Information Retrieval using Relevance Feedback

Shuntaro ISOGAI Shigeki OHIRA Katsuhiko SHIRAI
School of Science and Engineering, Waseda University
Okubo 3–4–1, Shinjuku-ku, Tokyo, 169–8555 Japan
isogai@shirai.info.waseda.ac.jp

Abstract

Our concern is to construct 'Audio news retrieval system' . When we think of this theme, it is important to consider how efficiently the system shows the user relevant docments as text retrieval system . Then ,it is assumed that the result of retrieval contains irrelevant documents .

So,in this reserch, we deal with the method to show the user relevant docments efficiently. This paper focuses on analyzing the ranking of relevance documents in retrieval results by using relevance feedback .

Keywords: *Relevance feedback*

1 Introduction

So far ,I have been researching on language models in large vocaburary continuous speech recognition. Recently, the research of audio news retrieval which uses the techniques of speech recognition have been receiving increasing attention. Then we are trying to construct 'Audio news retrieval system' . We approach to information retrieval in order to aaply its technique into speech recognition ,especially into language models.

When we consider to construct 'Audio news retrieval system' ,it would be the matter that how efficiently the system shows the user relevant docments .In this point ,it is assumed that the result of retrieval contains irrelevant documents .

So,in this reserch, we go with the method to show the user relevant docments efficiently. This paper concentrates on analyzing the ranking of relevance documents in retrieval results by using relevance feedback .

2 Outline of Retrieval

2.1 Data Preprocessing

The following processes are made on the abstract data for preparation.

- All the half-size (8bit character)symbols in abstract data are removed.
- Chasen ver2.02 is used for the morphological analysis of each data.No changes are applied to the program environment or to the dictionary.

  Nouns(common,proper,verbal,time,name,place), numerals,adjectives,noun prefixes,noun nature noun suffixes,and undefined words are extracted for indexing.

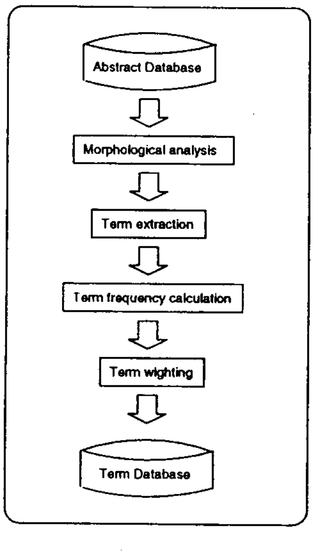

Figure 1. Indexing Method

2.2 Retrieval Model

In this research ,similarity $Sim(Q, D_d)$ between query Q and the document D_d is calculated by the following formula , based on the known as vector space model(w shows the term weight).

$$Sim(Q, D_d) = cos(Q, D_d) = \frac{\sum_{t=1}^{n} w_{q,t}w_{d,t}}{\sqrt{\sum_{t=1}^{n} w_{q,t}^2}\sqrt{\sum_{t=1}^{n} w_{d,t}^2}}$$

Terms in the sentence was scored using TF-IDF method.

$$w(t) = tf(d,t)idf(t)$$

where

$$tf(d,t)idf(t) = log(1 + tf(t))log(\frac{1+N}{df(t)})$$

N : the total number of the documents
$tf(t)$: frequency of term t in the docment
$df(t)$: the number of documents which contain the term t

2.3 Relevance Feedback

The following three formulas are known as scoring method of relevance feedback in vector space model.

$$Q_{i+1} = Q_i + \frac{1}{N^+}\sum_{j=1}^{N^+} D_j^+ - \frac{1}{N^-}\sum_{j=1}^{N^-} D_j^-$$

(Rocchio's formula)

$$Q_{i+1} = Q_i + \sum_{j=1}^{N^+} D_j^+ - \sum_{j=1}^{N^-} D_j^-$$

(Ide's formula)

$$Q_{i+1} = Q_i + \sum_{j=1}^{N^+} D_j^+ - D_1^-$$

(Ide dec-hi's formula)

where
$N+$: the number of relevant documents
$N-$: the number of non-relevant documents
D : weight vector of index terms

In this paper, only relevance document was used on relevance feedback.

$$Q_{i+1} = Q_i + \sum_{j=1}^{N^+} D_j^+ \qquad (1)$$

3 Experiment

In this experiment,retrieval is done in the same way to Figure2 with Formula(1)

Documents used in relevance feedback is followed by two types.

