A Decade after TREC-4
NTCIR-5 CLIR-J-J Experiments at Yahoo! Japan

Sumio FUJITA
Yahoo Japan Corporation,
Roppongi Hills Mori Tower, 6-10-1, Roppongi, Minato-ku,
Tokyo 106-6182, Japan

Abstract
This paper describes NTCIR-5 experiments of the CLIR-J-J task, i.e. Japanese monolingual retrieval subtask, at the Yahoo group, focusing on comparative studies of the feedback effectiveness with two retrieval methods, namely BM25TF*IDF and a KL-divergence language modeling approaches. An "automatic feedback from top k documents" strategy was surprisingly successful in this test collection. We compared behaviors of the systems with past NTCIR and TREC experiments and find out the characteristics of test collections where the strategy is especially effective.

Keywords: Information retrieval, Automatic feedback, Language modeling approach to IR.

1. Introduction
An "automatic feedback from top k documents strategy", which is referred to by "pseudo relevance feedback" or "blind feedback", gains an improvement of a mean average precision (MAP hereafter) as much as, or even more than, 20% from these baseline runs in test collection based evaluations. In effect, no other single method achieves such big gains against reasonably carried out baseline runs.

Despite such success in test collection based studies, the strategy does not seem to overwhelm one-shot feedback-less approaches in operating search systems. Even the qualifying prefixes of the naming, i.e. "pseudo" or "blind", themselves seem to show reluctance to adopt the strategy as a killer technology. Nevertheless, improvements gained by the feedback are getting larger in recent NTCIR experiments as well as in TREC experiments. The barrier against the usage in operating search systems, i.e. computational cost becomes relatively low given that cheap clusters of high performance PCs are available. The strategy dates back to TREC-2 and was especially successful in TREC-4 [9][10]. We started reexamining the strategy in order to boost the search effectiveness when the input search query is insufficiently poor, as is often the case in operating commercial search systems.

We examined comparatively two types of search models namely a TF*IDF approach with Okapi BM25 TF [16] and a Kullback-Leibler divergence (KL-divergence hereafter) language modeling approach [13] against Japanese newspaper test collections as Japanese monolingual runs of the NTCIR-5 CLIR task.

In NTCIR-3, 4 and 5 CLIR-J-J test collections, BM25 TF*IDF and KL-divergence runs perform similarly when no feedback is applied, whereas feedback gain is larger in BM25 TF*IDF with a Rocchio feedback. This makes us refrain from submitting KL-divergence runs as official results.

The rest of the paper is organized as follows:

Section 2 describes our experiment environment and retrieval system.

Section 3 discusses our official runs and post submission experiments.

Section 4 concludes the paper.

2. System description
Our evaluation environment: YLMS system developed based on Lemur toolkit 4.0 for indexing system[14], which is being developed by the Lemur project.

2.1 Indexing language
Chasen version 2.2.9 Japanese morphological analyzer with IPADIC dictionary version 2.5.1 are utilized for Japanese text segmentation and output single words are used as indexing units.

Stop word lists for newspaper documentation are prepared.

2.2 Retrieval models
The following two retrieval models are examined in experiments:

-KL-divergence of probabilistic language models with Dirichlet prior smoothing

-TF*IDF with BM25 TF

2.3 KL-divergence model
The adopted model is simple: estimate a language model for each document and rank documents by the

likelihood of generating the submitted query. This is exactly a retrieval version of a Naïve Bayes classifier, which estimates a language model for each class and ranks classes by the likelihood of generating the document to be classified. Applying Bayes' theorem, and eliminating document independent part, we have:

$$p(d \mid q) \propto p(d)\,p(q \mid d)$$

Assuming a simple uni-gram model of documents, p(q|d) is:

$$p(q \mid d) = \prod_i p(q_i \mid d)$$

Taking log, the retrieval function becomes:

$$\log(p(d)\,p(q \mid d)) = \log p(d) + \sum_i \log p(q_i \mid d)$$

A document dependent prior probability p(d) can be either uniform probability or any document dependent factors that may affect the relevance such as document length or hyper link related information. Assuming a uniform prior probability and dropping the first term, transforming the summation over query term positions into a summation over words in the vocabulary, dividing by the query length, we have:

$$\sum_{w \in V} p(w \mid q) \log(p(w \mid d))$$

This is exactly the negative cross entropy of a query language model with a document language model, which measures the difference between the two probability distributions and this is equivalent to KL-divergence of the query language model from the document language model in view of ranking documents against the given query.

