Fusing manual and machine feedback in biomedical domain

[1]Jainisha Sankhavara, [1]Fenny Thakrar, [2]Shamayeeta Sarkar, [1]Prasenjit Majumder

[1]DA-IICT, Gandhinagar, Gujarat, India
[2]Ramananda College, Bankura, West Bengal, India

Abstract. For our participation in CDS task of TREC, our first objective was to obtain efficient biomedical document retrieval. We focused on fusing manual and machine feedback runs. Fusion run performs better and gives consistent results for considered evaluation metrics. Also, the categories 'diagnosis' and 'treatment' are giving good results compared to 'test'.

1 Introduction

CDS-TREC 2014 Task[i]: CDS (Clinical Decision support) task is introduced for the first time in TREC 2014. This is a single task with a focus on retrieval of biomedical articles relevant for answering generic clinical questions about medical records. There are 30 topics provided, each consists of a case report and one of three generic clinical question types ('diagnosis', 'treatment' and 'test'). The topic consists of description and summary of the case and the participant is expected to use either for a particular run. The task is to retrieve full text biomedical articles that answer one of the asked generic clinical questions.

The participants had to submit at most 5 different runs. We generated 3 more runs where manual relevance feedback was used (top 5 documents were manually judged). we have applied two types of fusions: CombSUM and Z_fusion [1] [2]. After the relevance judgement files were released we compared our runs and presented an analysis.

Data Statistics: The data provided for the CDS track is the Open Access Subset of PubMed Central [ii]. To maintain the consistency of data set, we are provided with a snapshot of open access subset taken on January 21, 2014. It contained a total of 733,138 articles (47.2 GB). The article is

[i] (http://www.trec-cds.org/2014.html)
[ii] (http://www.ncbi.nlm.nih.gov/pmc/)

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Fusing Manual and Machine Feedback in Biomedical Domain	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT),Gandhinagar, Gujarat, India,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT
For our participation in CDS task of TREC, our rst objec- tive was to obtain e cient biomedical document retrieval. We focused on fusing manual and machine feedback runs. Fusion run performs better and gives consistent results for considered evaluation metrics. Also, the categories 'diagnosis' and 'treatment' are giving good results compared to 'test'.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**12**	

represented as NXML file. Each article in the collection is identified by a unique number PMCID. The article is named using the same PMCID number.

We have described runs in section 2, results and analysis in section 3, and we conclude in section 4.

2 System and Runs

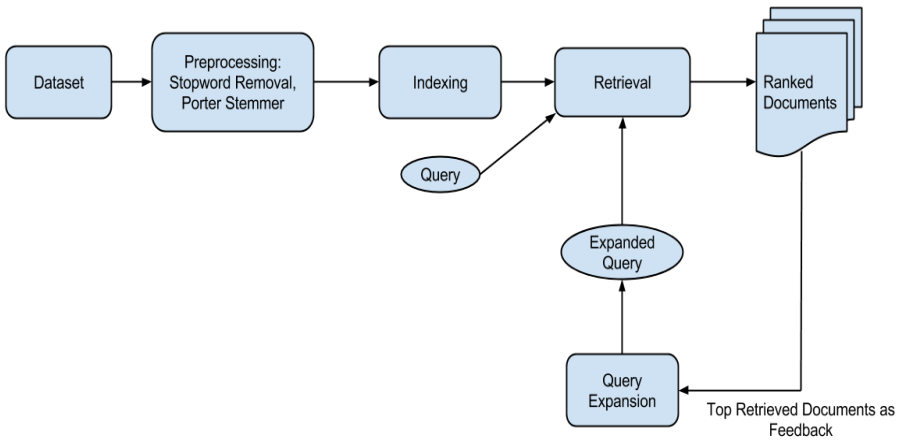

Fig. 1. System Overview Diagram

Our System consists of various modules. Preprocessing of dataset consists of stop-word removal process and porter stemmer. Then the index is built on processed data. Providing the query, the system retrieves relevant documents and assigns ranks to them. Top retrieved documents are given as blind feedback to expand the query and again the documents are retrieved using expanded query.

