Automatic Query Expansion for Odia Language

Santosh Kumar Behera
Department of Computer Science & Engg,
Institute Of Technical Education & Research(ITER), Bhubaneswar,
santoshbehera@iter.ac.in.

Abstract
In information retrieval field automatic query expansion is one of the most natural and successful technique to expand the query which best captures the original user intent to produce the most useful query that is more likely to retrieve relevant documents. We use pseudo relevance feedback to automate the process, so that user get improved retrieval performance without an extended interaction. Our experimental results shows that Rocchio helps to achieve better precision and recall.

Keywords: Query Expansion, Pseudo-relevance feedback, Forum for Information Retrieval Evaluation(FIRE), Precision, Recall, MAP(Mean Average Precision).

1 Intoduction

Current information retrieval systems, including web search engines, usually have a standard interface consisting of a single input box that accepts keywords. The keywords submitted by the user are matched against the collection index to find the documents that contain those keywords, which are then sorted by various methods[1]. The user queries are usually short and that the natural language is inherently ambiguous, this simple retrieval model is in general prone to errors and omissions. The relative ineffectiveness of information retrieval systems is largely caused by the inaccuracy with which a query formed by a few keywords models the actual user information need. One well known method to overcome this limitation is automatic query expansion (AQE)[1], whereby the users original query is augmented by new terms with a similar meaning.

2 Query Expansion

Query expansion (QE) is the process of reformulating an initial query to improve retrieval performance in information retrieval process. It evaluate the user input and expand the search query to match additional documents. There are different ways in which a system can help with query refinement, either fully automatically or with the user in the loop.

2 .1 Pseudo Relevance Feedback

Pseudo relevance feedback, also known as blind relevance feedback[2], provides a method for automatic local analysis. The method is to do normal retrieval to find an initial set of most relevant documents, to then assume that the top ”k” ranked documents are relevant, and finally to do relevance feedback as before under this assumption. The procedure is:

- Take the results returned by initial query as relevant results (only top k with k being between 10 to 50 in most experiments).
- Select top 20-30 terms from these documents using tf-idf weights or using some algorithm.
- Do query expansion, add these terms to query, and then match the returned documents for this query and finally return the most relevant documents.

The main objective of Relevance feedback is to improve precision and recall.

2 .2 Evaluation Measures

We describe some important performance measures of an IR system that we have used for our evaluation purpose.

Precision is the ability to retrieve top-ranked documents that are mostly relevant. Precision is the measure of how many of these hits are actually relevant.

$$Precision = |\{relevant\ documents\} \cap \{retrieved\ documents\}/\{retrieved\ documents\}|$$

Recall is the ability of the search to find all of the relevant items in the corpus. It gives how much of the content that is in our domain of interest is actually searchable?

$$Recall = |\{relevant\ documents\} \cap \{retrieved\ documents\}/\{relevant\ documents\}|$$

MAP (Mean Average Precision) is a single value measure based on the harmonic mean of precision and recall for a given set of queries.

3 Dataset

For our experiment, we used dataset of FIRE 2012 [URL: http*://www.isical.ac.in/ fire/*] for Odia. The corpus contains 32,871 documents from news domain.

Document format: It is adapted from the TREC [URL: http://trec.nist.gov] document format. Each document is stored physically in a separate file and it has 2 fields, DOCNO and TEXT. DOCNO is a unique identifier assigned to each document. TEXT field contains entire news article in plain text. The example is shown below.

```
<DOC>
<DOCNO>or_00001</DOCNO>
<TEXT>
ନୂଆଦିଲ୍ଲୀ, ତା ୧୭ା ୪:  କେନ୍ଦ୍ର ବିଦେଶ ରାଷ୍ଟ୍ରମନ୍ତ୍ରୀ ଶଶୀ ଥରୁରଙ୍କ ପାଇଁ ଏବେ କୋଟି ଆଇପିଏଲ ଫ୍ରାଞ୍ଚାଇଜ  ମହଙ୍ଗା ପଡିଛି| ପ୍ରେମିକା ତଥା ଭାବୀପତ୍ନୀ ସୁନନ୍ଦା ପୁଷ୍କରଙ୍କୁ କୋଟି ଆଇପିଏଲ ପାଇଁ ଆଥର୍ିକ ସହାୟତା ଦେଇ ଅଧିକ ଆପଣାର ହେବା ଓ ପରୋକ୍ଷରେ ଲାଭବାନ ପାଇବା ଆଶାରେ ନିଜ କ୍ଷମତାର ଅପବ୍ୟବହାର କରିଥିବା ଅଭିଯୋଗରେ ଥରୁରଙ୍କୁ ବିବାଦୀୟ ହେବାକୁ ପଡିଛି|
….
</TEXT>
</DOC>
```

Topics: Out of 50 Topics relevance judgment was available for set of 21 Topics. The topics are numbered from 176-225.

