Blog Feed Search in Persian Blogosphere

MohammadSadegh Zahedi*
Department of Electrical and Computer Engineering, University of Tehran, Tehran. Iran
sadeghzahedi@ut.ac.ir
Abolfazl Aleahmad
Department of Electrical and Computer Engineering, University of Tehran, Tehran. Iran
aleahmad@ut.ac.ir
Masoud Rahgozar
Department of Electrical and Computer Engineering, University of Tehran, Tehran. Iran
rahgozar@ut.ac.ir
Farhad Oroumchian
Department of Computer Science and Engineering, University of Wollongong in Dubai, Dubai. UAE
farhadoroumchian@uowdubai.ac.ae

Received: 21/May/2014 Revised: 28/Jun/2014 Accepted: 31/Jul/2014

Abstract
Recently, user generated content is growing rapidly and becoming one of the most important sources of information in the web. Blogosphere (the collection of blogs on the web) is one of the main sources of information in this category. User's information needs in blogosphere are different from those of general web users. So, it is necessary to present retrieval algorithms to the meet information need of blog users. In this paper, we focus on the blog feed search task. The goal of blog feed search is to rank blogs regarding their recurrent relevance to the topic of the query.

In this paper, the state-of-the-art blog retrieval methods are surveyed and then they are evaluated and compared in Persian blogosphere. Then we introduce our proposed method which is an extension of the voting model that one of the best blog retrieval methods. We have addressed the excessive dependency of the voting model on a list of initially retrieved top N relevant posts by using data fusion methods.

Evaluation of the proposed algorithm is carried out based on a standard Persian weblogs dataset with 45 diverse queries. Our comparisons show considerable improvement over existing blog retrieval algorithms. The results show 67% increase in MAP score for the CompVote data fusion method.

Keywords: Blog Feed Search; Blog Retrieval; Persian Blogosphere; Voting model.

1. Introduction

Emerging Web 2 enables the internet users to share their knowledge in fast and easy manner. This led to a large expansion of the Web content. Weblogs are one of the main technologies of Web 2 that have an important role in content generation on the Web. According to BlogPulse[1] (one of the major blog search engines), there were more than 182 Million Blogs in January 2012. Since the number of blogs is increasing extremely and the scale of the blogosphere has grown dramatically, this phenomenon cannot be ignored by the researchers [1].

Enormous number of blogs and their popularity make answering the users' information needs in the blogosphere context a challenging problem. Users' information needs in the context of the blogosphere are different from those of general web users. Based on a blog search engine query log analysis, Mishne and de Rijke divided blog queries into two categories called context and concept queries [2]. In the context queries, users are looking for named entities with special interest in new events ("find what people say about a given entity"). While in concept queries users are looking for blogs or blog posts related to one of their topic of interests. For context queries, the users have well defined events that try to access some information and discussions about them. Therefore, these behaviors of users are similar to when they looking for blog posts to read. While for concept queries, the searcher has broader information need, and this is likely to be interested in blogs that they can subscribe to.

In the context of the blogosphere, to answer of context and concept queries, different researches have been done such as: opinion retrieval, topic detection and tracking, top news stories identification, blog post search and blog feed search.

In this paper, we focus on *blog feed search*, also known as *blog retrieval* or *blog distillation* where the goal is to answer the concept queries. The *blog retrieval* task is defined in [3] as follows: "Knowing that users searching in the blogosphere often wish to identify blogs that are about a given topic *X; blog retrieval* algorithms must aim to find those blogs that are principally devoted to the topic *X* over the timespan of the collection at hand".

While blogs share some similar features with traditional web pages, they also have some distinct characteristics in that distinguish blog retrieval from typical ad hoc document retrieval:

[1] BlogPulse: http://blogpulse.co/

* Corresponding Author

- Blog retrieval is a task of ranking document collections rather than single documents.
- There exist amount of noise in the blog content and the language of blog posts is informal and conversational
- Blogs have some properties such as post’s comment or the time of the post which could be used for blog retrieval

Also in Iran, due to ease of publishing information in weblogs, a large number of internet users are active in this category of social media. According to the latest statistics reported by the popular ranking website (e.g., Alexa[1]), among top 10 websites in Iran, 3 websites belong to blog service providers. Despite the high importance of blogs in Iran, very few studies have been devoted to Persian blogs. In addition, to the best of our knowledge, there exists no research on blog retrieval in Persian blogosphere. Therefore, in this study we try to investigate and compare different blog retrieval methods in Persian blogosphere using *irBlogs* dataset [4].

Most researches in *blog feed search* have been started since 2007, when the organizers of TREC conference initiated a related task. Different methods from similar problems to blog feed search have been proposed like ad-hoc search methods, expert search algorithms or methods from resource selection in distributed information retrieval that are summarized in section 3.

One of the best *blog feed search* methods is *voting model* [5]. The voting model, calculates each blog's score based on a list of initially retrieved top N relevant posts that named R(Q) set, therefore, its accuracy depends on R(Q) set. This method assumes R(Q) contains relevant posts to the query; as a result, if a post is irrelevant in the list or a relevant post is absent in the list, the accuracy of the voting model decreases. To solve this critical problem, in this paper, combined data fusion methods are used to provide the actual relevant posts of R(Q) set.

So, the main contributions of this paper are as follows:

- Cross validation of blog feed search methods using new standard data collection in Persian blogosphere.
- The voting model is extended using combined data fusion methods to provide actual top relevant post in R(Q) set.

The rest of this paper is organized as follows: Persian blogosphere is described in section 2, a general review and classification of blog retrieval algorithms are presented in section 3, our method is presented in section 4 and the experimental setup is discussed in section 5. The results and comparisons are reported in section 6 and finally the paper is concluded in section 7.