We have used terrier[3] for our experiments. While indexing, terrier was a little modified to set document id-tag. For the parsing of .nxml files, we have used a number of dtd files[iii]. Also, we have encountered some of the documents with duplicate PMCIDs. There were 138 files whose document name and PMCID were not matching and PMCID was repeated. So, we have removed those 138 files from the data set and proceeded further.

We have submitted five runs (numbered 1 to 5). Along with them, we have analysed and compared run 6 to 8.

[iii] (http://dtd.nlm.nih.gov/book/tag-library/n-tk72.html)

1. **DAIICTdqep:**
   The first run DAIICTdqep is using description as a query. The retrieval model used is In_expC2. It is a query expanded run using Bo1 expansion model and pseudo-relevance feedback of top 5 documents and 250 terms.
2. **DAIICTdqer8:**
   The second run DAIICTdqer8 is also using description as a query. The retrieval model used is In_expC2. It is a query expanded run using Bo1 expansion model, Rocchio model with beta value 0.8 and pseudo-relevance feedback of top 5 documents and 250 terms.
3. **DAIICTsqer8:**
   The third run DAIICTsqer8 is same as run DAIICTdqer8 but it has used description as query and this run is based on summary. This run uses summary as query.
4. **DAIICTf:**
   The fourth run DAIICTf is a fusion run of four different runs. CombSUM fusion method is used to fuse them. Those four results are of the following type.

Table 1. Details of Fused Runs

No	**Query**	**Retrieval Model**	**Expansion Model**	**Rocchio**	**Beta Value**	**Feedback**
i	Description	In_expC2	Bo1	-	-	blind
ii	Description	In_expC2	Bo1	✓	0.4	blind
iii	Description	In_expC2	Bo1	✓	0.8	blind
iv	Description	In_expC2	KL	-	-	blind

5. **DAIICTzf:**
   The fifth run DAIICTzf is also a fusion run of the same four results used in run 4 DAIICTf. Here, the runs are fused using z-fusion method.
6. **DAIICTrelfb:**
   Here, we gave relevance feedback to the system to expand the query. The relevance judgement used in the process is the manual judgement given by the expert from biomedical domain. Using this manual judgement and the same set-up parameters as of run 2 (DAIICTdqer8), the result is taken.
7. **DAIICTnewf:**
   This run is an extension of run 4 (DAIICTf). Instead of 4, total 5 results are fused using CombSUM method in this run. Out of those 5 runs, 4 are the same runs fused in run 4 (DAIICTf) and fifth run is run 6 (DAIICTrelfb).

8. **DAIICTnewzf:**
   This run is same as run 7 DAIICTnewf, but the fusion method used here is z-fusion.

3 Official results and Discussion

The query-wise results of infAP, infNDCG, R-Prec and p@10 are provided for each submitted run (5 runs) by the officials of TREC-CDS task. After the release of relevance judgement, we have evaluated our other three runs (DAIICTrelfb, DAIICTnewf, DAIICTnewzf) for the same evaluation measures. We have also considered MAP for the comparison of runs. The table 2 summarizes all results and figure 2 is a graphical representation of it.

Table 2. Evaluation measures for all runs

No	Run	MAP	infAP	infNDCG	R-prec	p@10
	Best		0.1805	0.5197	0.3496	0.7100
	Median		0.0316	0.1514	0.1257	0.2333
1	DAIICTdqep	0.1546	0.0745	0.2404	0.2059	0.3067
2	DAIICTdqer8	0.1523	0.0781	0.2382	0.1952	0.3233
3	DAIICTsqer8	0.1476	0.0728	0.2562	0.1990	0.3733
4	DAIICTf	0.1559	0.0766	0.2442	0.2000	0.3167
5	DAIICTzf	0.1559	0.0765	0.2436	0.1995	0.3167
6	DAIICTrelfb	0.1428	0.0759	0.2374	0.1891	0.3533
7	DAIICTnewf	0.1565	0.0773	0.2493	0.2031	0.3333
8	DAIICTnewzf	0.1563	0.0767	0.2464	0.2030	0.3300