4 Pre-Processing

4.1 Corpus Pre-Processing

The odia corpus was made available as a part of FIRE 2012 is in text format. A cleaning process was applied on the news corpus to extract the <DOCNO> and <TEXT> i.e body of every document. The process of stop word removal first remove the stop words from the document. We have created our own stop word list and some of the stop words are given below:

ମାନେ, ଟିକୁ, ଙ୍କୁ, ପାଇଁ, ଠାରୁ, ଙ୍କର, ଠାରେ, ଟାଏ, ଟିଏ, ଙ୍କ, ଦ୍ୱାରା, ସକାଶେ, ଗୁଡିକ, ଗୁଡାକ, ରୁ, ରି, ରେ, ମାନଙ୍କଠାରୁ, ମାନଙ୍କଲାଗି etc.

4.2 Queries Pre-processing

In Query Pre-Processing the three components of the query i.e title, description and narration are considered. The three components are separated in the text query file with appropriate tags as shown below:

<top lang='or'>
<num>185</num>
<title>ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ଇସ୍ତଫା</title>
<desc>ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ପରେ ମୂଖ୍ୟମନ୍ତ୍ରୀଙ୍କ ଇସ୍ତଫା</desc>
<narr>ଆବଶ୍ୟକୀୟ ଦସ୍ତାବିଜଗୁଡ଼ିକରେ ଜୋଧପୁରର ଚାମୁଣ୍ଡା ଦେବୀ ମନ୍ଦିରରେ ଲୋକମାନଙ୍କର ଅକସ୍ମିକ ଅତଙ୍କଜନିତ ପଳାୟନରେ ଆହତଙ୍କ ସଂଖ୍ୟା ବିଷୟକ ତଥ୍ୟ ସାମିଲ ହେବା ଉଚିତ ।</narr>
</top>

The stop words were removed from both title and desc tag and the initial query q_0 was formulated. So from the above topic 185 we will have our initial query q_0="ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ଇସ୍ତଫା ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ପରେ ମୂଖ୍ୟମନ୍ତ୍ରୀଙ୍କ ଇସ୍ତଫା".

5 Related Work

5.1 Automatic Query Expansion using Rocchio

The Rocchio algorithm is the classic algorithm for implementing relevance feedback. It models a way of incorporating relevance feedback information into the vector space model. This mechanism introduced in and popularized by Salton's SMART system around 1970[3][2]. In real IR query context, we have a user query and partial knowledge of known relevant and non relevant documents. The algorithm proposes using the modified query $\overrightarrow{q_m}$:

$$\overrightarrow{q_m} = \alpha\overrightarrow{q_0} + \beta\frac{1}{|Dr|}\sum_{\overrightarrow{dj}\in Dr}\overrightarrow{dj} - \gamma\frac{1}{Dnr}\sum_{\overrightarrow{dj}\in Dnr}\overrightarrow{dj}$$

where q_0 is the original query vector, Dr and Dnr are the set of known relevant and non relevant documents respectively, and α, β and γ are weights attached to each term. Reasonable values might be α = 1, β = 0.75, and γ = 0.15. We have used the reasonable value for our experiments.

6 Ranking

In monolingual retrieval, the user enters a query describing the desired information. The system then return a ranked list of documents. The present work has concentrated on system that rank document according to estimated relevance to the query.

Apache Lucene, a open source free customized search engine has been used here as the base system. Lucene provides ranking based on standard IR techniques.

6.1 Lucene-based System

Lucene is an extremely rich and powerful full-text search library written in Java. Lucene can be used to provide full-text indexing across both database objects and documents in various formats (Microsoft Office documents, PDF, HTML, text, and so on). Lucene allows to add indexing and searching capabilities to different application.

6.1.1 Vector Space Model

In vector space model, documents and queries are represented as weighted vectors in a multidimensional space, where each distinct index term is a dimension, and weights are tf-idf values. The result of the searching gives a set of documents according to the highest similarity between query and documents. The similarity score[1] between query and document is expressed as:

$$Sim(q,d) = \sum_{t \in q \cap d} w_{t.q} \, . w_{t.d}$$

$w_{t.q}$ is the query term weight and $w_{t.d}$ is the document term weight. The weight is the tf.idf score.
The term frequency tf defined as number of times term t appear in a document d. Document that have more occurrences of a given term receive a higher score.