Documents which are ranked above x_{th} in the first normal retrieval. Besides

- x documents automatically

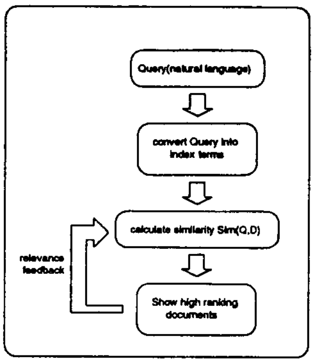
Query(natural language)
convert Query into index terms
calculate similarity Sim(Q,D)
Show high ranking documents
relevance feedback

Figure 2. Retrieval system

- documents that the user considered as relevant

In this section,the result of two queries' are shown. Large number of relevant documents are contained in both two results of first retrieval which is ranked from 1 to 100.

Figure1 presents the number of relevant documents in first retrieval.The number of relevant documents which are ranked from 1_{st} to 100_{th} are 18 (query A),and 40(query B).

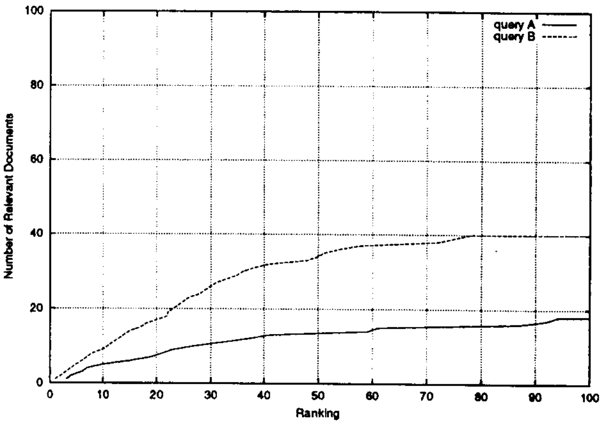

Figure 3. Number of Relevant Documents

3.1 Ingeractive Feedback

Interactive feedback is proceeded as follows. At first,the system executes first retrieval with the query,and then system shows the user high order result . The number of documents which are shown to the user are set from 1 to 100. Then, the user determines if documents are relevant. Finally, relevance feedback is proceeded with using documents which are determined as relevant by the user. (Usually,it would be greatly difficult to determine if100 documents are relevant.)

Figure4 shows the ranking of relevant documents (y) when relevance feedback is proceeded by using the relevant documents which were selected in the top x at the first retrieval by query A. In this case, relevant documents which are selected in the top 20 at the first retrieval are focused.

Figure5 shows the difference between the ranking at the first retrieval and the ranking after proceeding relevance feedback , in which we use the relevant documents which are selected in the top x at the first retrieval.In this case, relevant documents which are selected in the top 20 at the first retrieval are focused.

Similar to Figure4,Figure6 shows the ranking of relevant documents (y) when relevance feedback is proceeded by using the relevant documents which are selected in the top x at the first retrieval by query A. In this case, relevant documents which are selected in the ranking from 20 to 100 at the first retrieval are focused.

In the same way to Figure5 ,Figure7 shows the difference between the ranking at the first retrieval and the ranking after proceeding relevance feedback , in which we use the relevant documents which are selected in the top x at the first retrieval.In this case, relevant documents which are selected in ranking from 20 to 100 at the first retrieval are focused.

Figure 4 to 7 is the analysis of the retrieval results using query A. Similarly ,Figure 8 to 11 is the analysis of the retrieval results using query B.