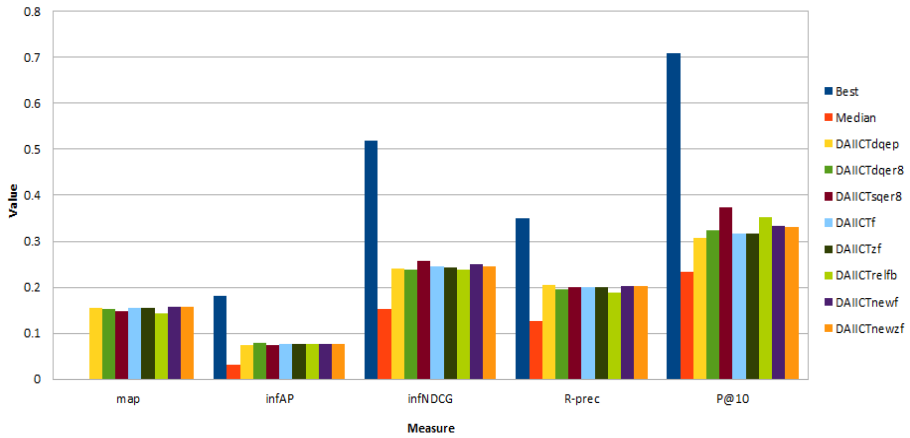

Fig. 2. Comparison of all five measures for eight runs

Here, best and median values are the average of all query-wise best and median values respectively.

The analysis shows that the run 2(DAIICTdqer8) gives best infAP, while the run 3(DAIICTsqer8) gives best for infNDCG. Also, the run 7(DAIICTnewf) shows consistent results in infAP and infNDCG.

The fusion runs 7(DAIICTnewf) and 8(DAIICTnewzf) having manual feedback (DAIICTrelfb) included, perform better. There was a significant improvement as compared with our previously fused runs. The statistical hypothesis testing - paired t-test for five evaluation matrices' values of DAIICTf and DAIICTnewf confirms that they are significantly (p=0.0766) different. the runs DAIICTzf and DAIICTnewzf are also significantly (p=0.0841) different.

The query-wise differences of our best runs with best and median values of infAP and infNDCG are graphically represented in figures 3 to 8.