The idf(t) stands for Inverse Document Frequency. This value correlates to the inverse of docFreq (the number of documents in which the term t appears).

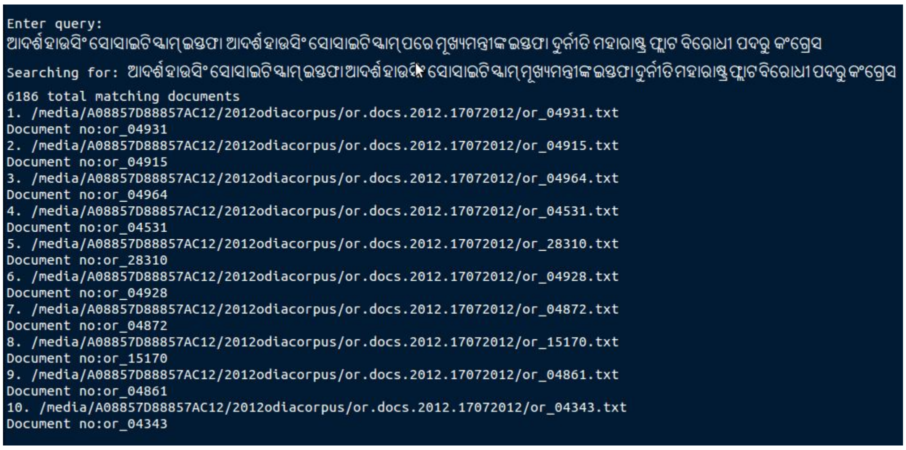
7 Experimental Results

7.1 Precision and Recall

As discussed in *section 3* out of 50 only 21 relevance judgement is available. So for some of the queries we draw the p-r curve.
As discussed in *section 3.2* the topic no. 185 result the initial query q_0="ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ଇସ୍ତଫା। ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ପରେ ମୂଖ୍ୟମନ୍ତ୍ରୀଙ୍କ ଇସ୍ତଫା।".

With the initial query q_0 1203 no of document are retrieved out of which we took top 10 as relevant document and the bottom 100 document as non relevant for Rocchio.

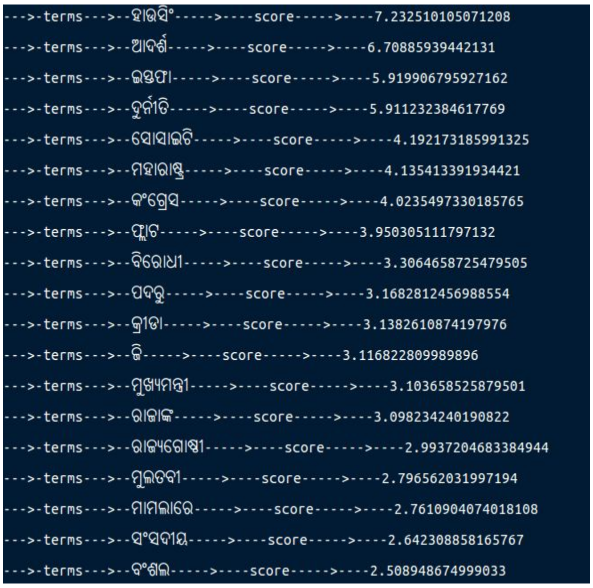
The snap shot of top 10 documents are shown in *fig 7.1*.



```
Enter query:
ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ଇସ୍ତଫା ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ପରେ ମୁଖ୍ୟମନ୍ତ୍ରୀଙ୍କ ଇସ୍ତଫା ଦୁର୍ନୀତି ମହାରାଷ୍ଟ୍ର ଫ୍ଲାଟ ବିରୋଧୀ ପଦରୁ କଂଗ୍ରେସ
Searching for: ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ଇସ୍ତଫା ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ମୁଖ୍ୟମନ୍ତ୍ରୀଙ୍କ ଇସ୍ତଫା ଦୁର୍ନୀତି ମହାରାଷ୍ଟ୍ର ଫ୍ଲାଟ ବିରୋଧୀ ପଦରୁ କଂଗ୍ରେସ
6186 total matching documents
1. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04931.txt
Document no:or_04931
2. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04915.txt
Document no:or_04915
3. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04964.txt
Document no:or_04964
4. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04531.txt
Document no:or_04531
5. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_28310.txt
Document no:or_28310
6. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04928.txt
Document no:or_04928
7. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04872.txt
Document no:or_04872
8. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_15170.txt
Document no:or_15170
9. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04861.txt
Document no:or_04861
10. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_04343.txt
Document no:or_04343
```