2. Persian blogosphere

The exact number of Persian blogs is not known but based on a research in 2005, there exist 700,000 Persian blogs [6]. Also, the Persian language was introduced as the tenth most widely used language in blog writing by Technorati in 2007 [7]. Currently, Blogger, Persian Blog, Blog sky, ParsiBlog, MihanBlog, Blogfa, IranBlog and Wordpress are known to be the major blogging service providers for Iranian Internet users. Since many Iranian Internet users prefer to be anonymous bloggers, there is no exact statistics about the users’ gender, age and population. According to a research by the Harvard University in 2008 [8], most Iranian bloggers are 25-30 years old and among the blog users 24% are women and 76% are men. In [6], the authors used some samples from the blogs to study and analyze the Persian blog contents sociologically. Also, it is claimed that due to the relative freedom of the Persian blogosphere, their content reflects the society's opinion well. Kelly and Etling studied Persian blogosphere from political and cultural point of view [8]. They used automated content analysis tools to cluster Persian weblogs and claimed that the blogs cover four main areas: reformist and political, religious, Persian literature and poetry and blogs with various subjects. The study regarded the blogosphere as the fourth most widely used in the world. Also, in [9] the authors analyzed Persian weblogs from political aspect and conclude that Persian weblogs are a suitable space for Iranian to express their opinion about social matters.

On the other hand, some papers have studied the language style of the Persian blogosphere. A project in MITRE investigated morphological analysis of conversational Persian in weblogs [10]. They could detect many new words that are devised by Persian bloggers. The new words are mainly borrowed from other language like English and French, created by combining English and Persian words or postfixes, or changing already existing Persian words. In [11] the authors analyzed Persian weblogs morphologically and argue that Persian weblogs contain different formal and conversational texts and even the used formal language is different from commonly-used Persian formal language.

3. Related works

In this section, different blog retrieval algorithms will be explained concisely. The following categorization is present for the algorithms proposed previously for the blog retrieval task:

- Resource selection approaches
- Expert finding approaches
- Methods that use structural properties
- Methods that use temporal properties

3.1 Blog feed search using resource selection approaches

If we consider a blog as a collection of posts, then the problem of blog retrieval would be similar to the problem of resource selection in distributed information retrieval [12].

[1] Alexa - The Web Information Company: http://www.alexa.com

Elsas et. al., [13] propose large document (LD) and small document models (SD) for blog retrieval. The LD model regards the whole blog as a single document and calculates its relevancy score with regard to any input query. The SD model considers each blog as a collection of posts; after calculating the relevancy score of each post, the blog is scored using a combination of the individual post scores.

Seo & Croft [14] also approached blog distillation as a resource selection problem. Similar to the approach of Elsas et. al., they consider different representations of blogs: Global Representation (GR) and Pseudo-Cluster Selection (PCS). Similar to the LD model the GR model simply treats blogs as a concatenation of their posts.Also, PCS is similar to the SD model of Elsas et al., but it uses a different principle. In PCS, a blog is considered as a query-dependent cluster containing only highly ranked blog posts for the given query.

Lee et. al., [15], present global and local evidences of blog feeds to calculate blog scores, which correspond to the document-level and passage-level evidences used in passage retrieval. They estimated global evidence similar to the SD model of Elsas et al [13] by using all posts within a blog feed. Accordingly, local evidence is estimated similar to the PCS model of Seo & Croft [14] that uses highly relevant posts of a blog feed in response to a given query. They propose a series of methods for evaluating the relevance between a blog feed and a given query, using the two types of evidences.

3.2 Blog feed search using expert finding approaches

The task of expert finding aims at identifying persons with relevant expertise or experience for a given topic. So, if we consider each blog as an individual or expert and consider the content of each blog as expertise of each expert then blog retrieval is mapped into expert finding task. The only difference is that each expert writes only about his/her specialty, but a blogger may write about various subjects.

This connection to the expert search task was explored by Weerkamp et. al., [16] using a language modeling framework. They adopted two expert search models for blog distillation task; in *blogger model*, the probability of a query being generated by a given blog is estimated by representing this blog as a multinomial distribution of terms. In the *posting model*, this probability is estimated by combining the probability of each blog posts generating the query. In both models, the prior probability of choosing a given blog is considered to be uniform as well as the probability of choosing a post given its blog. Macdonald and Ounis [5] propose an approach that is similar to expert finding task. They used different voting models which have been used in [17] for expert finding. Some aggregation methods that they applied to blog distillation are used as our baselines and are discussed in more details later in section 3.

3.3 Blog feed search using structural properties

Some blog feed search methods try to use structural properties of blogs such as inter blog links. In [18] two different graphs are created based on the links which exist in blogs and the link information that is available in individual blog posts. Then they boost post scores using the in-degree of each post and the h-index of each blog. In [19] a graph is created based on blogs, posts and the input query words. Then by performing a random walk on this graph most relevant blogs for each query is extracted. In [20] the effect of content similarity between posts of a blog is investigated and the relevancy score of a post is smoothed based on the connection between the post and the other posts of the blog.

3.4 Blog feed search using temporal properties

Time has been used in blog feed search in various methods. Keikha et. al., [21] used temporal properties of posts in different ways for blog retrieval. They proposed a time-based query expansion method that selects terms for query expansion based on most relevant days for a given topic.

Nunes et. al., used temporal evidence as an extra feature of blogs in addition to their content [22]. They use temporal span and temporal dispersion as two measures of relevance over time, and show that these features can improve blog retrieval. In [23], Keikha et al proposed a framework, named TEMPER, which selects time-dependent terms for query expansion, generates one query for each point of time and calculate a distance measure based on temporal distributions.

Some models are designed to retrieve the blogs that are written about a topic more frequently. Such models show some improvement over the baseline that uses only the content of blogs [24, 25]. MacDonald and Ounis tried to capture recurring interests of blogs over time [5]. Following the intuition that a relevant blog will continue to publish relevant posts throughout the timescale of the collection, they divide the collection into a series of equal time intervals. Then blogs are scored based on their number of relevant posts in different time intervals. Keikha et. al., [26] aim to measure the stability of a blog relevance to a query over time. Their idea is that a blog which has many related posts only during one short period of time, is not highly relevant. Thus they defined the TRS (Temporal Relevance Stability) that scores a blog higher if it has more related posts in more time intervals which is said to possess more stability.

4. Our proposed methods

4.1 Motivation

Voting model has some properties that make it a suitable method for blog feed search. First, the Voting model is efficient methods because it uses individual blog posts indexing.