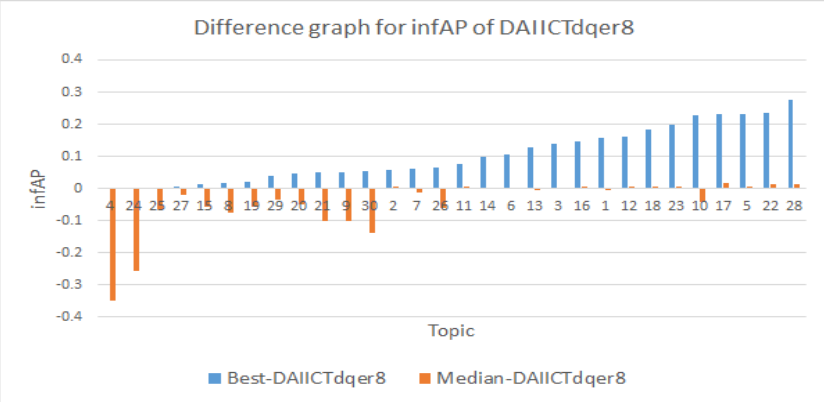

Fig. 3. Difference graph of infAP of DAIICTdqer8

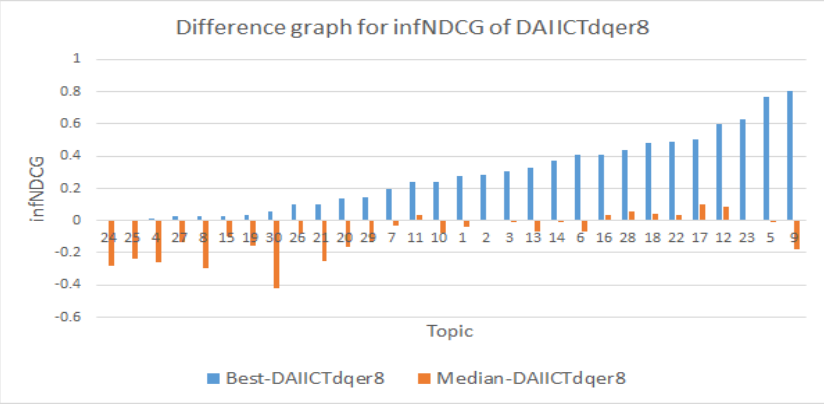

Fig. 4. Difference graph of infNDCG of DAIICTdqer8

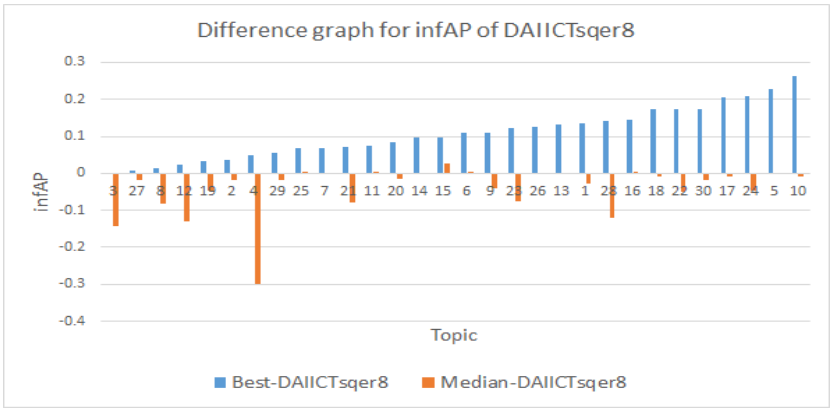

Fig. 5. Difference graph of infAP of DAIICTsqer8

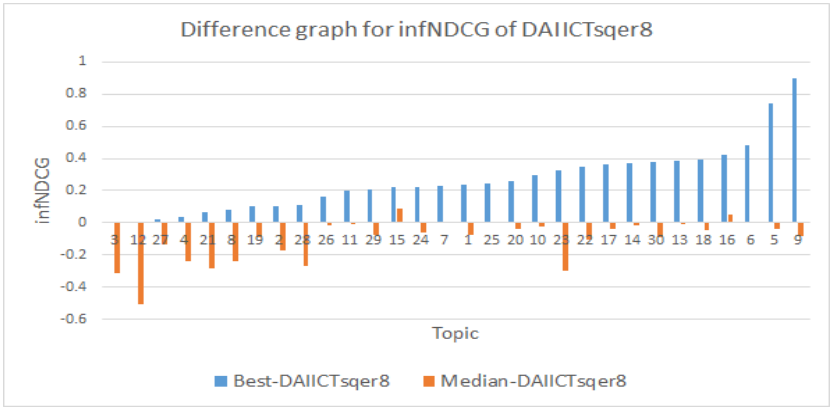

Fig. 6. Difference graph of infNDCG of DAIICTsqer8

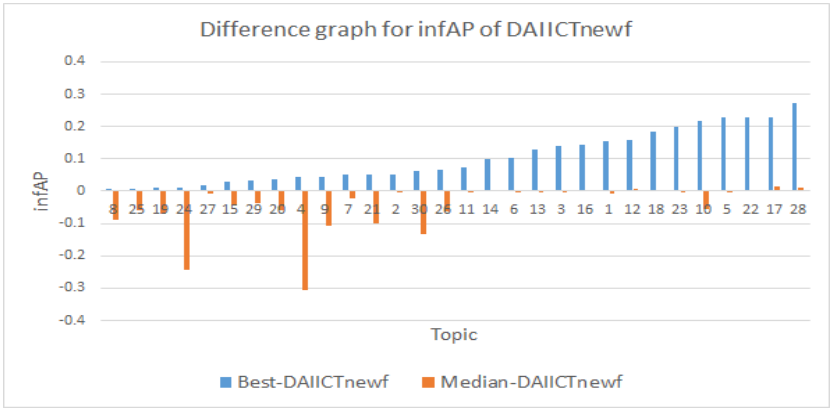

Fig. 7. Difference graph of infAP of DAIICTnewf

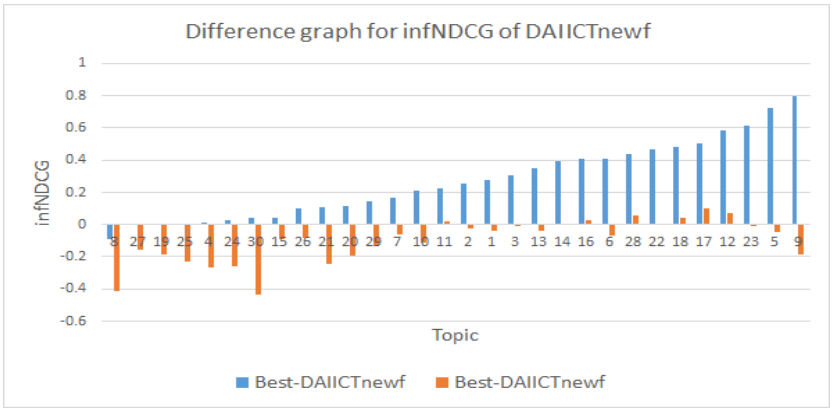

Fig. 8. Difference graph of infNDCG of DAIICTnewf

Category-wise analysis of results:

