2009

Fusion of retrieval models at CLEF 2008 Ad Hoc Persian Track

Zahra Aghazade
University of Tehran

Nazanin Dehghani
University of Tehran

Leili Farzinvash
University of Tehran

Razieh Rahimi
University of Tehran

Abolfazel AleAhmad
University of Tehran

See next page for additional authors

Authors
Zahra Aghazade, Nazanin Dehghani, Leili Farzinvash, Razieh Rahimi, Abolfazel AleAhmad, Hadi Amiri, and Farhad Oroumchian

Fusion of Retrieval Models at CLEF 2008 Ad hoc Persian Track

Zahra Aghazade*, Nazanin Dehghani*, Leili Farzinvash*,
Razieh Rahimi*, Abolfazel AleAhmad*, Hadi Amiri*, and
Farhad Oroumchian**

*University of Tehran, School of Electrical and Computer Engineering,
North Karegar Street, Tehran, Iran
{z.aghazadeh,n.dehghany,l.farzinvash,
r.rahimi,a.aleahmad,h.amiri}@ece.ut.ac.ir
**University of Wollongong in Dubai,
PO Box 20183, Dubai, UAE
FarhadOroumchian@uowdubai.ac.ae
http://ece.ut.ac.ir/dbrg

Abstract. Metasearch engines submit the user query to several underlying search engines and then merge their retrieved results to generate a single list that is more effective to the users information needs. According to the idea behind metasearch engines, it seems that merging the results retrieved from different retrieval models will improve the search coverage and precision. In this study, we have investigated the effect of fusion of different retrieval techniques on the performance of Persian retrieval. We use an extension of Ordered Weighted Average (OWA) operator called IOWA and a weighting schema, NOWA for merging the results. Our experimental results show that merging by OWA operators produces better MAP.

Key words: Information Retrieval, Information Fusion, Persian Text Retrieval.

1 Introduction

With the rapid growth of the volume of the data, improving the effectiveness of information retrieval systems is essential. In this study, we try to use the idea behind metasearch engines in order to improve the results of Persian information retrieval. We consider each retrieval model as a decision maker and then fuse their decisions with an OWA operator in order to increase the effectiveness. This work has been done as our first participation in the CLEF evaluation campaign [1]. For the *ad hoc* Persian track we submitted eleven experiments (runs). Our main goal was to study the effect of fusion operators and whether fusing retrieval models can bring additional performance improvements. The collection that is used in this study is a standard test collection of Persian text which is called Hamshahri and was made available to CLEF by University of Tehran [2], [3].

In Section 2, we present a brief description of the retrieval methods that have been used in our experiments. Previous experiments have demonstrated that these methods have good performance on Persian retrieval.In Section 3, OWA operator and its extensions that are used for merging the results are described. One key point in the OWA operator is to determine its associated weights. In this study, we use a weighting model which is based on Normal distribution and an IOWA extension. There are two approaches to fuse the retrieved lists: (1) Combine the results of distinct retrieval methods, (2) Combine the results of the same method but with different types of tokens. Runs that submitted to CLEF 2008 use the first approach and results show that using this approach does not lend itself to a significant improvement. It seems although the retrieval methods are different but their performances and result sets are similar. In another word, those retrieval methods provide the same vision of the data. After CLEF results were published, we tried the second approach and we were able to improve the effectiveness up to 5.67% and reached the 45.22% MAP on the test set. Section 4 describes the experiments and their results.

2 Retrieval Methods

In this work, for the purpose of fusion, we needed different retrieval methods. After studying different retrieval toolkits, finally we choose *Terrier* [4]. Different methods have been implemented in *Terrier* toolkit. Among these methods, we selected nine of them. The weighting models and a brief description of them (from [5]) are illustrated in Table 1. Table 2 depicts the result obtained from running the above nine methods described in Table 1 on the training set of queries.

3 OWA Fuzzy Operator

This section describes the Order Weighted Average (OWA) operator, normal distribution-based weighting and IOWA extension.

3.1 OWA Definition

An OWA operator of dimension n is a mapping $OWA : R^n \rightarrow R$, that has an associated n vector $w = (w_1, w_2, ..., w_n)^T$ such that $w_j \in [0, 1]$ and $\sum_{j=1}^{n} w_j = 1$. Furthermore,

$$OWA(a_1, a_2, ..., a_n) = \sum_{j=1}^{n} b_j w_j \tag{1}$$

where b_j is the j^{th} largest element of the collection of the aggregated objects $a_1, a_2, ..., a_n$ [6].