Fig. 7.1 Top 10 documents of topic 185 for the modified query q_m

Then we re-weight the terms using the Rocchio formula as described in *section 5.1*. After term re-weighting the some of the top terms are shown in *figure 7.2*.



```
--->-terms--->--ହାଉସିଂ---->----score---->----7.232510105071208
--->-terms--->--ଆଦର୍ଶ---->----score---->----6.70885939442131
--->-terms--->--ଇସ୍ତଫା---->----score---->----5.919906795927162
--->-terms--->--ଦୁର୍ନୀତି---->----score---->----5.911232384617769
--->-terms--->--ସୋସାଇଟି---->----score---->----4.192173185991325
--->-terms--->--ମହାରାଷ୍ଟ୍ର---->----score---->----4.135413391934421
--->-terms--->--କଂଗ୍ରେସ---->----score---->----4.0235497330185765
--->-terms--->--ଫ୍ଲାଟ---->----score---->----3.950305111797132
--->-terms--->--ବିରୋଧୀ---->----score---->----3.3064658725479505
--->-terms--->--ପଦରୁ---->----score---->----3.1682812456988554
--->-terms--->--କ୍ରୀଡା---->----score---->----3.1382610874197976
--->-terms--->--ଜି---->----score---->----3.116822809989896
--->-terms--->--ମୁଖ୍ୟମନ୍ତ୍ରୀ---->----score---->----3.103658525879501
--->-terms--->--ରାଜାଙ୍କ---->----score---->----3.098234240190822
--->-terms--->--ରାଜ୍ୟଗୋଷ୍ଠୀ---->----score---->----2.9937204683384944
--->-terms--->--ମୁଲତବୀ---->----score---->----2.796562031997194
--->-terms--->--ମାମଲାରେ---->----score---->----2.7610904074018108
--->-terms--->--ସଂସଦୀୟ---->----score---->----2.642308858165767
--->-terms--->--ବଂଶଲ---->----score---->----2.508948674999033
```

Fig 7.2:Top term generated by Rocchio Approach of Topic no. 185

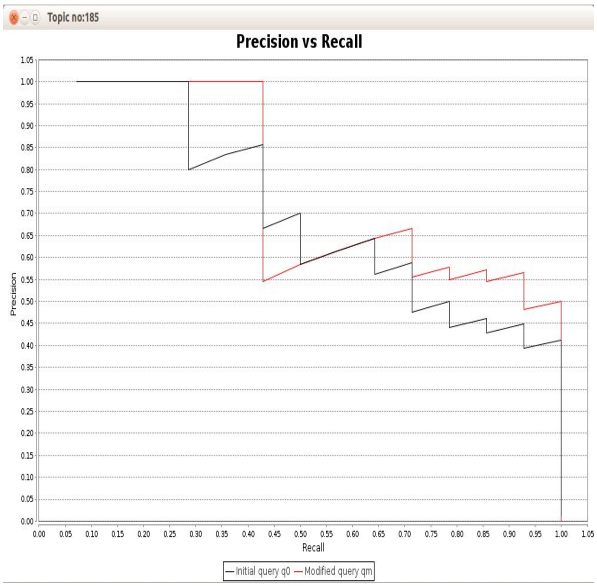


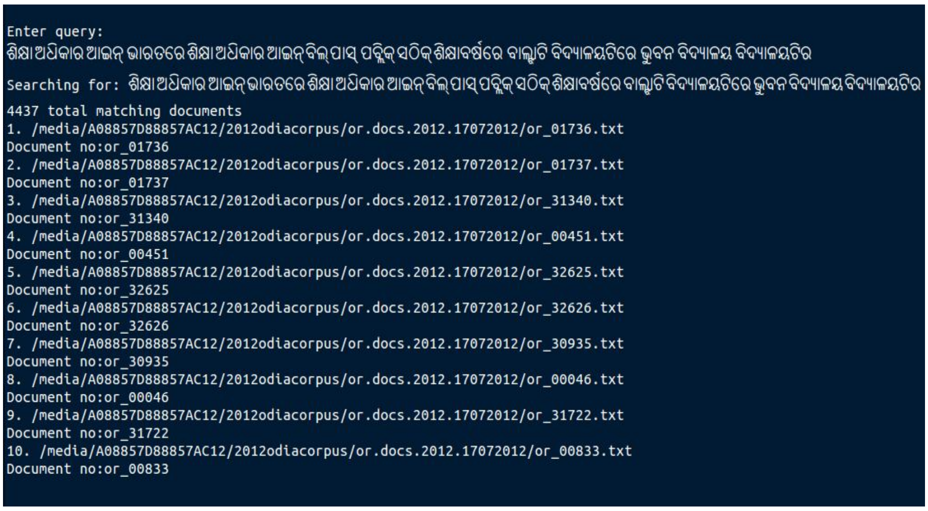
Fig 7.3:Comparative P-R curve between q_0 *and* q_m *of Topic no. 185*