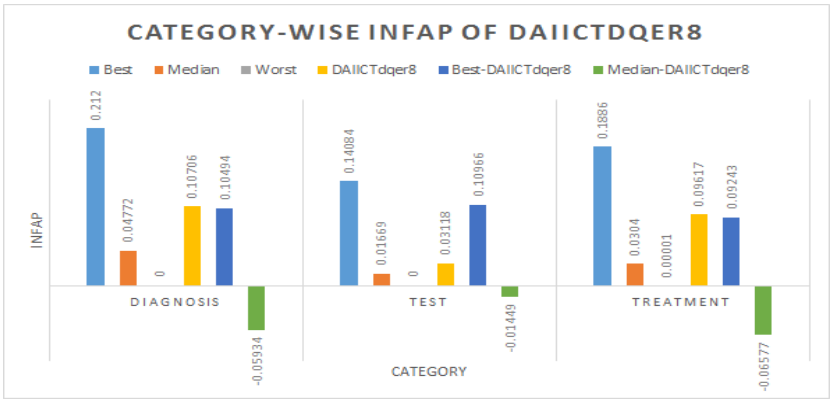

Fig. 9. Category-wise infAP of DAIICTdqer8

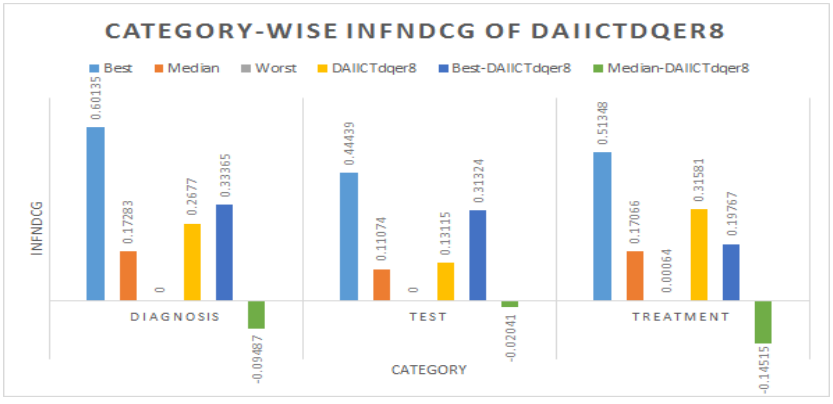

Fig. 10. Category-wise infNDCG of DAIICTdqer8

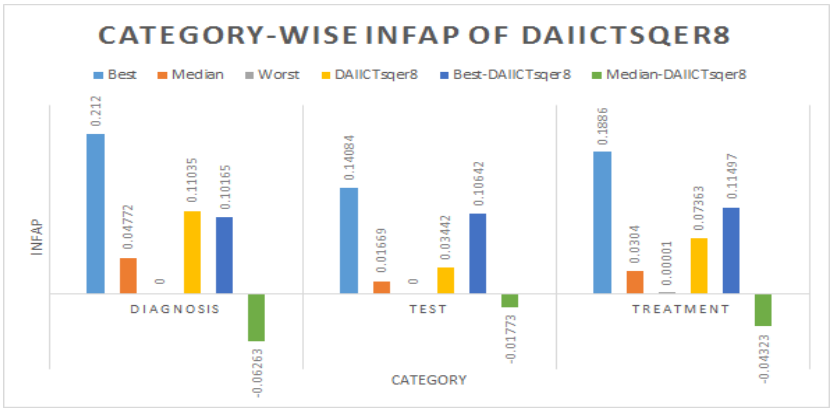

Fig. 11. Category-wise infAP of DAIICTsqer8

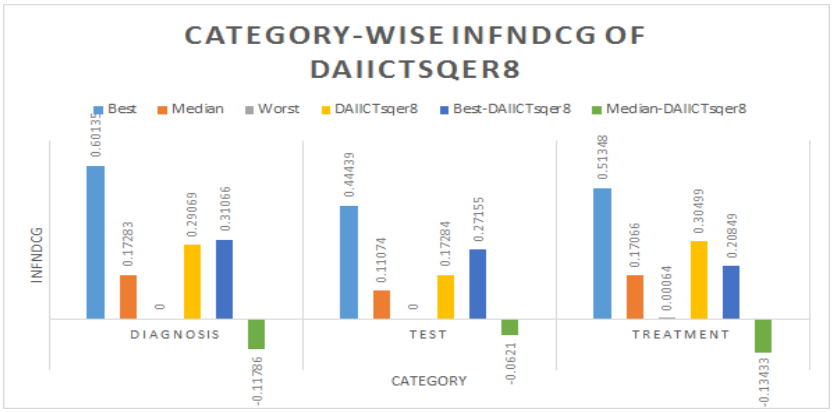

Fig. 12. Category-wise infNDCG of DAIICTsqer8