One of the most important part of blog feed search systems is indexing. In the context of blog feed search, there are three methods for indexing:

- Indexing individual blog posts.
- Indexing a whole blog or homepage of blog. It means concatenating each blog's posts into a single pseudo-document and indexing these pseudo-documents.
- Finally, we can consider hybrid indexing that uses both individual blog posts and single pseudo-document for indexing.

Since the voting model calculates each blog's score based on a list of initially retrieved top N relevant posts, we can use individual blog posts for indexing. So, the individual post index of the voting model allows simple incremental indexing and does not require frequent re-computations of pseudo-documents that are meant to represent an entire blog. Therefore, it has led the Voting model to be a more efficient blog feed search method. Also the Voting model can be used for blog post retrieval.

Information retrieval in the context of the blogosphere usually involves one of the following two main tasks:

- Blog post retrieval: identifying relevant blog posts
- Blog feed search: identifying relevant blogs.

Since the voting model uses individual blog posts score for calculating final score of blog, it can be used also for blog post retrieval.

Finally, the voting model is also an efficient method for blog feed search in Persian blogosphere. NasiriShargh in [27] shows that there exit some differences between Persian weblog and non-Persian language weblog. Due to cultural characteristics of Iranian bloggers, there exists high topical diversity in Persian blogosphere. For example a number of weblogs, that are categorized as "scientific" or "educational", may write about some of the common celebrations such as "Nowruz" or "Yalda" in a separate post. Therefore, a method like large document model or resource selection that considers a whole blog as a document may not have enough accuracy and other models that apply individual post scores such as voting model perform better.

Therefore, in this paper we try to extend voting model to propose efficient blog feed search methods in the Persian blogosphere.

On the other hand, there exist some problems with the voting model: The *voting model* [5], calculates each blog's score based on the posts that are present in the R(Q) set. Therefore, its accuracy depends on R(Q) set. This method assumes R(Q) contains relevant posts of a query; As a result, if a post is irrelevant in the list or a relevant post is absent in the list, the accuracy of the voting model decreases.

In the next section, we extend the voting model to solve this problem. First, the voting method is described and then the proposed solutions for this problem are explained.

4.2 Voting model for blog retrieval

The authors of [5] propose an approach which adopts the voting model of [17] for blog retrieval. It works based on a list of initially retrieved top N relevant posts for a query, named R(Q), which is supposed to be the set of probably relevant posts. Then each blog's score is calculated based on its posts that exist in R(Q). In this model, bloggers are considered as experts in different topics. A blogger that is interested in a particular topic, writes regularly about that topic and it is highly probable that his/her posts are retrieved in response to a related query. In this way, blog feed search can be modeled as a voting process: a post that is retrieved in response to a query is considered as a weighted vote for the expertise of its blogger. In [5], different fusion methods are used to aggregate the weighted votes and finally rank related blogs. *ExpCombMNZ* is the best fusion method in their experiments which is calculated as follows:

$$Score_{expCombMNZ}(B,Q) = |R(Q) \cap Post(B)| \sum_{p \in R(Q) \cap Post(B)} exp(Content\_Score(p,Q)) \quad (1)$$

Where, Post(B) is the set of posts of blog B, $|R(Q) \cap Post(B)|$ is the number of posts from blog B that exist in R(Q) and $Content\_Score(p, Q)$ is relevancy score of the content of post P regarding to query Q which is calculated using PL2 retrieval model.

4.3 Proposed method

In order to solve the problem of *voting model*, data fusion methods can be used to improve the accuracy of the R(Q) set. In other words, top 1000 relevant posts are initially retrieved using four different information retrieval models such as BM25 [28], PL2 [28], PL2F [29], LM [30]. Let R(Q)i presents the four retrieved list. Then, we form a new set of posts by using all of the R(Q)i sets and call it *New_R(Q)*. Provided that all R(Q)i sets have no posts in common, the final *New_R(Q)* would contain 1000*4=4000 posts . Then we assign a score to each of the posts in the *New_R(Q)* set using data fusion techniques, and use this new set, named *Optimal_R(Q)* instead of R(Q) in the voting method. But to calculate the score of the posts in *Optimal_R(Q)* set, the following procedure is used:

Since four different information retrieval methods are used to form R(Q)i sets, we initially normalize the scores of the posts in each list in a way that the scores fall in the range of (0-1). Then the final score of each post is computed based on different data fusion techniques. In other words, the results of the four retrieved sets are combined using formulas 3 to 5:

$$New\_R(Q) = \bigcup_{i=1}^{4} R(Q)_i \quad (2)$$

$$Score_{Combvotes}(P,Q) = |P \in New\_R(Q)| \quad (3)$$

$$Score_{CombRR}(P,Q) = \sum_{R(Q)_i \in New\_R(Q)} \frac{1}{Rank(P,R(Q)_i)} \quad (4)$$

$$Score_{CombMNZ}(P,Q) = Score_{votes}(P,Q) \sum_{R(Q)_i \in New\_R(Q)} Score(P,R(Q)_i) \quad (5)$$

In Formula 3, P represents a post and Q is the user query and $|P \in New\_R(Q)|$ is the number of times that post P appears in $New\_R(Q)$ and for computing the new score of each single post, the total frequency of the post in different R(Q)i sets is used. Each post is scored between 1 and 4; If a post is present in all the R(Q)i sets, it's score is 4. Thus a post will be scored higher if it is retrieved by more retrieval methods.

In formula 4, $Rank(P, R(Q)_i)$ specifies the rank of post P in $R(Q)_i$. In this fusion method, the new score of each post is computed by summing the inverse rank of post P in all $R(Q)_i$ lists. As a result, a post will be scored higher if it is retrieved by more methods or its ranking is higher in any of the methods.

In formula 5, $Score(P, R(Q)_i)$ shows the score of post P in $R(Q)_i$. In order to calculate the new score for each post, the total frequency of the post in all $R(Q)_i$ sets is multiplied by the total score of the post in all $R(Q)_i$ sets. So a post can be scored higher if it is retrieved by more methods and its score is also higher in the lists.