The top 10 terms from Rocchio term list are added to initial query q_0. Then the modified qyery q_m="ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ଇସ୍ତଫା ଆଦର୍ଶ ହାଉସିଂ ସୋସାଇଟି ସ୍କାମ୍ ପରେ ମୁଖ୍ୟମନ୍ତ୍ରୀଙ୍କ ଇସ୍ତଫା ଦୁର୍ନୀତି ମହାରାଷ୍ଟ୍ର ଫ୍ଲାଟ ବିରୋଧୀ ପଦରୁ କଂଗ୍ରେସ ".

Then we draw the comparative p-r curve before query expansion i.e initial query q_0 and after query expansion i.e modified query q_m as shown in *fig 7.3*.

Similarly after query pre-processing topic no. 189 will result the initial query q_0=" ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ଭାରତରେ ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ବିଲ୍ ପାସ୍".

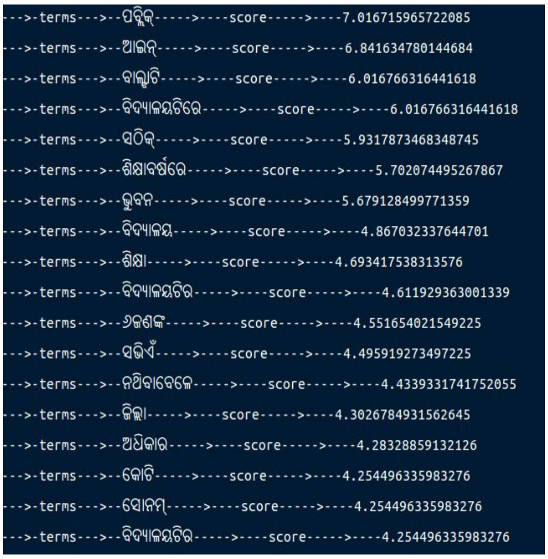
With the initial query q_0 3850 no. of document are retrieved out of which we took top 10 as relevant document and the bottom 100 document as non relevant for Rocchio.
The snap shot of top 10 documents are shown in *fig 7.4*.



```
Enter query:
ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ଭାରତରେ ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ବିଲ୍ ପାସ୍ ପବ୍ଲିକ୍ ସଠିକ୍ ଶିକ୍ଷାବର୍ଷରେ ବାଲୁଟି ବିଦ୍ୟାଳୟଟିରେ ଭୁବନ ବିଦ୍ୟାଳୟ ବିଦ୍ୟାଳୟଟିର
Searching for: ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ଭାରତରେ ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ବିଲ୍ ପାସ୍ ପବ୍ଲିକ୍ ସଠିକ୍ ଶିକ୍ଷାବର୍ଷରେ ବାଲୁଟି ବିଦ୍ୟାଳୟଟିରେ ଭୁବନ ବିଦ୍ୟାଳୟ ବିଦ୍ୟାଳୟଟିର
4437 total matching documents
1. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_01736.txt
Document no:or_01736
2. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_01737.txt
Document no:or_01737
3. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_31340.txt
Document no:or_31340
4. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_00451.txt
Document no:or_00451
5. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_32625.txt
Document no:or_32625
6. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_32626.txt
Document no:or_32626
7. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_30935.txt
Document no:or_30935
8. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_00046.txt
Document no:or_00046
9. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_31722.txt
Document no:or_31722
10. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_00833.txt
Document no:or_00833
```

