ASSESSMENT OF QUERY REWEIGHING, BY ROCCHIO METHOD IN FARSI INFORMATION RETRIEVAL

F. Saboori, M.S.[1]
Department of Computer Engineering
Bu-Ali Sina University, Hamedan, I. R. of Iran
email: farzaneh_2002_c@yahoo.com

H. Bashiri, M.S.[2]
Department of Computer Engineering
Bu-Ali Sina University, Hamedan, I. R. of Iran
Corresponding Author:
email: bashiri@gmail.com

F. Oroumchian, Ph.D.
Assisstant Professor
Faculty of IT Wollongong University in Dubai, Dubai, UAE
email: foroumchian@acm.org

ABSTRACT - Due to the lack of users knowledge of the collections used by search engines and in general retrieval systems, users can not express their information need appropriately in queries. In other words, they do not have enough experience to formulate their needs to find related documents. The idea of user's query expansion aims to help users to improve and correct the queries. In fact, retrieval system, regarding the feedback it receives from user at the first stage, moves the query in set space to more related documents. Different approaches in information retrieval systems have been used; however, there has not been any assessment of efficacy of query expansion in Farsi information retrieval systems. In this paper, expansion of basic model of Rocchio, assessed as the primary model to retrieve Farsi documents, has been presented. As a matter of fact, the purpose of this study is to determine the effect of a standard and basic model on query expansion to retrieve Farsi documents, so that the researchers can compare their achievements of query expansion with the findings of this paper which showed a straightforward and positive effect on Farsi document retrieval.

Keywords: Information Retrieval, Query, Query Expansion, Precision, Feedback**,** Rocchio

INTRODUCTION

One of the factors of low precision and recall of information retrieval systems, especially in the first retrieval phase, is to pose the query inefficiently. Indeed, users are novice in expressing their needs produced due to lack of their enough knowledge of collections used in information retrieval systems. For users spend a lot of time finding

their favorite documents relying a wide set of lexicons and collection terms, information retrieval systems can help to improve and correct their query[3, 4].

The commonest strategy to remake query is the feedback of relationship with user [2]. In relationship feedback cycle, user faces last retrieved documents and after examining introduces some of them as related documents. The main idea consists of choosing effective phrases related to documents introduced as relevant, so that a new query is presented based on the efficacy of words and phrases. Likewise, new query can be directed toward more relevant documents in collection space and get far from non–relevant documents. Here, three approaches can be used [1, 3]:

- Expansion of query (adding new words of related documents)
- Reweighing the words (changing the word weight based on user's feedback)
- Combination of two expansion and reweighing method

In this study, we have used word reweighing method to change the query and query expansion has considered the change of word weight. In the next studies, we are to use the method of adding related words.

ROCCHIO APPROACH

Supposing that document collection space is partitioned into two relevant and non–relevant documents for each inlet query, we will be informed that relevant documents vector has a different behavior with non-relevant one. Therefore, we must correct the query vector in order to approach the relevant documents set. This partition is divided into two sets of C_r, C_n in which C_r is all documents related to user's query out of the whole collection. The best query vector to recognize related documents out of non–related ones is identified by Formula 1:

$$\vec{q}_{opt} = \frac{1}{|c_r|} \sum_{\forall \vec{d}_j \in c_r} \vec{d}_j - \frac{1}{N - |C_r|} \sum_{\forall \vec{d}_j \notin c_r} \vec{d}_j \qquad (1)$$

Optimum Query

Lack of access to the documents of C_r makes the optimal query difficult. Rocchio model tries to make the partition more exact, regarding user's feed back[1].

$$\vec{q}_m = \alpha \vec{q} + \frac{\beta}{|D_r|} \sum_{\forall \vec{d}_j \in D_r} \vec{d}_j - \frac{\gamma}{|D_n|} \sum_{\forall \vec{d}_j \in D_n} \vec{d}_j \qquad (2)$$

Modified Query Regarding User's Feedback

D_r, D_n are relevant and non-relevant documents set respectively which are

introduced to user in the elementary retrieved.