Table 1. A description of retrieval methods

Weighting Model	Description
BB2	Bose-Einstein model for randomness, the ratio of two Bernoulli's processes for first normalization, and Normalization 2 for term frequency normalization
BM25	The BM25 probabilistic model
DFR_BM25	This DFR model, if expanded in Taylor's series, provides the BM25 formula, when the parameter c is set to 1.
IFB2	Inverse Term Frequency model for randomness, the ratio of two Bernoulli's processes for first normalization, and Normalization 2 for term frequency normalization
In_expB2	Inverse expected document frequency model for randomness, the ratio of two Bernoulli's processes for first normalization, and Normalization 2 for term frequency normalization
In_expC2	Inverse expected document frequency model for randomness, the ratio of two Bernoulli's processes for first normalization, and Normalization 2 for term frequency normalization with natural logarithm
InL2	Inverse document frequency model for randomness, Laplace succession for first normalization, and Normalization 2 for term frequency normalization
PL2	Poisson estimation for randomness, Laplace succession for first normalization, and Normalization 2 for term frequency normalization
TF_IDF	The tf*idf weighting function, where tf is given by Robertson's tf and idf is given by the standard Sparck Jones' idf

3.2 IOWA

An IOWA operator is defined as follows:

$$IOWA(< u_1, a_1 >, < u_2, a_2 >, ..., < u_n, a_n >) = \sum_{j=1}^{n} w_j b_j \qquad (2)$$

where $w = (w_1, w_2, ..., w_n)^T$ is a weighting vector, such that $\sum_{j=1}^{n} w_j = 1$, $0 \leq w_j \leq 1$ and b_j is the a_i value of the OWA pair $< u_i, a_i >$ having the j^{th} largest u_i, and u_i in $< u_i, a_i >$ is referred to as the order inducing variable and a_i as the argument variable. It is assumed that a_i is an exact numerical value while u_i can be drawn from any ordinal set Ω [7]. The weighting vector which is used in our experiment will be defined in Section 4.

3.3 NOWA

Suppose that we want to fuse n preference values provided by n different individuals. Some individuals may assign unduly high or unduly low preference values to their preferred or repugnant objects. In such a case, we shall assign very low weights to these false or biased opinions, that is to say, the closer a

Table 2. Comparison between different weighting models

Weighting Model	MAP	R-Precision
BB2	0.3854	0.4167
BM25	0.3562	0.4009
DFR_BM25	0.3562	0.4347
IFB2	0.4017	0.4328
In_expB2	0.3997	0.4329
In_expC2	0.4190	0.4461
InL2	0.3832	0.4200
PL2	0.4314	0.4548
TF_IDF	0.3574	0.4017

preference value (argument) is to the mid one(s), the more the weight it will receive; conversely, the further a preference value is from the mid one(s), the less the weight it will have.

Let $w = (w_1, w_2, ..., w_n)^T$ be the weight vector of the OWA operator; then we define the following [8]:

$$w_i = \frac{1}{\sqrt{2\Pi}\sigma_n} e^{-[(i-\mu_n)^2/2{\sigma_n}^2]} \tag{3}$$

where μ_n is the mean of the collection of $1, 2, ..., n$, σ_n, $(\sigma_n > 0)$ is the standard deviation of the collection of $1, 2, ..., n$. μ_n and σ_n are obtained by the following formulas, respectively:

$$\mu_n = \frac{1}{n}\frac{n(n+1)}{2} \tag{4}$$

$$\sigma_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(i-\mu_n)^2} \tag{5}$$

Consider that $\sum_{j=1}^{n} w_j = 1$ and $0 \leq w_j \leq 1$ then we have:

$$w_i = \frac{\frac{1}{\sqrt{2\Pi}\sigma_n} e^{-[(i-\mu_n)^2/2{\sigma_n}^2]}}{\sum_{j=1}^{n} \frac{1}{\sqrt{2\Pi}\sigma_n} e^{-[(j-\mu_n)^2/2{\sigma_n}^2]}} = \frac{e^{-[(i-\mu_n)^2/2{\sigma_n}^2]}}{\sum_{j=1}^{n} e^{-[(j-\mu_n)^2/2{\sigma_n}^2]}} \tag{6}$$

4 Experiment

For the experiments, CLEF has obtained the standard Persian test collection which is called Hamshahri. Hamshahri collection is the largest test collection of Persian text. This collection is prepared and distributed by University of Tehran. The third version of Hamshahri collection is 660 MB in size and contains more than 160,000 distinct textual news articles in Persian [9]. There were 50 training queries with their relevance judgments and 50 test queries prepared for the Persian *ad hoc* track. For the CLEF, we choose nine methods of document retrieval described above and fuse the top hundred retrieved results from each of them. The evaluation of the single IR models are depicted in Table 2.

