IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 7, Issue 6, June 2020

Study of Query Expansion for Information Retrieval

Kompal

Govt. College, Panchkula

Abstract: Query expansion is a process in information retrieval where the original user query is modified to improve search results. This can involve adding synonyms, related terms, or relevant terms to enhance the query's ability to retrieve relevant documents.

Relevance feedback is a process in information retrieval where users provide feedback on the relevance of retrieved documents. The primary goal is to iteratively refine the search results based on user preferences, making the retrieval process more accurate and personalized.

I. INTRODUCTION

Whenever a query is sent, each term in the query has a score and its occurrence is calculated so that the relevance can be judged based on the total rank/score of information. In the IR process, there are different scoring methods for retrieved relevant information for a given query

II. SCORING METHODS

The various scoring methods for retrieving relevant information are as follows:

- **Term frequency (tf):**

If a term occurs frequently in a document, that document is considered more relevant to a query containing that term than other documents with fewer or no occurrences of the same term.

It can be expressed as

tf(i,j)= no. of occurrences of i in j

Occurrences where i represents term and j represents the document.

- **Inverse document frequency (idf):**

Let the document frequency df (i) be the number of documents in the collection which contains the term i and N be the total number of documents in the collection, then the inverse document frequency will be.

idf(i, j)= log(N / df(i,j))

The reason for incorporating inverse-document factor is to diminish the weight of the terms that occur very frequently in the collection and increase the weight of terms that occur rarely. In a multiple-word query, the rarer terms (those that occur in very few documents) receive more weight in determining document relevance.

These two definitions can be combined to produce a composite weight for each term in each document called tf-idf weighting as given below.

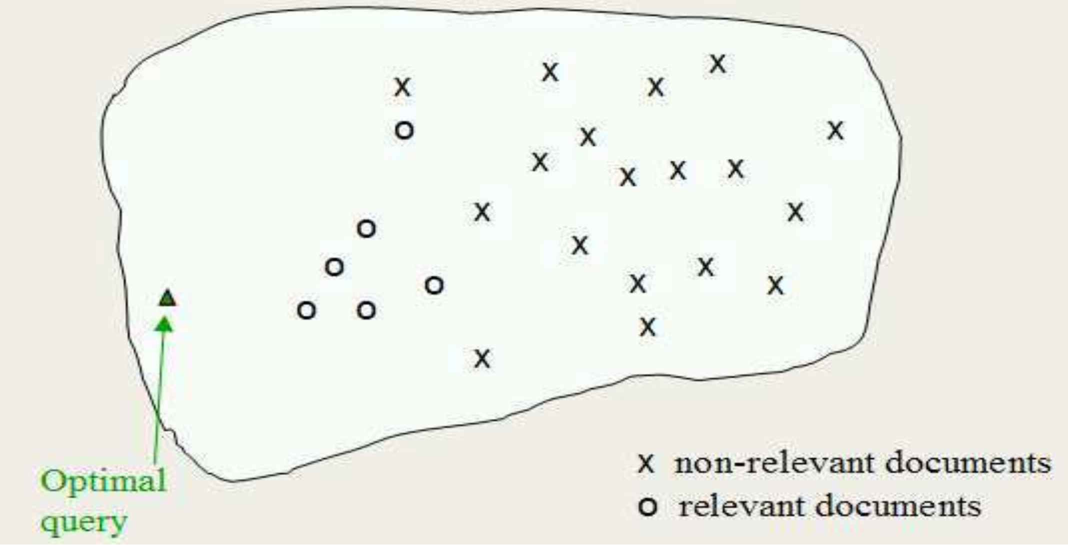
w(i,j) = tf(i,j) * idf(i,j)

Where i is the term and j is the document.