DESIGNING AND IMPLEMENTING OF FARSI QUERY REWEIGHING BASED ON ROCCHIO METHOD

The procedure of research and access to the results before and after the user's query reweighing has been followed according to activity diagram in Figure 1.

To weigh the query words, we have used Ltu method (Formula 3). Lnu method has been used to weigh document words in the collection under assessment, explained in Formula 4. The reason to use Lnu. Ltu in weighing document and query words is the results of the study done by Oroumchian and Mazhar [5], where various weighing methods for Farsi texts have been assessed and eventually Lnu. Ltu has been introduced as the best method to weigh the words.

$$W_{iq} = \frac{\left(\ln(tf)+1.0\right)\times \ln\frac{N}{n}}{\left(slope\times NUT\right)+\left(1-slope\right)\times Pivot} \qquad (3)$$

Ltu Weighting Method in Query

Where:

w_{iq} : the weight of query words

t_f: term frequency

N: the whole number of collection documents

n: the number of documents in which the words have been used

slope: the curve slope of collection document space equal to 0.25 in our research

NUT: the number of unique terms

Pivot: the parameter of collection equals to 35.273 for the average of document length.

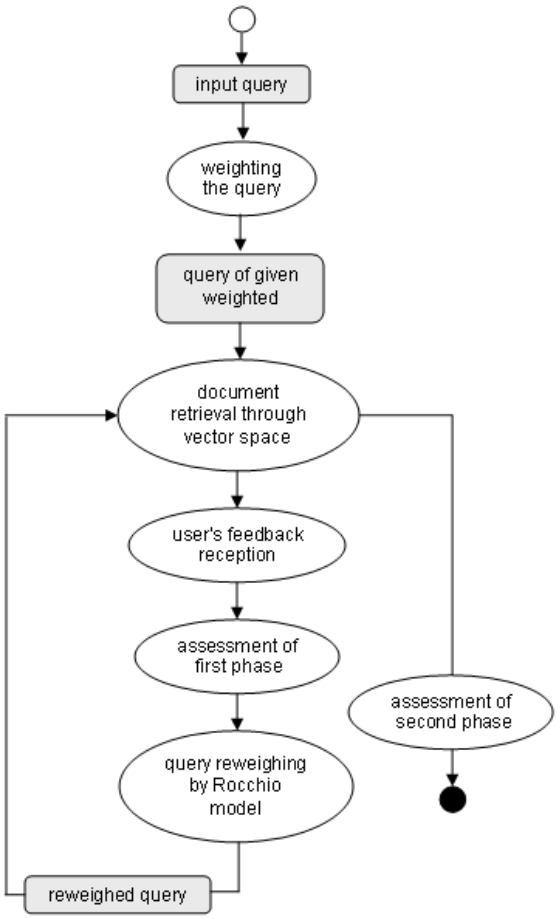

Figure 1: Activity diagram.

$$W_{ij} = \frac{\dfrac{1+\log(tf_i)}{1+\log(averagetf_j)}}{(slope \times NUT_j) + (1-slope) \times Pivot} \quad (4)$$

Ltu Weighting Method in Collection Terms

To retrieve documents, vector space model has been employed according to Formula 5:

$$Sim(q_i, d_j) = \frac{\overrightarrow{q_i} \cdot \overrightarrow{d_j}}{|\overrightarrow{q_i}| \times |\overrightarrow{d_j}|} = \frac{\sum_{k=1}^{t} w_{ki} \times w_{kj}}{\sqrt{\sum_{k=1}^{t} w_{ki}^2} \times \sqrt{\sum_{k=1}^{t} w_{kj}^2}} \quad (5)$$

Vector Space Model

SYSTEM ASSESSMENT

- COLLECTION UNDER EXPRIMENTS

The used collection in this survey was Qavanin collection, a unique collection in

assessing retrieval models of Farsi texts, consisting of the ninety years of law in Iran with a great variety of lengths. For example, a small law with one or more paragraphs has been paid attention to as a document similar to annual budget of country with all its subdivisions. To examine all retrieval models, the articles of Qavanin collection have been divided into passages consisting of a section or subsection of a law, making some paragraphs. There were formed 177089 sections in the collection [5].

