. ACM Symposium on Applied Aomputin, 2004.

[7] Farhad Oroumchian, Firooz Mazhar Garamaleki. An Evaluation of Retrieval performance Using Farsi Text. First Eurasia Conference on Advances in Information and Communication Technology, Tehran, Iran, 29-31 October 2002.

[8] K. Taghva, J. Coombs, R. Pareda, T. Nartker. Language Model-Based Retrieval for Farsi Documents. International Conference on Information Technology: Coding and Computing (ITCC'04 2004.

[9] K. Taghva, R. Beckley, M. Sadeh. A Stemming Algorithm for the Farsi Language. International Conference on Information Technology: Coding and Computing (ITCC 2005), 2005.

[10] 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.

[11] G. Soltan, A. Wang, C. S. Yang. A Vector Space Model for Automatic Indexing, Communications of the ACM, 1975.

[12] Ricardo A. Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval, Addison Wesley Pub Co Inc /1999, ISBN: 020139829X.