We use OWA operator based on normal distribution weighting for merging the lists. In this problem, we have nine decision makers, so the weighting vector is as the following:

$$n = 9, \mu_9 = 5, \sigma_9 = \sqrt{\frac{20}{3}}, ornes(w) = 0.5, disp(w) = 2.1195, \tag{7}$$

$$w = (0.0506, 0.0855, 0.1243, 0.1557, 0.1678, 0.1557, 0.1243, 0.0855, 0.0506)^T \tag{8}$$

The precision-recall diagram obtained after submitting the OWA run to CLEF is illustrated in figure 1. IOWA extension was also tested. We used 50 training

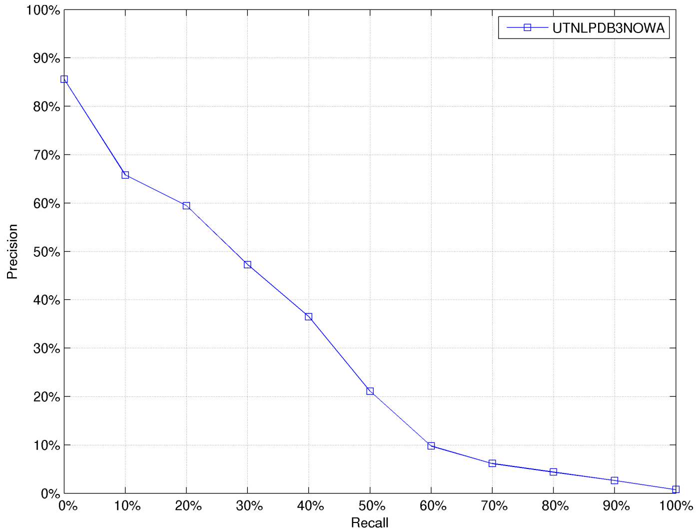

Fig. 1. The result of running NOWA published by CLEF 2008

queries in order to calculate the weighting vector for this method. We ran the nine selected retrieval methods on the collection. The following weighting vector is obtained by using the average precision of each method as its weight. These percisions are obtained from Table 2:

0.4167/3.8409, 0.4009/3.8409, 0.4347/3.8409, 0.4328/3.8409, 0.4329/3.8409, 0.4461/3.8409, 0.42/3.8409, 0.4548/3.8409, 0.402/3.8409 (3.8409 is the sum of the obtained average precisions)

Figure 2 illustrates the precision-recall diagram of IOWA run with the above weighting vector.

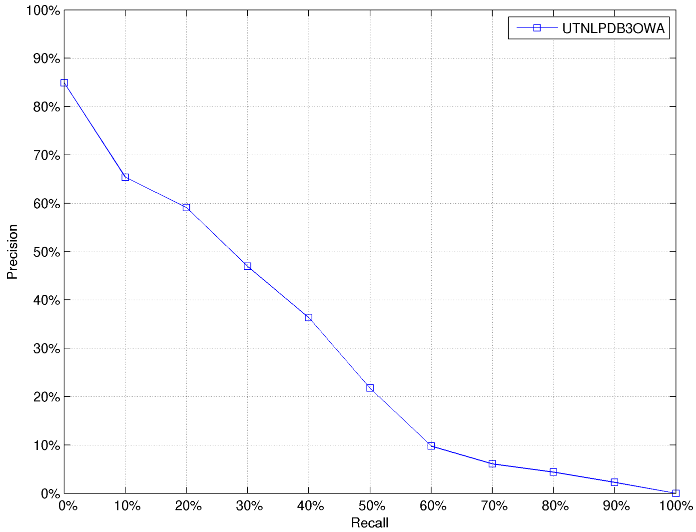

Fig. 2. The result of running IOWA published by CLEF 2008

5 Analyzing the Results and More Experiments

We submitted top hundred retrieved documents for our runs to CLEF, while CLEF evaluates the results by top thousand documents which decreased average precision about 10% in average. Therefore, in future we intend to calculate our Precision-Recall charts and other measurements based on the retrieved documents. The results published by CLEF for our fusion runs show that using fusion techniques on these methods does not yield to improved results over the individual methods. By analyzing the lists obtained from the retrieval methods, we observed that these result lists for these nine different methods have high overlap among them. On the other hand, fusion methods work well when there are significant differences between decision makers. Therefore, we have concluded that although the methods are different they are not significantly different from each other and basically they provide the same view of the collection.