5. Experimental setup

In order to evaluate and compare the presented blog retrieval models, *irBlogs* dataset [4] is used. It is a standard dataset prepared for evaluation of blog retrieval algorithms that includes: (1) a set of blogs together with their posts, (2) a set of standard topics and (3) relevance judgments (ground truth). Table 1 shows some general information about *irBlogs* data set:

Table 1. Statistics of irBlogs collection

Number of blogs	**602671**
Number of posts	4983365
Average number of posts for a blog	8.2

5.1 Document collection

The document collection of this data set contains permalinks (html version of the blog posts) and feed (xml version). In our experiments the xml version is used.

5.2 Topic set

irBlogs contains 45 manually created queries about different subject categories in TREC standard format, consisting of a title (a few keywords), a description (a few sentences on what the topic is), and a narrative (a short story on which documents should be considered relevant or irrelevant). For our experiments, we are only interested in the title of a topic, which is comparable to a query submitted to a search engine by an end user

The collection also contains manual relevance judgments for each query in a four level scale of: highly relevant, relevant, irrelevant and spam. Table 2 shows the number of relevant and highly relevant blogs for the queries of *irBlogs*. Moreover, the Table 3 compares number of relevant weblogs for topics in *irBlogs*, *TREC2007* and *TREC2008* collections. This table shows that these collections are similar in terms of the number of relevant weblogs per topic. Note that good topics are those that have enough relevant blogs, so that different algorithms can be compared fairly based on them.

Table 2. Some statistics about relevance judgments of irBlogs

Number of query	**45**
Average high relevant blogs per query	23.28
Average relevant blogs per query	39.04
Number of high relevant blogs	1048
Number of relevant blogs	1773

Table 3. Categorization of topics based on the number of relevant weblogs in Trec and irBlogs data set

Topics with . . .	**Trec 2008**	**Trec 2007**	**irBlogs**
<5 relevant blogs	5	0	1
<10 relevant blogs	11	5	2
<20 relevant blogs	20	12	9
>100 relevant blogs	3	6	11
Between 5 and 100 relevant blogs	27	27	25

5.3 Inverted indexes

The collection is indexed by the Terrier (version 3.5) open source software package. Statistics of the created index are summarized in Table 4.

Table 4: Statistics of irBlogs index

Number of posts	**4746536**
Number of blogs	388994
number of tokens	598659118
size of vocabulary	6019379
Index size	6 GB

5.4 Evaluation and Comparison Criteria

For evaluation of our proposed method, various standard metrics are utilized that are commonly used for comparing blog retrieval algorithms. The metrics are: R-Precision, Mean Average Precision (MAP), Precision at rank 5 (P@5), Precision at rank 10 (P@10) and Normalized Discounted Cumulative Gain (NDCG). Also, in the following comparisons, MAP1 denotes MAP of a run, when only the highly relevant blogs are considered to be relevant and MAP2 denotes the MAP of a run, when both highly relevant and relevant blogs are considered to be relevant. In order to compare the proposed method with other existing methods, some of the best known blog retrieval methods are chosen. The algorithms and their parameters are listed in Table 5:

Table 5. The best blog feed search methods that exist in the literature

Name	**Parameters Value**	**Blog feed Search Method**
Blogger	β_{Bloger}= average Blog length	Blogger Model[16]
Posting	$\beta_{posingt}$= average post length	Posting model[16]
Voting	C=1	Voting Model[5]
Local	K=2,α=0.62	Local evidence [15]
Large	Feed Prior=Uniform,μ=2500	Large Document [13]
Resource Selection	M=4,π=1	Resource Selection +Diversity Penalty[14]
Temp	α=0.9	Temporal Evidence[22]

6. Experimental results

Performance of several blog retrieval methods (listed in table 5) are analyzed based on top 1000 retrieved blogs.

6.1 Evaluation and comparison of the state of the art methods

Figure 1 compares already existing blog retrieval methods of table 5 based on MAP2 for TREC 2007 data set. As it is obvious *local evidence* [15], *resources selection* [14] and *large documents model* [13] have the best accuracy. Also, Figure 2 provides a comparison in the performance of those methods based on the measures discussed before using the irBlogs dataset. One can observe that ranking of the methods are almost the same on both collections.

Figure 3 indicates the MAP1 scores of a number of queries for a more in-depth analysis. Figure 3.a indicates that for general queries, with a large number of highly relevant blogs, such as "learning to cook", "mountaineering training", "women's veil" and "stock market analysis", *large documents* and *resource selection models* perform better. This could be attributed to the fact that there are very few irrelevant posts in these highly relevant blogs. Thus, the *large documents model* that considers the whole blog as a document performs better for such queries. Figure 3.b indicates that for more specific queries, that have a few highly relevant blogs, such as "Computer M.Sc. Entrance exam", "Urmia Lake drying", "Fajr film festival" and "Tehran international book fair", other models that apply local post scores such as the *voting model* or the *local evidence model* perform better. The score of a blog is calculated based on its individual post scores in these models. Since the majority of these blogs are relevant (not highly relevant); so, a method like *large document model* assigns a lower score to these blogs due to the existence of irrelevant posts.