Fig. 7.4 Top 10 documents of topic 189 for the modified query q_m

Then we re-weight the term using the Rocchio formula as described in *section 5.1*. After term re-weighting the some of the top terms are shown in *figure 7.5*.



```
--->-terms--->--ପବ୍ଲିକ୍---->----score---->----7.016715965722085
--->-terms--->--ଆଇନ୍---->----score---->----6.841634780144684
--->-terms--->--ବାଲ୍ଲୁଟି---->----score---->----6.016766316441618
--->-terms--->--ବିଦ୍ୟାଳୟଟିରେ---->----score---->----6.016766316441618
--->-terms--->--ସଠିକ୍---->----score---->----5.9317873468348745
--->-terms--->--ଶିକ୍ଷାବର୍ଷରେ---->----score---->----5.702074495267867
--->-terms--->--ଭୁବନ---->----score---->----5.679128499771359
--->-terms--->--ବିଦ୍ୟାଳୟ---->----score---->----4.867032337644701
--->-terms--->--ଶିକ୍ଷା---->----score---->----4.693417538313576
--->-terms--->--ବିଦ୍ୟାଳୟଟିର---->----score---->----4.611929363001339
--->-terms--->--୬ଜଣଙ୍କ---->----score---->----4.551654021549225
--->-terms--->--ସଭିଏଁ---->----score---->----4.495919273497225
--->-terms--->--ନଥିବାବେଳେ---->----score---->----4.4339331741752055
--->-terms--->--ଜିଲ୍ଲା---->----score---->----4.3026784931562645
--->-terms--->--ଅଧିକାର---->----score---->----4.28328859132126
--->-terms--->--କୋଟି---->----score---->----4.254496335983276
--->-terms--->--ସୋନମ୍---->----score---->----4.254496335983276
--->-terms--->--ବିଦ୍ୟାଳୟଟିର---->----score---->----4.254496335983276
```

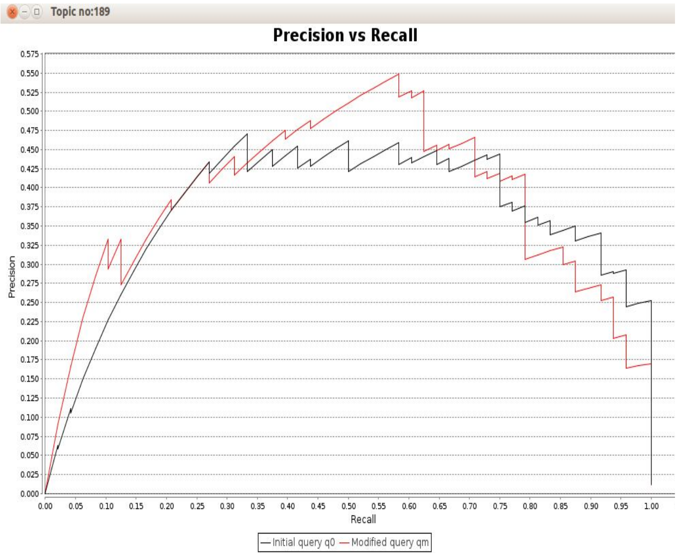


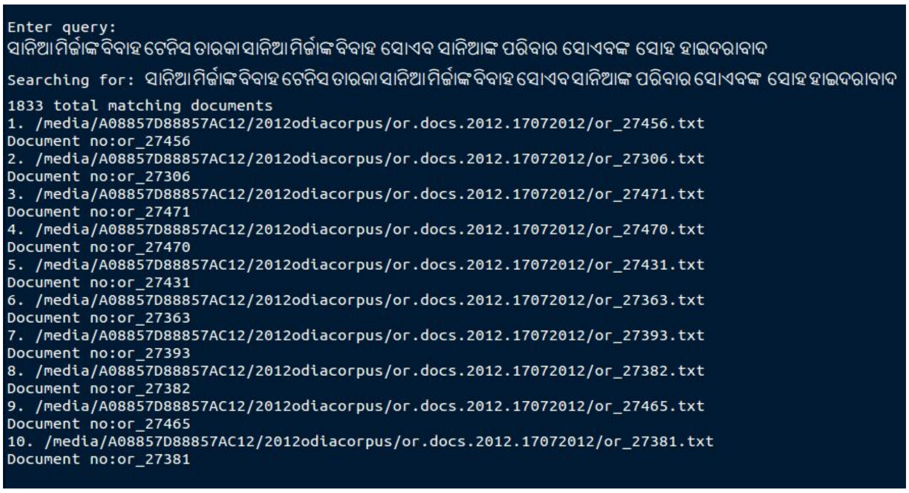
Fig 7.5:Top term generated by Rocchio Approach of Topic no. 189

Fig 7.6:Comparative P-R curve between q_0 and q_m Topic no. 189

The top 10 terms from Rocchio term list are added to initial query q_0. Then the modified qyery q_m=" ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ଭାରତରେ ଶିକ୍ଷା ଅଧିକାର ଆଇନ୍ ବିଲ୍ ପାସ୍ ପବ୍ଲିକ୍ ସଠିକ୍ ଶିକ୍ଷାବର୍ଷରେ ବାଲ୍ଲୁଟି ବିଦ୍ୟାଳୟଟିରେ ଭୁବନ ବିଦ୍ୟାଳୟ ବିଦ୍ୟାଳୟଟିର ".