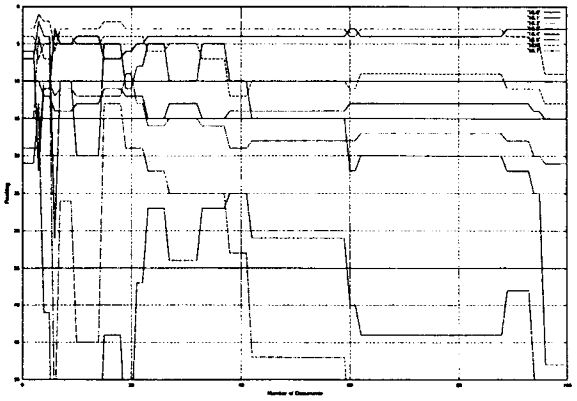

Figure 4. Ranking of relevant documents (interactive ,top20 ,query A)

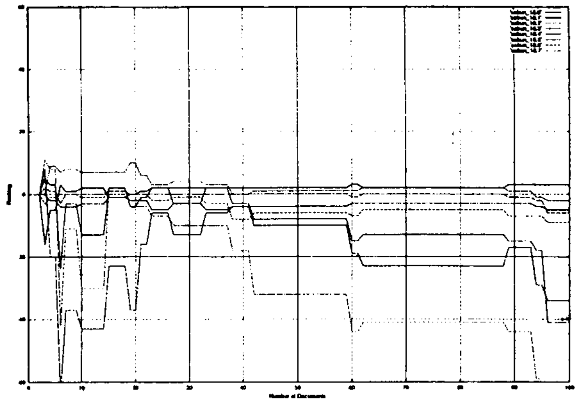

Figure 5. Ranking Difference before and after feedback(interactive,top20,query A)

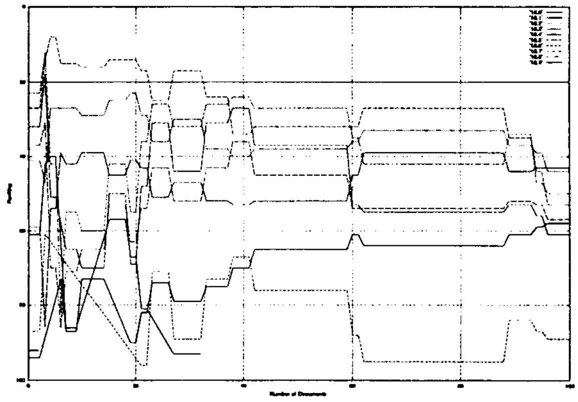

Figure 6. Ranking of relevant documents (interactive ,ranking 20 to 100 ,query A)

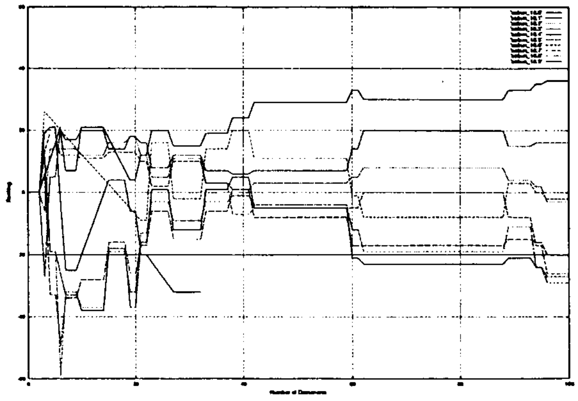

Figure 7. Ranking Difference before and after feedback(interactive,ranking 20 to 100,query A)

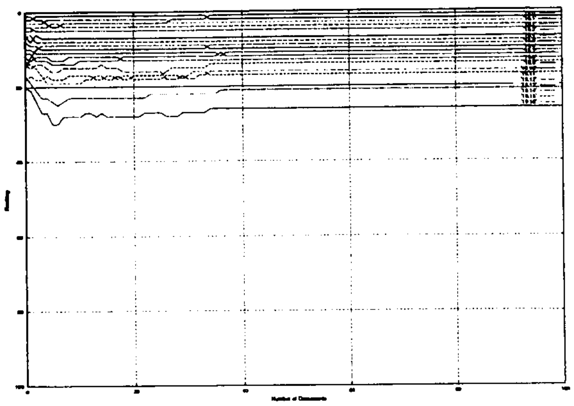

Figure 8. Ranking of relevant documents (interactive ,top 20 ,query B)