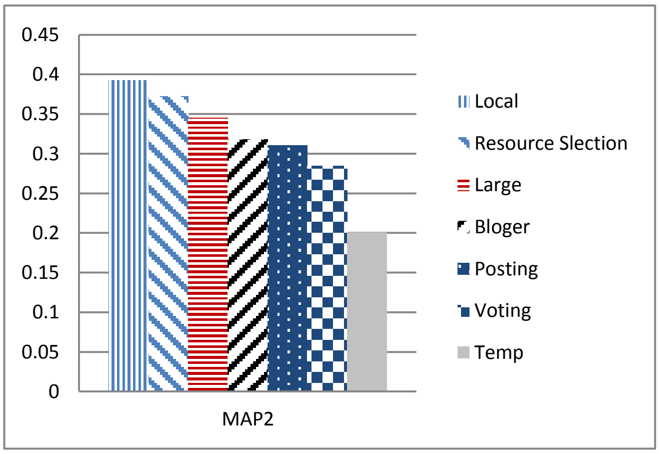


Fig 1. Evaluation results for the existing blog retrieval methods on TREC07 data set

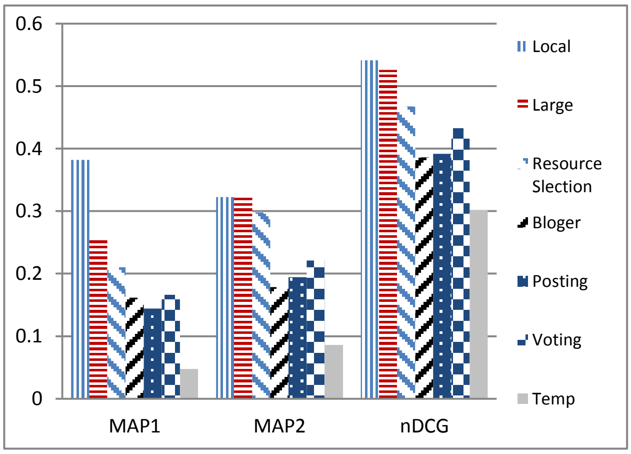
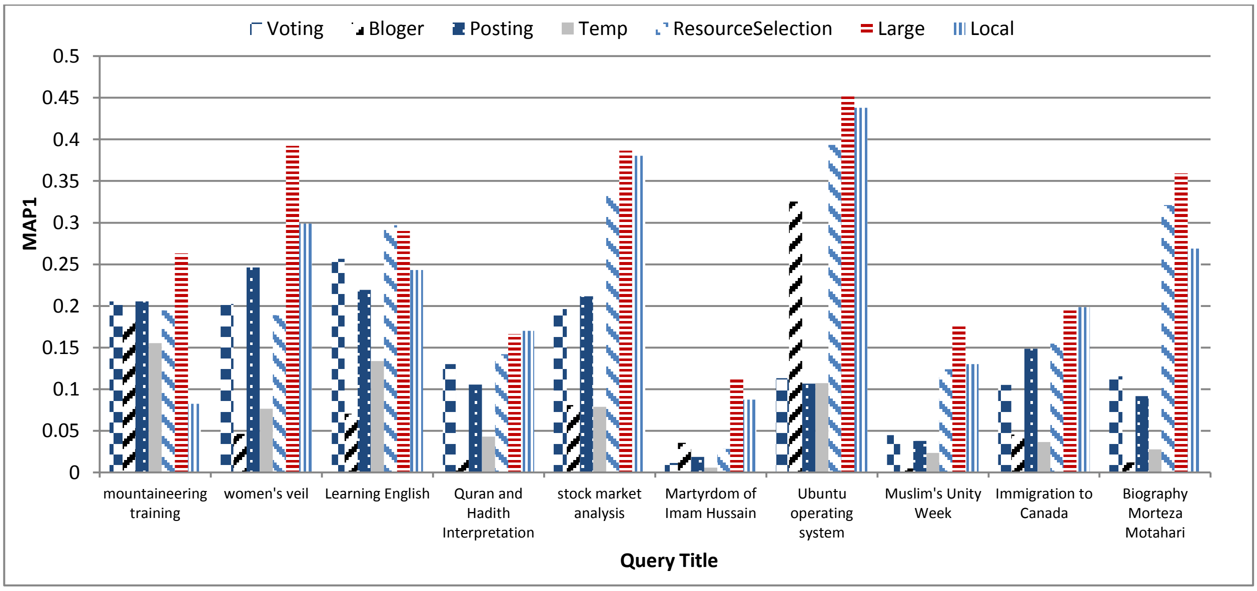
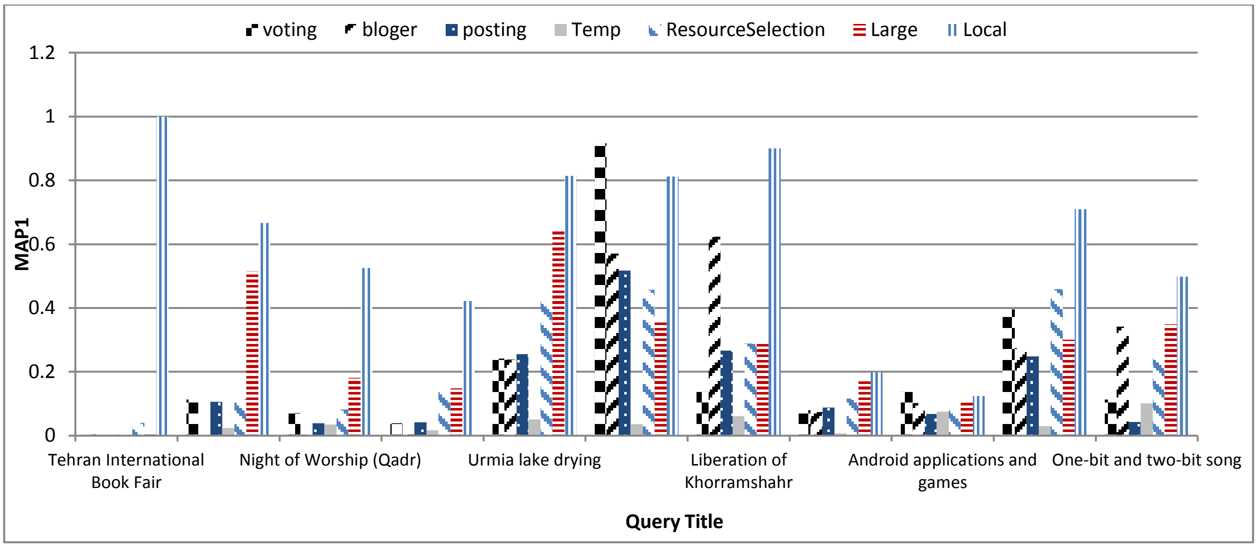


Fig 2. Evaluation results for the existing blog retrieval methods on irBlogs data set



(a)



(b)

Fig 3. Evaluation results of blog retrieval methods per topic (a) general topics (b) more specific topics of irBlogs

6.2 Evaluation and comparison of the proposed method

Table 6 summarizes the results of the *voting model* for different values of N for initially retrieved top N relevant posts of R(Q) set. As it is shown, by increasing N, the accuracy of the voting method increases but when N reaches near to 4000 the accuracy remains constant and does not increase. The best performance of the *voting method* is obtained with N = 4000 and for the rest of the experiments we have used these optimal value.