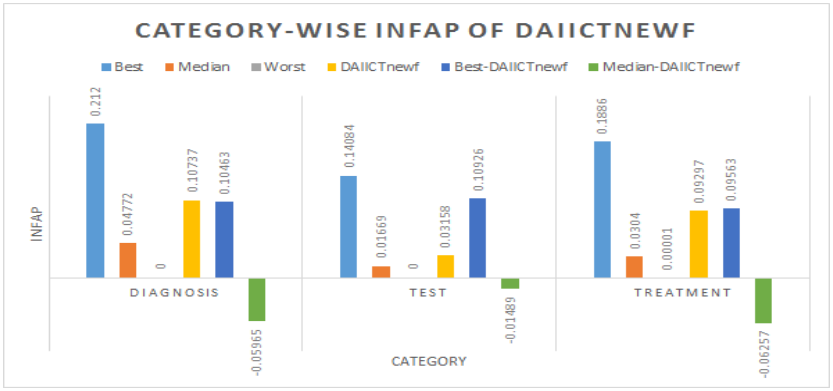

Fig. 13. Category-wise infAP of DAIICTnewf

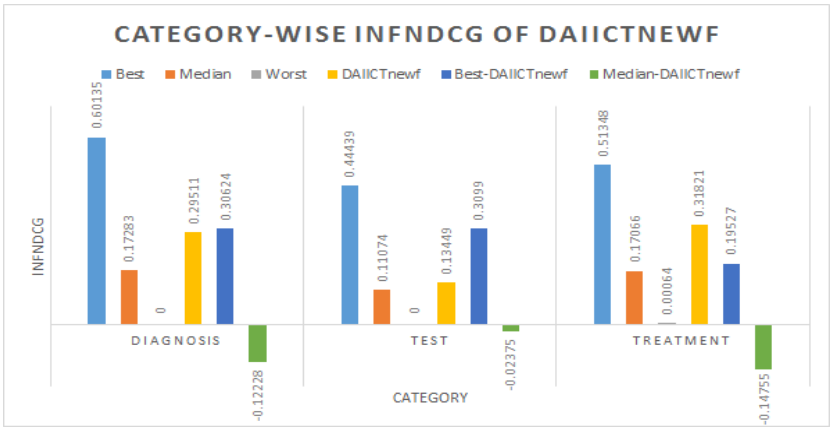

Fig. 14. Category-wise infNDCG of DAIICTnewf

From the figure 9 to 14, we can say that among three categories, the test category performs less. Its difference from the median is very less as compared to diagnosis and treatment.