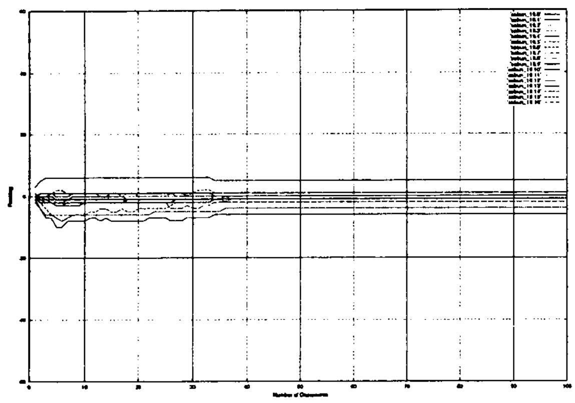

Figure 9. Ranking Difference before and after feedback(interactive,top 20 ,query B)

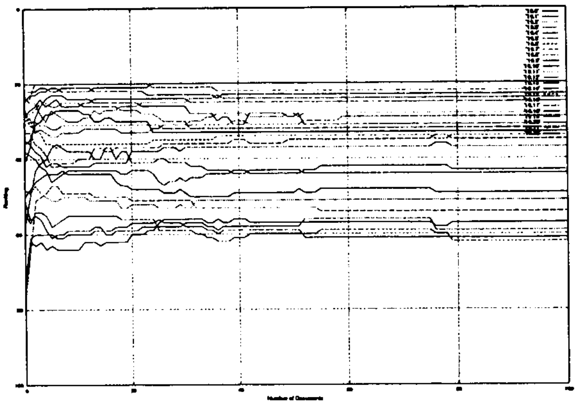

Figure 10. Ranking of relevant documents (interactive ,ranking 20 to 100 ,query B)

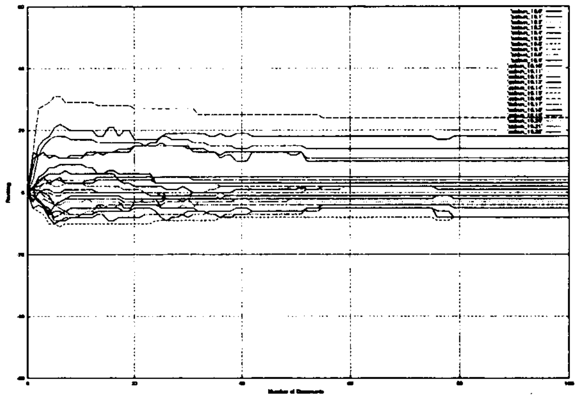

Figure 11. Ranking Difference before and after feedback(interactive,ranking 20 to 100,query B)

3.2 Automatic Feedback

Automatic feedback is proceeded as follows. At first , the system execute first retrieval with the query. And then relevance feedback is proceeded using the documents in the top ranking.

The same as section 3.1 ,the analysis of the retrieval results using query A and B are shown here.

Figure 12 shows the ranking of relevant documents (y) when relevance feedback is proceeded using the documents in the top ranking at the first retrieval by query A. In this case, relevant documents which are selected in the top 20 at the first retrieval are focused.On the other hand ,Figure 14 focuses relevant documents which are selected in the ranking from 20 to 100 at the first retrieval.

Figure13 shows the difference between the ranking at the first retrieval and the ranking after proceeding relevance feedback , in which we use the documents which are in high ranking at the first retrieval. In this case, relevant documents which are selected in the ranking from 20 to 100 at the first retrieval are focused. On the other hand ,Figure 15 focuses relevant documents which are selected in the ranking from 20 to 100 at the first retrieval.

Figure 12 to 15 is the analysis of the retrieval results using query A. Similarly ,Figure 16 to 19 is the analysis of the retrieval results using query B.

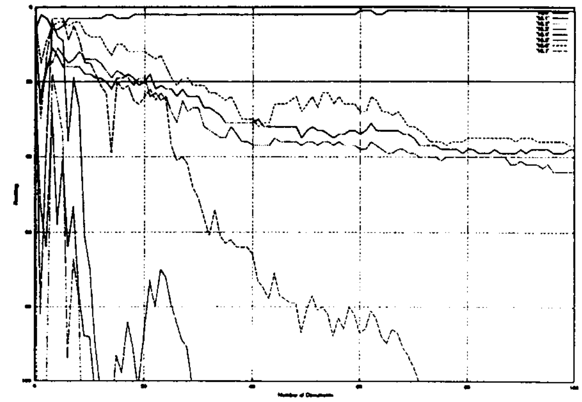

Figure 12. Ranking of relevant documents (automatic ,top20 ,query A)