III. RELEVANCE FEEDBACK BASED ROCCHIO ALGORITHM

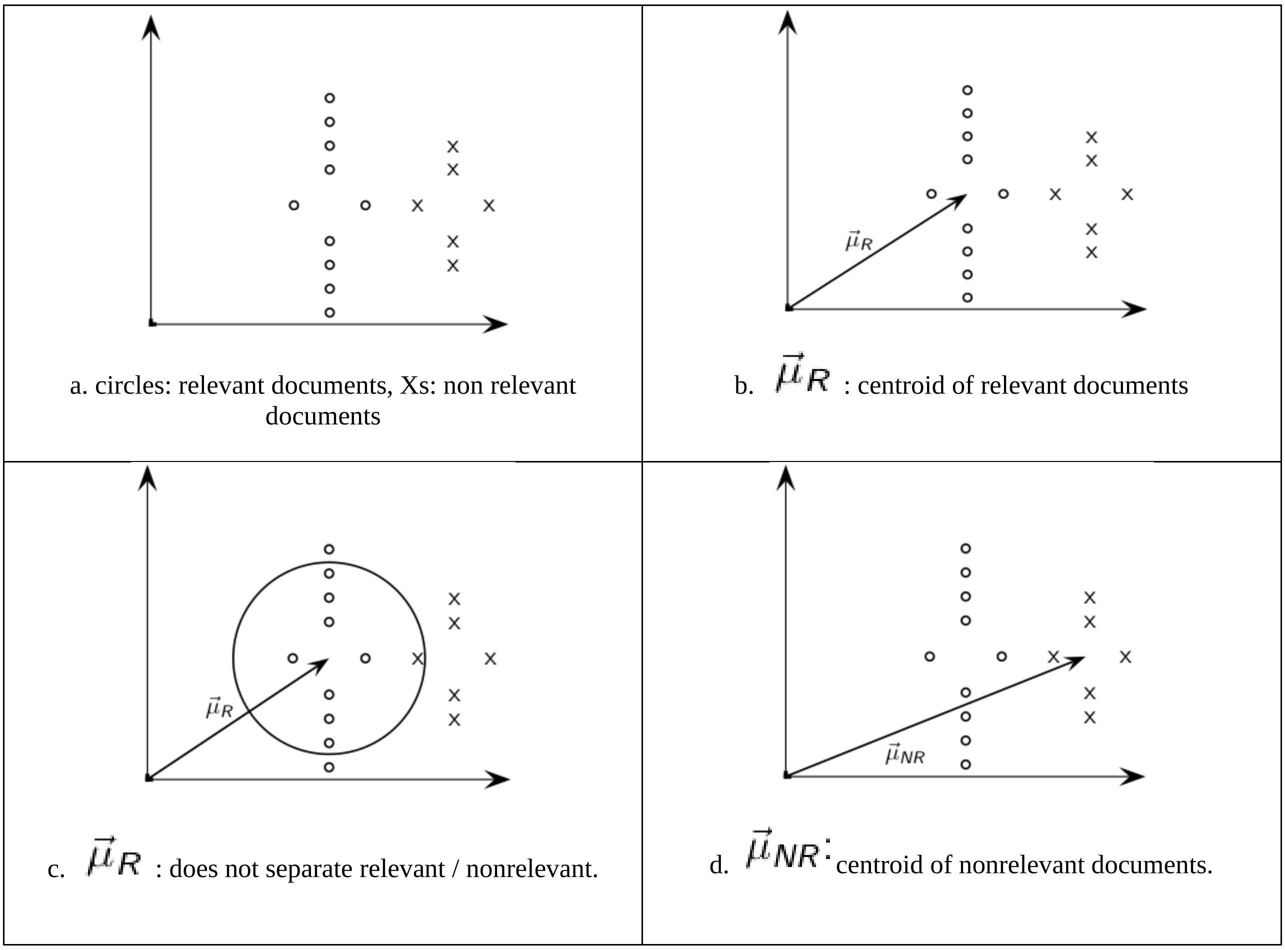
The relevance feedback concept is implemented by Rocchio Algorithm. In this algorithm, the information about relevance feedback is incorporated into the vector space model .

DOI 10.17148/IARJSET.2020.7632



The purpose is to find a query vector so that relevant documents similarity is maximized and nonrelevant documents similarity is minimized. The vector difference that exist between the relevant and nonrelevant documents centroid is the optimal query. The new query moves toward the relevant documents centroid and away from the nonrelevant documents centroid

The associated weights in Rocchio approach are responsible for moving the original query towards the related documents and away from non-related documents. The centroid of the relevant documents is moved by the difference between the two centroids. The other documents which are not used before may get selected or unselected based on their difference from the new centroid as the centroid movement explained in figure below: -