Query-wise analysis of results:

Figure 15 shows query-wise infAP value for runs DAIICTdqer8 (description as query) and DAIICTsqer8(summary as query). Similarly figure 16 shows query-wise infNDCG of DAIICTdqer8 and DAIICTsqer8. Here, number of queries performing better while using description is 16 and other 14 queries perform better using summary.

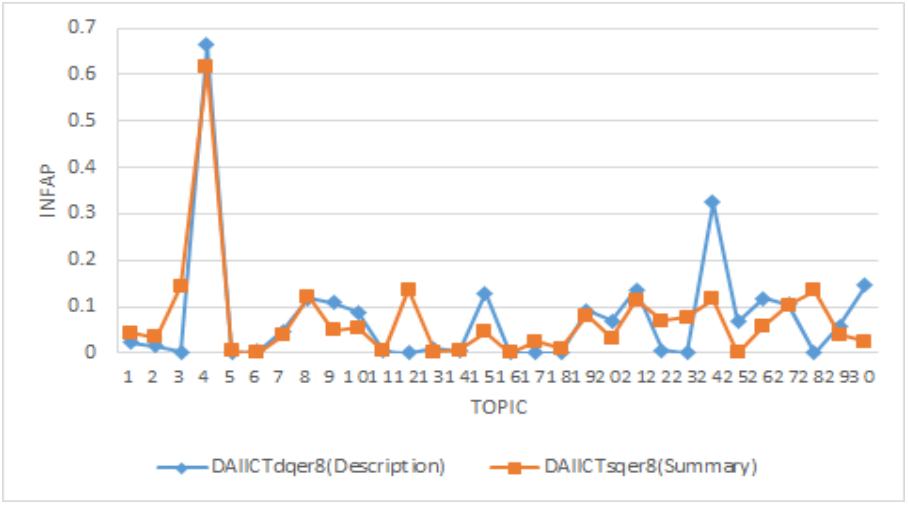

Fig. 15. Query-wise infAP of DAIICTdqer8 and DAIICTsqer8

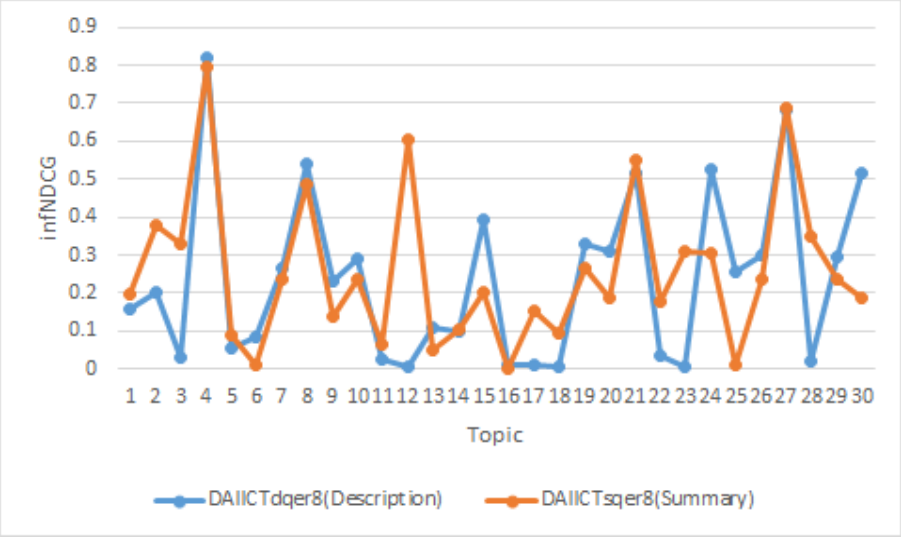

Fig. 16. Query-wise infNDCG of DAIICTdqer8 and DAIICTsqer8

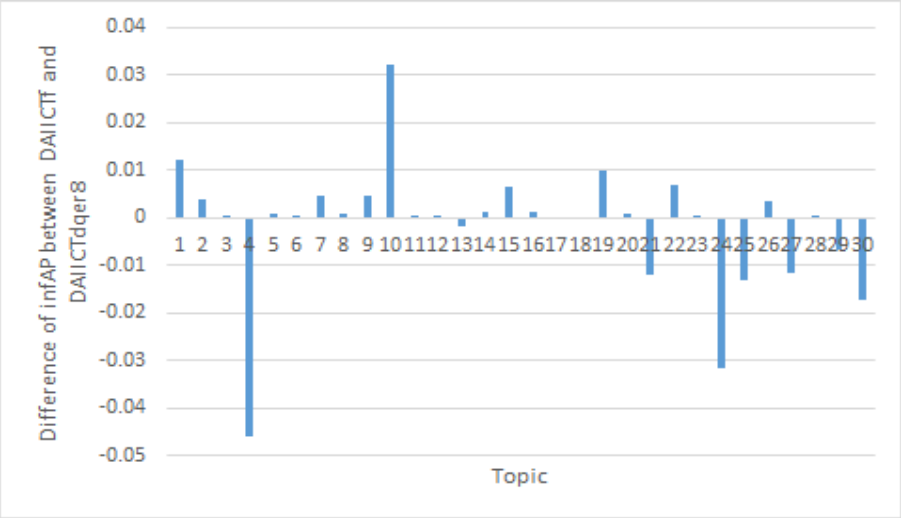

Fig. 17. Query-wise difference of between fusion run and normal run

While considering fusion runs(DAIICTf, DAIICTzf) and the run DAIICTdqer8, performance of fusion runs increases for 21 queries among 30 as seen from figure 17. While for other 9 queries, fusion runs gives less result than DAIICTdqer8.

Poor performing queries:

We did first level investigation for poor queries. Overall, the queries 16,17,18,23 and 28 are performing very low. Query 17:

```
<topic number="17" type="test">
<description>
A 48-year-old white male with history of common variable immun-
odeficiency (CVID) with acute abdominal pain, fever, dehydra-
tion, HR of 132 bpm, BP 80/40. The physical examination is re-
markable for tenderness and positive Murphy sign. Abdominal ul-
trasound shows hepatomegaly and abundant free intraperitoneal
fluid. Exploratory laparotomy reveals a ruptured liver abscess,
which is then surgically drained. After surgery, the patient is
taken to the ICU.</description>
<summary>
48-year-old man with common variable immunodeficiency presents
with abdominal pain and fever. Ultrasound reveals hepatomegaly
and free intraperitoneal fluid. A ruptured liver abscess is found
and drained during exploratory laparotomy.</summary>
</topic>
```

The query vector of this query in the run 2(DAIICTdqer8) has top weighted terms like