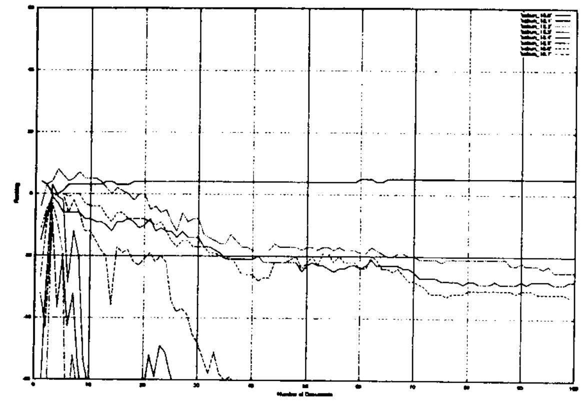

Figure 13. Ranking Difference before and after feedback(automatic,top20,query A)

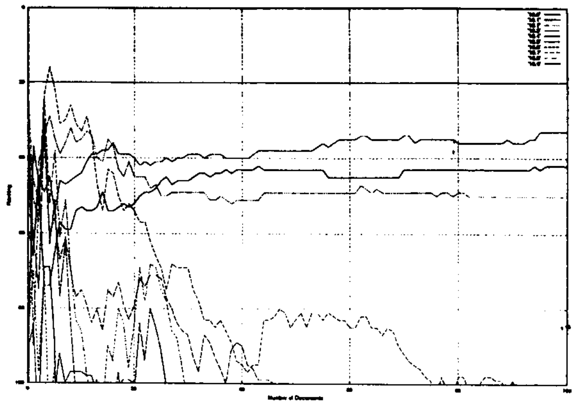

Figure 14. Ranking of relevant documents (automatic ,ranking 20 to 100 ,query A)

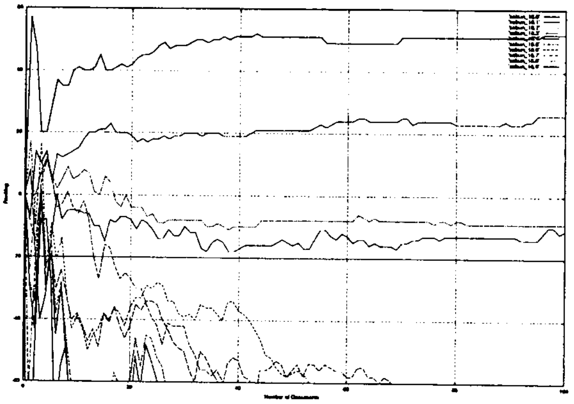

Figure 15. Ranking Difference before and after feedback(automatic,ranking 20 to 100,query A)

3.3 Prospect

Considering result using query A,in spite of using only relevant documents which the user judged ,similarity between some relevant documents and query is decreased . However , the ranking of some of the other relevant documents tends to rise . There would be some couses.

- There are little common index words between the query and documents because both the query and documents are short.
- There are some general nouns in the common index words.

Figure7 shows that more than half documents rise its' ranking. It indicates that the relevance feedback was effective.

In the case using all the higher ranking documents , similarity between some relevant documents and query is decreased because the higher ranking documents include non-relevant documents .

In constrast to that if the feedback is proceeded only when the user judges the documents as relevant,we can see the increase of the ranking.

Specially, in the case of documents ranked in between 20 to 100 by the first retrieval,the effect of the feedback is greatly marked. Even by the autimatical feedback it is noted that 30 % of documents are increased of their rankings.These figures tell us that great differense is found between query A and B. We can see that Figure3 which indicates the number of relevant documents illustrates relatively stable result of feedback . Because query B contains more relevant documents which occupy high rankings by the first retrieval than query A.At the same time,it is possible to examine that