Then we draw the comparative p-r curve before query expansion i.e initial query q_0 and after query expansion i.e modified query qm as shown in *fig 7.6*. Similarly after query pre-processing topic no. 207 will result the initial query q_0= "ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ ଟେନିସ ତାରକା ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ".

With the initial query q_0 1541 no of document are retrieved out of which we took top 10 as relevant document and the bottom 100 document as non relevant for Rocchio.
The snap shot of top 10 documents are shown in *fig 7.7*.

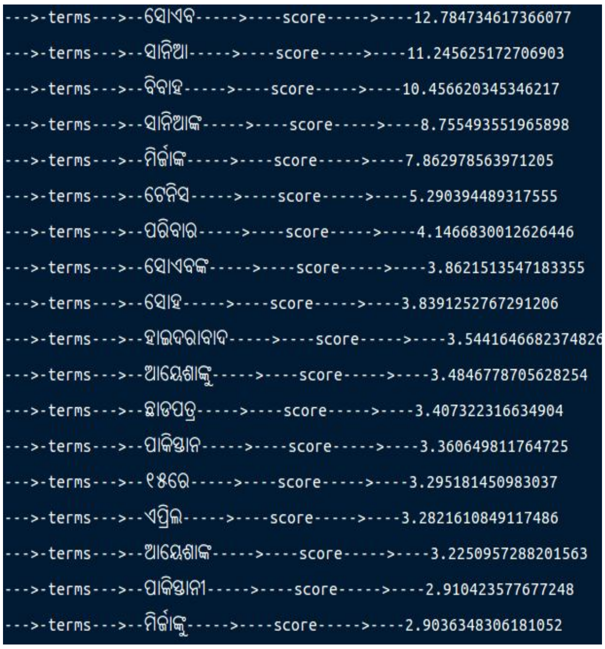


```
Enter query:
ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ ଟେନିସ ତାରକା ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ ସୋଏବ ସାନିଆଙ୍କ ପରିବାର ସୋଏବଙ୍କ ସୋହ ହାଇଦରାବାଦ

Searching for: ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ ଟେନିସ ତାରକା ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ ସୋଏବ ସାନିଆଙ୍କ ପରିବାର ସୋଏବଙ୍କ ସୋହ ହାଇଦରାବାଦ
1833 total matching documents
1. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27456.txt
Document no:or_27456
2. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27306.txt
Document no:or_27306
3. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27471.txt
Document no:or_27471
4. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27470.txt
Document no:or_27470
5. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27431.txt
Document no:or_27431
6. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27363.txt
Document no:or_27363
7. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27393.txt
Document no:or_27393
8. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27382.txt
Document no:or_27382
9. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27465.txt
Document no:or_27465
10. /media/A08857D88857AC12/2012odiacorpus/or.docs.2012.17072012/or_27381.txt
Document no:or_27381
```

Fig. 7.7 Top 10 documents of topic 207 for the modified query q_m

Then we re-weight the term using the Rocchio formula as described in *section* 5.1. After term re-weighting the some of the top terms are shown in *figure 7.8*.



```
--->-terms--->--ସୋଏବ---->----score---->----12.784734617366077
--->-terms--->--ସାନିଆ---->----score---->----11.245625172706903
--->-terms--->--ବିବାହ---->----score---->----10.456620345346217
--->-terms--->--ସାନିଆଙ୍କ---->----score---->----8.755493551965898
--->-terms--->--ମିର୍ଜାଙ୍କ---->----score---->----7.862978563971205
--->-terms--->--ଟେନିସ---->----score---->----5.290394489317555
--->-terms--->--ପରିବାର---->----score---->----4.1466830012626446
--->-terms--->--ସୋଏବଙ୍କ---->----score---->----3.8621513547183355
--->-terms--->--ସୋହ---->----score---->----3.8391252767291206
--->-terms--->--ହାଇଦରାବାଦ---->----score---->----3.5441646682374826
--->-terms--->--ଆୟେଶାଙ୍କୁ---->----score---->----3.4846778705628254
--->-terms--->--ଛାଡପତ୍ର---->----score---->----3.407322316634904
--->-terms--->--ପାକିସ୍ତାନ---->----score---->----3.360649811764725
--->-terms--->--୧୪ରେ---->----score---->----3.295181450983037
--->-terms--->--ଏପ୍ରିଲ---->----score---->----3.2821610849117486
--->-terms--->--ଆୟେଶାଙ୍କ---->----score---->----3.2250957288201563
--->-terms--->--ପାକିସ୍ତାନୀ---->----score---->----2.910423577677248
--->-terms--->--ମିର୍ଜାଙ୍କୁ---->----score---->----2.9036348306181052
```