Table 6. Accuracy of voting model with different values of N for R (Q) set

	Map1	**Map2**	**NDCG**	**R_prec**	**P@5**	**P@10**
N=500	0.157	0.194	0.364	0.244	0.417	0.36
N=1000	0.166	0.221	0.410	0.264	**0.426**	0.375
N=2000	0.171	0.233	0.437	0.273	0.395	0.377
N=3000	**0.174**	0.237	0.445	0.280	0.408	**0.386**
N=4000	**0.174**	**0.239**	**0.450**	**0.281**	0.413	0.382

Figures 4 and 5 show the results of the proposed methods using data fusion for different measures. As it is shown, all the proposed methods outperform the original *voting model* based on the aforementioned measures.

Table 7 shows the improvement percentage achieved by the proposed methods in comparison with the original voting model. As one can observe, *CombVote_fusion* method gained the highest improvement rate.

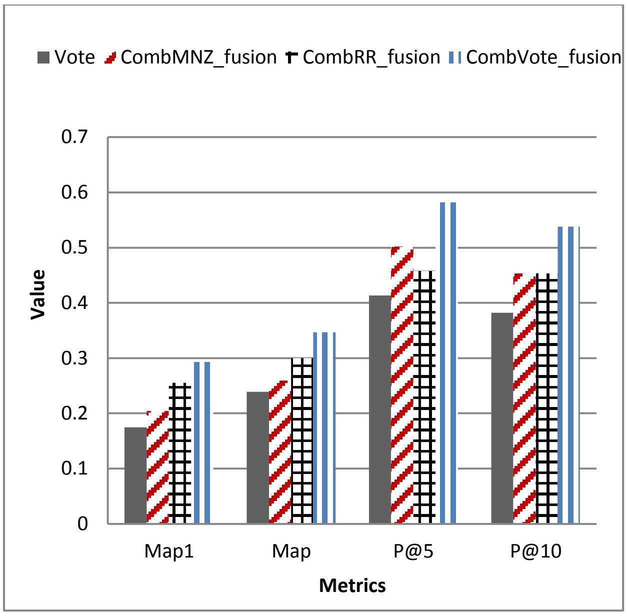


Fig 4. Evaluation results for the proposed methods and original voting model

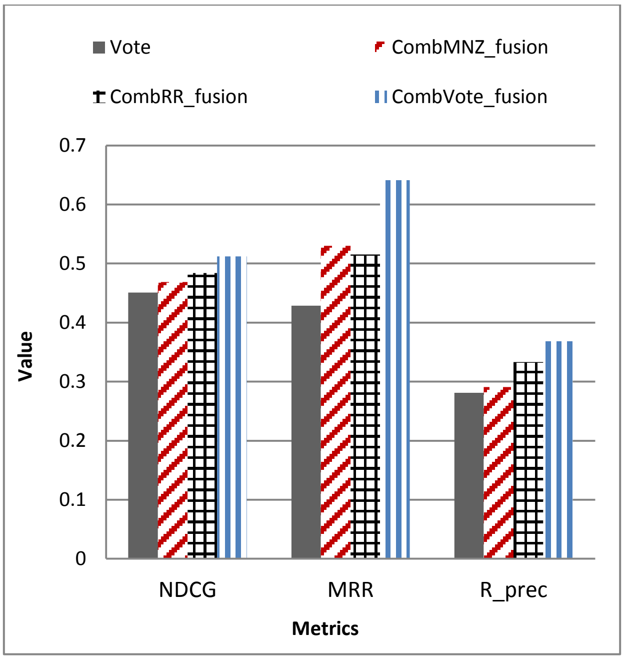


Fig 5. Evaluation results for the proposed methods and the original voting model

Table 7: Improvement percentage of proposed methods compare to the original voting model

	Percent improvement compared with Voting model				
	Map2	**Map1**	**NDCG**	**P@5**	**P@10**
CombVote_fusion	**44.98%**	**67.50%**	**13.59%**	**40.86%**	**40.71%**
CombRR_fusion	25.29%	45.99%	7.27%	10.76%	18.60%
CombMNZ_fusion	8.44%	16.70%	3.9%	21.50%	18.60%

For an in-depth analysis, Figure 6 and 7 provide per topic comparison of *CombVote_fusion* methods and the original voting model based on the MAP1 measure.

Figure 6 shows the results of the queries for which the proposed method performed better than the voting method. As it can be seen, the proposed method shows significant improvement for most of the queries. Also, Figure 7 indicates queries for which the proposed method shows lower accuracy; even for these queries, the proposed method is comparable with the original voting model. In the latter case, the original voting model turns out to have insufficient accuracy; therefore, the proposed method cannot obtain good accuracy.