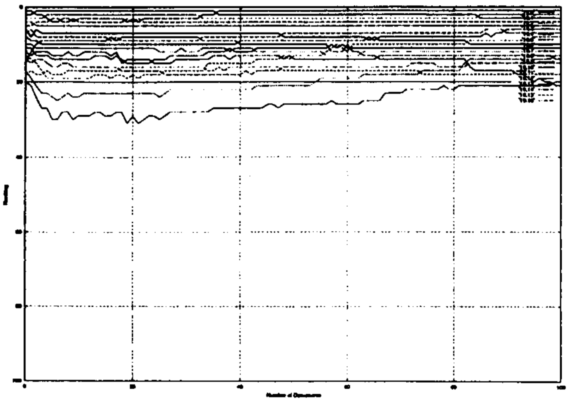

Figure 16. Ranking of relevant documents (automatic ,top 20 ,query B)

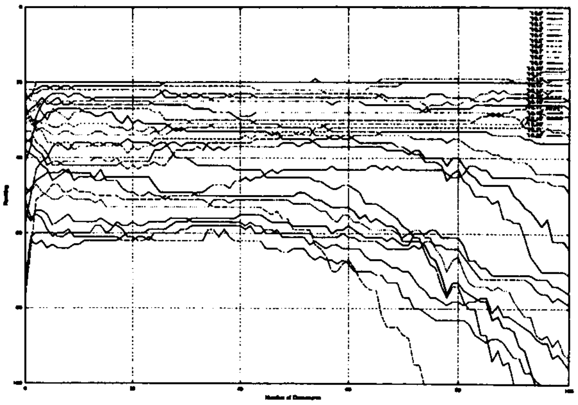

Figure 18. Ranking of relevant documents (automatic ,ranking 20 to 100 ,query B)

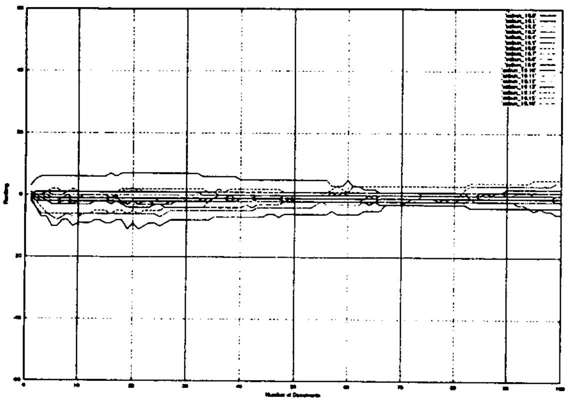

Figure 17. Ranking Difference before and after feedback(automatic,top 20 ,query B)

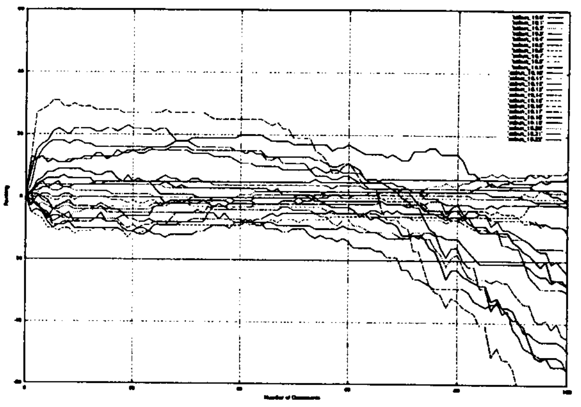

Figure 19. Ranking Difference before and after feedback(automatic,ranking 20 to 100,query B)

the effect of feedback is small in Figure3 since the number of index words in query B ,5 is fewer than that in query A,8.

4 Conclusion

In this paper,we have been analized how the relevance feedback would effect the retrieval.We could see the great difference in the effect of feedback depending on the query , even if we use the same methods. Query A shows that the effect of feedback has worked negativery. Whereas query B indicates the positive result,which is not greatly marked. Our method might not be effective enough,since we can see several points to be improved. First of all, stop-list is necessary to remove the negative effect caused by general nouns. In addition , collocations should be included into index words,because some technical terms make sense in the form of a collocation . The aim of our research is to improve the part of the speech recognition which constructs news audio retrieval system. It is worth while considering which method can give the effective information to the speech recognition especially langage models .

References

[1] 中川聖一,西崎博光,"音声入力によるニュース音声検索システム",音声言語情報処理 26-3, 自然言語処理 131-3,1999.5.28

[2] 岸田和明, "情報検索の理論と技術", 勁草書房,1998

[3] 徳永健伸,"情報検索と言語処理",東京大学出版会,1999