A group of lawyers have supervised all inputs and system operators. Forty one questions were used to assess this system, for each of which the first 20 retrieved documents were judged. The range of scores was 0-4 by a human judge, in which zero and 4 were considered as non–relevance and full–relevance to the questions respectively. The number of words was 77889, which along with the documents formed a great collection to assess Farsi information retrieval systems. Fifteen questions out of 41 were chosen to examine the efficacy of their reweighing in document retrieval.

- ASSESSMENT BEFORE QUERY REWEIGHING

The assessment was carried out based on the precision calculation in five sets of retrieved documents in cut-offs 5, 10, 15, 20, 25, 30, 35 and 40.

The average precision was calculated in each cut–off point for all queries and used to compare retrieval before and after query reweighing. The collection documents ranked 2, 3, 4 and those ranked 0, 1 were assessed as relevant and irrelevant, respectively. The diagram of precision average of retrieval system is shown in Figure 2. This diagram is based on vector space model on 15 input questions to system in mentioned cut–off points and shows the calculated precision before the query reweighing, with a precision interval of 0.7 to 1.

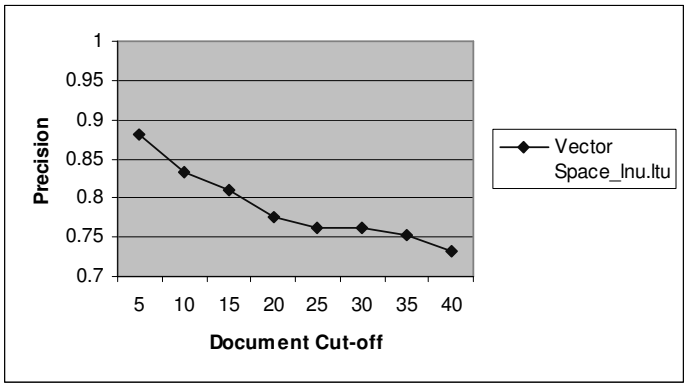

Figure 2: Diagram of average precision before query reweighing.

- ASSESSMENT AFTER QUERY REWEIGHING

In this step, all the 15 queries under experiment were corrected based on the Formula 2

and word weight changed according to feedback received from system about relevant and non–relevant documents. To enhance experiment speed and precision, feedback delivery was performed automatically, that is the examined documents in the primary retrieval were used to specify the relevance of document in the collection.

The diagram of average precision after the application of change in query and representing to system, has been shown in Figure 3.

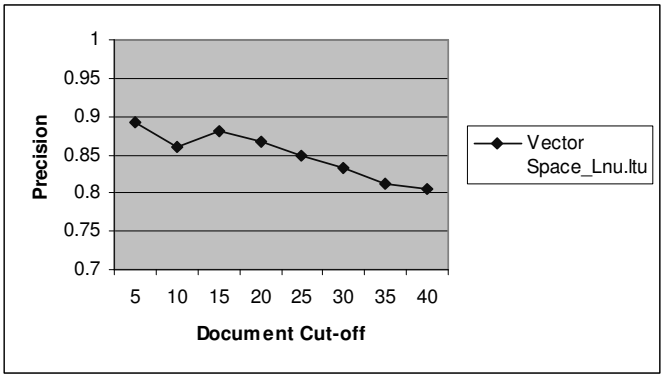

Figure 3: Diagram of average precision after query reweighing.

- COMPARISION OF AVERAGE PRECISION BEFORE AND AFTER QUERY REWEIGHING

The results of average precision in each cut–off point before and after query reweighing have been presented in Figure 4 and Table 1.