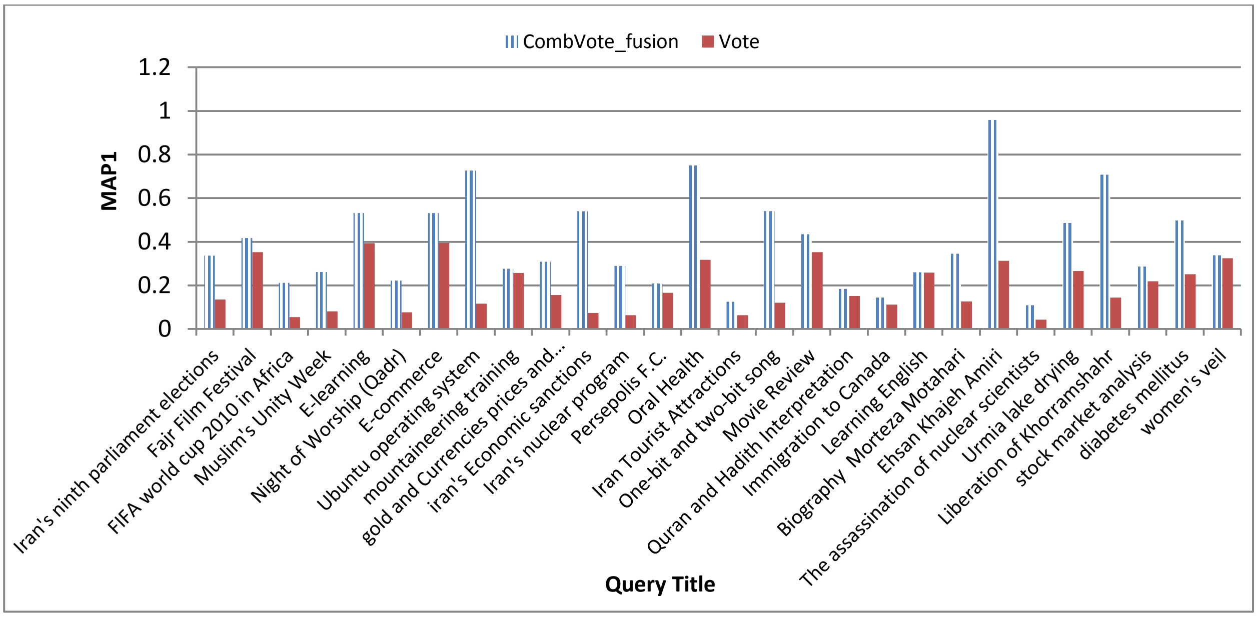


Fig 6. Per topic comparison of *CombVote_fusion* method and the original voting model based on MAP1 scores

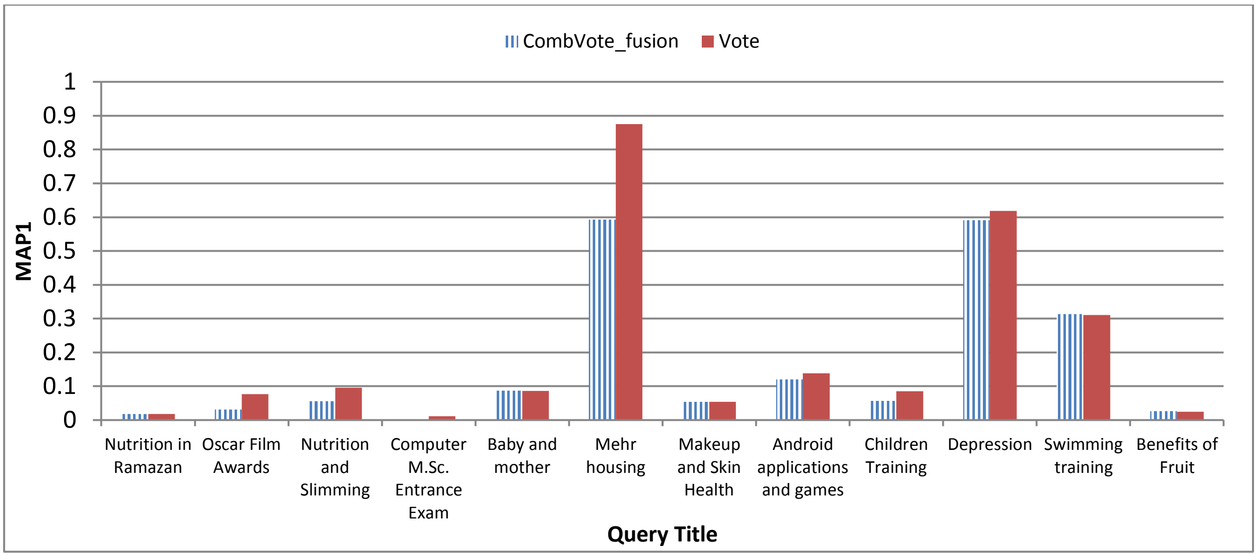


Fig 7. Per topic comparison of *CombVote_fusion* methods and the original voting model based on MAP1

7. Conclusions

In this paper, different blog retrieval methods were compared and evaluated by use of a standard Persian weblogs dataset. The results showed that the best state of the art methods shows very similar accuracy both on English and Persian blogosphere, however their ranking shows a few changes. Also, we can conclude that for general queries, methods such as large document model and resource selection model are better, but for specific queries, local method and the voting model showed better results.

Also, the voting method was developed in this paper; the problem is excessive dependency of the voting model on the R(Q) set that was solved using combined data fusion methods. Our experimental results obtained showed considerable improvement over the voting model and other blog retrieval methods.

References

[1] P. W inn. State of the Blogos pher e, intr o duction, 2008. http://technorati.com/blogging/article/state-of-the-blogosphere-introduction

[2] G. Mishne and M. de Rijke. A study of blog search. In CIR , pages 289{301, 2006.

[3] C. Macdonald, R. L. Santos, I. Ounis, and I. Soboroff, "Blog track research at TREC." In ACM SIGIR Forum, vol. 44, no. 1, pp. 58-75. ACM, 2010.

[4] A. AleAhmad, M. Zahedi, M. Rahgozar, and B. Moshiri, “irBlogs: A Standard Collection for Studying Persian Weblogs , Journal of Language Resources and Evaluation, Springer, submited on December 2013.

[5] C. Macdonald, and I. Ounis, "Key blog distillation: ranking aggregates." In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1043-1052. ACM, 2008.

[6] N. Alavi, We are Iran: Soft Skull Press, 2005.

[7] "The State of the Live Web, April 2007," http://www.sifry.com/alerts/archives/000493.html.

[8] J. Kelly, and B. Etling, “Mapping Iran’ s Online Public: Politics and Culture in the Persian Blogosphere,” Berkman Center for Internet and Society and Internet & Democracy Project, Harvard Law School, 2008.

[9] S. Golkar, “Politics in Weblogs: A Safe Space for Protest,” Iran Analysis Quarterly, vol. 2, no. 3, 2005.