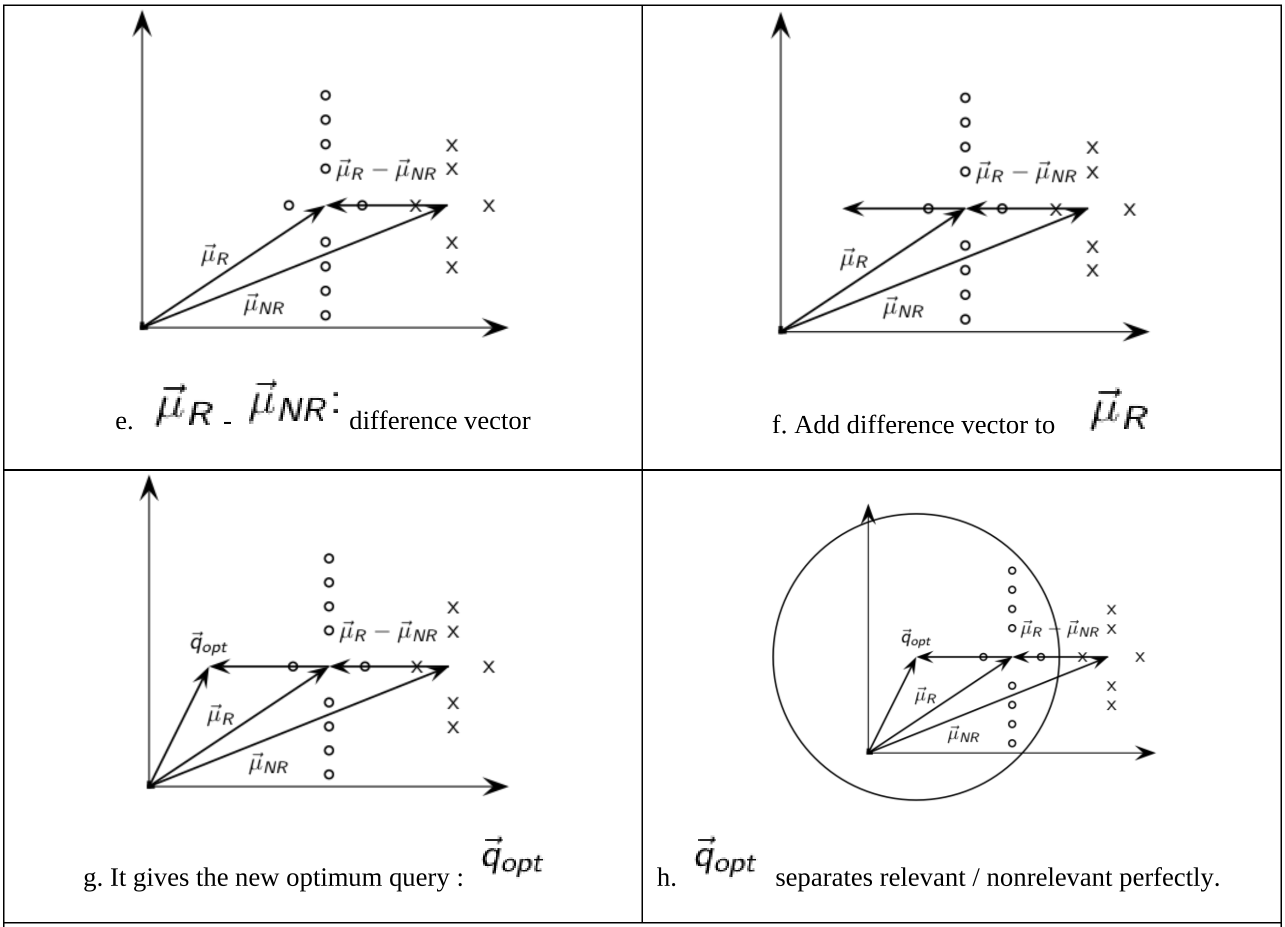


a. circles: relevant documents, Xs: non relevant documents

b. $\vec{\mu}_R$: centroid of relevant documents

c. $\vec{\mu}_R$: does not separate relevant / nonrelevant.

d. $\vec{\mu}_{NR}$: centroid of nonrelevant documents.



e. $\vec{\mu}_R$ - $\vec{\mu}_{NR}$: difference vector

f. Add difference vector to $\vec{\mu}_R$

g. It gives the new optimum query : $\vec{q}_{opt}$

h. $\vec{q}_{opt}$ separates relevant / nonrelevant perfectly.

Figure 3.1: Showing the relevant and non-relevant documents and the centroid movements using Rocchio algorithm

The purpose is to exclude the non-relevant documents from the results i.e. new query is moving away from the non-relevant documents centroid and in order to do so we have to move the relevant documents centroid by the difference existing between the two centroids (centroids of relevant and non-relevant documents). As some value is added to the centroid of relevant documents, so the new query is moving away from the relevant documents centroid also. No doubt, by doing so precision will improve as the non-relevant documents will be discarded, but recall will also decrease as the new query is moving away from the relevant documents centroid. Both recall and precision are improved by relevance feedback.

IV. CONCLUSION

For best query results, search engines have to work in faster and effective manner. For this, each crawler need to identify what is the requirement of the user. This can be taken as a profile i.e. experience of the user on search results. Many metrics like tf*idf, BM25 etc. are used to provide best documents for a given query. The query is reformulated to exclude non-relevant documents and include the relevant documents for the user query using rocchio framework.

REFERENCES

[1]. J. J. Rochio, 1971. Relevance Feedback in Information Retrieval. In The SMART Retrieval System Experiments in Automatic Document Processing, pages 313--323.

[2]. Salton and C. Buckley, “Improving retrieval performance by relevance feedback,” Readings in information retrieval, vol. 24, no. 5, pp. 355–363, 1997

[3]. G.I. Ruthven and M. Lalmas, “A survey on the use of relevance feedback for information access systems,” The Knowledge Engineering Review, vol. 18, no. 2, pp. 95–145, 2003.

[4]. Stephen Robertson and Hugo Zaragoza . The probabilistic Relevance Framework: BM25 and Beyond, Foundation and Trends in Information Retrieval,Vol.3, pg 333-389, 2009