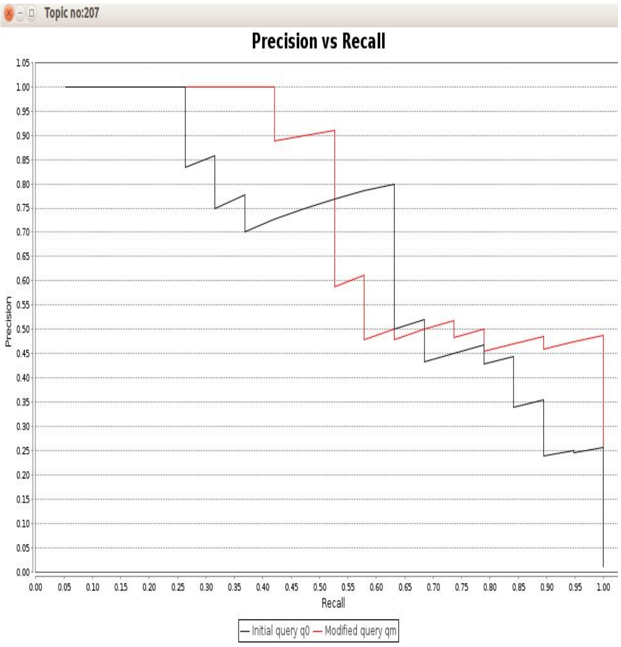


Fig 7.8:Top term generated by Rocchio Approach of Topic no. 207

Fig 7.9:Comparative P-R curve between q_0 and q_m Approach of Topic no. 207

The top 10 terms from Rocchio term list are added to initial query q_0. Then the modified qyery q_m=" ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ ଟେନିସ ତାରକା ସାନିଆ ମିର୍ଜାଙ୍କ ବିବାହ ସୋଏବ ସାନିଆଙ୍କ ପରିବାର ସୋଏବଙ୍କ ସାନିଆଙ୍କୁ ସୋହ ହାଇଦରାବାଦ". Then we draw the comparative p-r curve before query expansion i.e initial query q_0 and after query expansion i.e modified query q_m as shown in *fig 7.9*.

7.2 Mean Average Precision(MAP)

In this experiment we have calculated MAP from the result of the initial query q_0.

Retrieval model	**MAP**
TF-IDF	0.15622

Table 7.2 Mean Average Precision

8 Conclusion and Future Works

In this study, we explore ways to improve automatic query expansion via pseudo relevance feedback. We start with the initial query provided by FIRE 2012. We have used Rocchio approach drawn the comparative p-r curve for initial query q_0 and modified query q_m. The Rocchio approach gives better precision of modified query q_m as compared to initial query q_0.The main advantage of using pseudo relevance feedback is that it does not require the user input. It assume that top k documents are relevant and extract the top terms which help to automatically create the modified query.

The Rochio technique rely on several parameters. For the pseudo-relevance feedback method it is necessary to choose the number of pseudo relevant documents, the number of expansion terms, query reformulation. The retrieval performance of the over all method is usually dependent on the parameter setting.

The key aspects that need to be improved are the robustness of retrieval performance, the automatic setting of parameters, the computational efficiency of executing larger queries, and the usability of an IR system implementing Automatic Query Expansion.

9 References

[1] Carpineto C and Romano G. A survey of automatic query expansion in information retrieval. ACM Comput. Surv., 44(1):1:1–1:50, January 2012.

[2] D.Manning, Raghavan P, and Schutze H. An Introduction to Information Retrieval. Cambridge University Press, Cambridge, London, online edition, 2009.

[3] J.J.Rocchio. Relevance feedback in information retrieval. *In The SMART Retrieval System Experiments in Automatic Document Processing*, pages 313–323, Prentice Hall, 1971.

[4] Mitra M, Singhal A, and Buckley C. Improving automatic query expansion. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’98, pages 206–214, New York, NY, USA, 1998. ACM. 6.