test-0.333333333	output-0.333333333	undergo-0.333333333
6-0.344458871	0-0.333333333	major-0.333333333
month-0.350165797	2-0.347232117	surgeri-0.33333333
old-0.353759205	ml-0.333333333	examin-0.685316682
male-0.333333333	kg-0.333333333	gener-0.333333333
infant-0.333333333	hr-0.333333333	edema-0.370116354
urin-0.975785110	shortli-0.33333333	

Perhaps the reason behind the query 17 performing very less, is that the top weighted terms are more general terms, and the important terms are assigned very less weight (e.g. hypersensit is assigned the weight 0.011322403).

4 Conclusion & Future work

We conclude that fusion runs DAIICTnewf is consistent for all the four evaluation metrics. Though the fusion of manual feedback and machine(blind) feedback improves the retrieval performance here, we need more investigation for the same in biomedical domain. Also, we analysed that category wise, 'test' category performed poor compared to 'diagnosis' and 'treatment'.

Our future goal is to apply query expansion using MeSH (Medical Subject Heading) [iv] terms for the topics and query expansion by selecting terms for blind feedback by selective summarization. We are planning to apply manual query expansion for each topic of the CDS track.

5 Acknowledgments

We would like to specially thank IRLab DAIICT to provide the useful resources. We would also like to convey our regards to the TREC team for organizing the CDS Track.

References

1. Lee, J.H.: Analyses of multiple evidence combination. In: ACM SIGIR Forum. Volume 31., ACM (1997) 267–276
2. Wu, S., Crestani, F.: Data fusion with estimated weights. In: Proceedings of the eleventh international conference on Information and knowledge management, ACM (2002) 648–651
3. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A high performance and scalable information retrieval platform, Citeseer (2006)

[iv] http://www.nlm.nih.gov/mesh/