[10] K. Megerdoomian, "Analysis of Farsi weblogs," MITRE Corporation, Washington DC2008.

[11] K. Megerdoomian, "Extending a Persian morphological analyzer to blogs," presented at the Zabane Farsi va Rayane [Persian Language and Computers], Tehran, Iran, 2010.

[12] J. Callan, “Distributed information retrieval,” Advances in information retrieval, vol. 5, pp. 127-150, 2000.

[13] J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell, "Retrieval and feedback models for blog feed search." In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 347-354. ACM, 2008.

[14] J. Seo, and W. B. Croft, "Blog site search using resource selection." In Proceedings of the 17th ACM conference on Information and knowledge management, pp. 1053-1062. ACM, 2008..

[15] Y. Lee, S.-H. Na, and J.-H. Lee, “Utilizing local evidence for blog feed search,” Information retrieval, vol. 15, no. 2, pp. 157-177, 2012.

[16] W. Weerkamp, K. Balog, and M. de Rijke, “Blog feed search with a post index,” Information retrieval, vol. 14, no. 5, pp. 515-545, 2011.

[17] C. Macdonald, and I. Ounis, "Voting for candidates: adapting data fusion techniques for an expert search task." In Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 387-396. ACM, 2006.

[18] C. Ribeiro, “FEUP at TREC 2010 Blog Track : Using h-index for blog ranking,” Proceedings of TREC 2010

[19] M. Keikha, M. J. Carman, and F. Crestani, "Blog distillation using random walks." In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 638-639. ACM, 2009.

[20] M. Keikha, F. Crestani, and M. J. Carman, “Employing document dependency in blog search,” Journal of the American society for information science and technology, vol. 63, no. 2, pp. 354-365, 2012.

[21] M. Keikha, S. Gerani, and F. Crestani, "Time-based relevance models." In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp. 1087-1088. ACM, 2011.

[22] S. Nunes, C. Ribeiro, and G. David, Feup at trec 2008 blog track: Using temporal evidence for ranking and feed distillation, DTIC Document, 2008.

[23] M. Keikha, S. Gerani, and F. Crestani, "Temper: A temporal relevance feedback method," Advances in Information Retrieval, pp. 436-447: Springer, 2011.

[24] B. Ernsting, W. Weerkamp, and M. de Rijke, “Language modeling approaches to blog post and feed finding,” 2007.

[25] W. Weerkamp, K. Balog, and M. De Rijke, "Finding Key Bloggers, One Post At A Time." In ECAI, pp. 318-322. 2008.

[26] M. Keikha, S. Gerani, and F. Crestani, "Relevance stability in blog retrieval." In Proceedings of the 2011 ACM Symposium on Applied Computing, pp. 1119-1123. ACM, 2011.

[27] Aidin NasiriShargh, “Suggesting and Evaluating a New Content-Based Measure to Find Similarity between Persian Weblogs”, MSc. Thesis, Sharif University of Technology, 2009.

[28] G.Amati "Probability models for information retrieval based on divergence from randomness." PhD diss., University of Glasgow, 2003.

[29] C. Macdonald, V. Plachouras, B. He, C. Lioma, and I. Ounis. University of Glasgow at WebCLEF-2005: Experiments in per-field normalisation and language specific stemming. In Proc. of CLEF 2005

[30] Zhai & Lafferty,”A Study of Smoothing Methods for Language Models Applied to Information Retrieval”.ACM Transactions on Information Systems, Vol. 22, No. 2, Pages 179—214, April 2004.

Mohammad Sadegh Zahedi received the B.S. degree in Software Engineering (2010) from the University of Zanjan. He is currently M.S. student in Department of Electrical and Computer Engineering, University of Tehran, Iran and member of Database Research Group (DBRG). His area research interests include information retrieval, social network analysis and data mining. His email address is: sadeghzahedi@ut.ac.ir

Abolfazel Aleahmad received the B.S. degree in Computer Engineering (2003) from the Bahonar University of Kerman and M.S. degree in Computer Engineering, major in Information Technology (2008) from the University of Tehran. He is currently Ph. D student in Department of Electrical and Computer Engineering, University of Tehran, Iran. He has published more than 30 journal/conference on his research findings. His area research interests include information retrieval, graph theory, social network analysis. His email address is: aleahmad @ut.ac.ir

Masoud Rahgozar, He received B.Sc. degree on Electronics Engineering from Sharif University of Technology in Tehran in 1979 and, He received Ph.D. on Database Systems from Pierre and Marie Curie University in Paris in 1987. He is Associate Professor in the Department of Electrical and Computer Engineering, University of Tehran, Iran. He has over 19 years of professional career in French software house companies as R&D manager, senior consultant and published more than 100 journal/conference and book chapters on his research findings. His current fields of interests are: Database Systems, designing CASE tools for Object Oriented programming, designing CASE tools for Database normalization and modernization of legacy applications and their environments. His email address is: rahgozar@ut.ac.ir

Farhad Oroumchian received the B.S. degree in Computer Science from National University of Iran(1984) and M.S. degree in Computer Science (1987) from the Sharif University of Technology and Ph.D. Degrees from School of Computer & Information Science, Syracuse University (1995). He is an Associate Professor in the Faculty of Computer Science and Engineering, University of Wollongong in Dubai. His research specialty is in Artificial Intelligence, Information Retrieval and Natural Language Processing. He has developed an intelligent search engine which mimics Human reasoning in order to find relevant documents. He is also a prominent researcher on Persian text processing and retrieval and multi-lingual search engines. He has published more than 90 journal/conference and book chapters on his research findings. He serves on technical/program committees of many international journals and conferences. He is also a member of several Centers of Excellence and Research groups. He has worked as project manager and consultant for many years in U.S.A and Iran. His previous posts include Associate Dean of Faculty Computer Science and Engineering, University of Wollongong in Dubai, Chair of the Department of Software Engineering at the University of Tehran, Manager of IT Services in the Faculty of Engineering, University of Tehran, and Senior Research Engineer TextWise LLC., in Syracuse New York and Database/Computer Consultant for Defelsko Inc, Ogdensburg, New York.