After the CLEF results were published, we decided to investigate the second approach for fusion and looked the effect of different tokens in retrieval. For this purpose we chose a vector space model and ran it on the training set three times with three different types of tokens namely 4-grams, stemmed single terms and unstemmed single terms. To obtaining best results, we ran PL2 method of *Terrier* toolkit on 4-gram terms, Indri of Lemur toolkit [10] on stemmed terms and TF_IDF of *Terrier* toolkit on unstemmed terms. Then we applied the above OWA methods and as shown in Table 3, we obtained 9.97% improvements over individual runs.

After that, we continued this approach and did more experiments with the CLEF test set. On the test set, this approach lead only to 5.67% improvements

Table 3. Comparison between different weighting models on the training set

Retrieval Method	MAP	R-Precision	Dif
TF_IDF with unstemmed single terms	0.4163	0.4073	
PL2 with 4-gram terms	0.4100	0.3990	
Indri with stemmed terms	0.4100	0.4183	
IOWA	0.5160	0.4928	+9.97%
NOWA	0.5030	0.4839	+8.67%

on the average precision over individual runs using NOWA method and 5.6% using IOWA method. Table 4 demonstrate the obtained results.

Table 4. Comparison between different weighting models on the test set

Retrieval Method	MAP	R-Precision	Dif
TF_IDF with unstemmed single terms	0.3847	0.4122	
PL2 with 4-gram terms	0.3669	0.3939	
Indri with stemmed terms	0.3955	0.4149	
IOWA	0.4515	0.4708	+5.6%
NOWA	0.4522	0.4736	+5.67%

6 Conclusion

Our motivation for participation in the *ad hoc* Persian track of CLEF was to investigate the influence of fusion techniques on the effectiveness of Persian retrieval methods. First we used nine retrieval methods and then fused the results by NOWA and IOWA. The obtained results showed that although there were some improvements on the overall performance but it was not significant. In the second stage, we changed our approach to use different types of tokens with the same method. To reach this goal, we focused on working with different types of terms instead of different methods. Results indicates that fusion produces better results under such circumstances although this improvement was under 10% on the training set and 6% on the test set.

In future, we will continue investigating the effects of different token types and retrieval engines on Persian retrieval and will try to fine tune an engine based on fusion.

References

1. Agirre, E., Di Nunzio, G.M., Ferro, N., Mandl, T., Peters, C., Multilingual Textual Document Retrieval (Ad Hoc). In: Evaluating Systems for Multilingual and Multimodal Information Access 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark (2008)

2. Oroumchian, F., Darrudi, E., Taghiyareh, F., Angoshtari, N., Experiments with Persian Text Compression for Web. In: 13th international World Wide Web conference on Alternate track papers & posters, pp. 478–479. ACM, New York (2004)
3. Hamshahri Corpus, `http://ece.ut.ac.ir/dbrg/hamshahri`
4. Terrier Information Retrieval Platform, `http://ir.dcs.gla.ac.uk/terrier`
5. Amati, G., van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems, 20(4), 357–389 (2002).
6. Yager, R.R.: On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Transactions on Systems, Man and Cybernetics. 18, 183–190 (1988)
7. Yager, R.R., Filev, D.P.: Induced ordered weighted averaging operators. IEEE Transactions on Systems, Man and Cybernetics–Part B. 29, 141–150 (1999)
8. Min, D., Xu-rui, Z., Yun-xiang, C.: A Note on OWA Operator Based on the Normal Distribution. International Conference on Management Science and Engineering, 537–542 (2007)
9. Amiri, H., AleAhmad, A., Oroumchian, F., Lucas, C., Rahgozar, M.: Using OWA Fuzzy Operator to Merge Retrieval System Results. In: The Second Workshop on Computational Approaches to Arabic Script-based Languages, Stanford University, USA (2007)
10. INDRI - Language modeling meets inference networks, `http://www.lemurproject.org/indri`