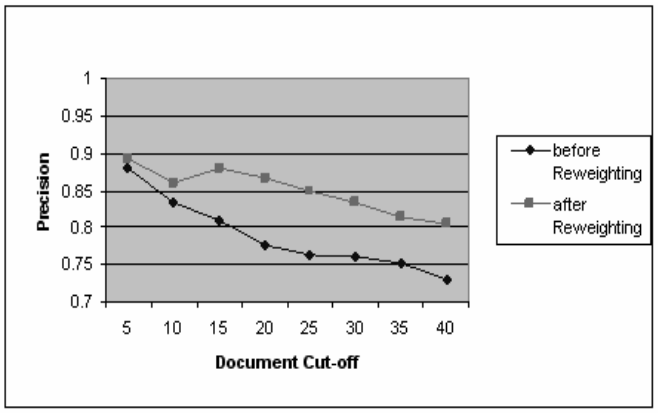

Figure 4: Comparison of average precision before and after query reweighing.

Table 1: Precision in 4 cut-offs before query reweighing.

Cut-off / Query No.	5	15	25	40
9	0.6	04667	0.44	0475
10	1	1	0.96	0.85
11	1	1	1	0.975
12	1	0.6667	0.6	0.65
13	0.8	0.9333	0.84	0.775
15	1	0.8	0.6	0.625
16	1	1	1	0.95
17	1	1	1	0.95
20	1	1	1	1
21	0.8	0.7333	0.68	0.7
28	1	1	0.92	0.825
32	1	0.9333	0.96	0.875
33	0.8	0.6	0.52	0.45
35	1	0.9333	0.8	0.8
37	0.2	0.0667	0.12	0.075
Average Precision	0.88	0.8089	0.7627	0.7317

CONCLUSION

Following findings were extracted from this study:

The average precision for each of eight cut-off points evaluated was higher after expansion. Although this effect was low in the first ranks, Figure 4 shows the increase of precision average after query expansion.

Table 2: Precision in 4 cut-offs before query reweighing.

Cut-off / Query No.	5	15	25	40
9	0.6	0.7333	0.72	0.675
10	1	1	0.96	0.95
11	1	1	1	1
12	1	0.6667	0.64	0.7
13	1	1	0.96	0.825
15	1	1	0.88	0.825
16	1	1	1	0.95
17	1	1	1	0.975
20	1	1	1	0.975
21	0.8	0.8	0.84	0.775
28	1	1	0.96	0.9
32	1	1	0.96	0.975
33	0.8	0.8	0.72	0.55
35	1	1	0.96	0.925
37	0.2	0.2	0.12	0.075
Average Precision	0.8933	0.88	0.848	0.805

The increase of average precision was observed from cut–off point of 5 onward and approaching the final cut–off point was concurrent with the increased effect of query

expansion on precision average.

Referring Tables 1 and 2 and examining queries 11 and 12, we understand that precision in each of eight cut–off points is equal to 1.

It seems logical that these queries must not re-enter in the system due to their high precision and Rocchio is used in the state to retrieve more relevant documents. Table 2 shows that query expansion didn't have any effect on retrieval precision for queries 11 and 20.

Query expansion could help to increase the average precision to 0.0614, which can be observed from the comparison in Figure 4 and Tables 1 and 2.

ENDNOTES

1. Student of Software Engineering
2. Instructor of Computer Engineering

REFERENCES

[1] Baeza-Yates, R. and Ribeiro-Neto, B., "Modern Information Retrieval." *ACM Press*, 1999.

[2] Bai, J. et al., "Query Expansion Using TERM Relationships in Language Models for Information Retrieval." *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005.

[3] Carpineto, C. et al., "An Information-Theoretic Approach to Automatic Query Expansion." *ACM Transactions on Information Systems*, 2001.

[4] Mitra, M. et al., "Improving Automatic Query Expansion." *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.

[5] Oroumchian, F. and Mazhar, F., "An Evaluation of Retrieval Performance Using Farsi Text." *Farsi Eurasia Conference on Advances in Information and Communication Technology*, Tehran, Iran, 2002.