2.4 Smoothing methods

Zhai and Lafferty presented that the smoothing method plays a crucial role in language modeling IR [22].

They analyzed the role of smoothing in language modeling IR from two aspects: to avoid zero probabilities for unseen words and "to accommodate generation of common words in a query". In this respect, smoothing plays a role similar to IDF in TF*IDF approach. They empirically showed that a Dirichlet-Prior method that computes maximum a posteriori parameter values with a Dirichlet prior (i.e. generalization of Laplace smoothing) performs better than Jelinek-Mercer method i.e. a linear interpolation of a document language model and a collection language model.

Dirichlet-Prior method is:

$$p_\mu(w \mid d) = \frac{freq(w,d) + \mu p(w \mid C)}{|d| + \mu}$$

$$0 < \mu < \infty$$

In this paper, we use KL-divergence retrieval model with Dirichlet smoothing method (KL-Dir hereafter).

2.5 BM25TF*IDF

RSV between a document d and a query q is calculated as a dot product between the document term vector and the query term vector, where each term is weighted by TF*IDF [18]. Okapi BM25 TF [17] is used.

$$RSV(q,d) = \sum_{t \in q \cap d} w(q,t)\,w(d,t)$$

$$w(d,t) = TF(d,t)\,IDF(t)$$

$$w(q,t) = TF(q,t)\,IDF(t)$$

$$TF(d,t) = \frac{(k1+1)\,freq(d,t)}{k1((1-b) + b\frac{dl_d}{avdl}) + freq(d,t)}$$

$$IDF(t) = (k4 + \log\frac{N}{df(t)})$$

d : document or query

t : term

N : total number of documents in the collection

$df(t)$: number of documents where t appears

$freq(d,t)$: number of occurrences of t in d

dl_d : document length of d

avdl : average document length in the collection

2.6 Feedback strategies

The strategy of "feedback from top k documents in a pilot search" is applied.

The Rocchio feedback for TF*IDF is adopted as term extraction method.

For KL-divergence runs, a mixture model query update method, where language models for pseudo-

	MAP-Rigid	RP-Rigid	Rel-Ret	P@10	P@20	MAP-Relax	RP-relax	Rel-Ret	P@10	P@20
YLMS-J-J-T-01	0.3472	0.3462	1879	0.4489	0.3862	0.4263	0.4228	3728	0.6106	0.5564
YLMS-J-J-D-02	0.3119	0.3214	1852	0.4000	0.3468	0.4008	0.4048	3666	0.5447	0.5053
YLMS-J-J-T-03	0.4193	0.4250	1959	0.5277	0.4309	0.5028	0.4911	3844	0.6915	0.6128
YLMS-J-J-D-04	0.3674	0.3734	1964	0.4553	0.3766	0.4641	0.4592	3846	0.5979	0.5468
YLMS-J-J-TDNC-05	0.4457	0.4426	1992	0.5404	0.4617	0.5302	0.5146	3914	0.6936	0.6447
TDNC-run	0.4433	0.4419	1992	0.5277	0.4553	0.5215	0.5114	3914	0.6851	0.6298

Table 1: Effectiveness of CLIR-J-J official runs

	MAP-Rigid	RP-Rigid	Rel-Ret	P@10	P@20	MAP-Relax	RP-relax	Rel-Ret	P@10	P@20
YLMS-J-J-T-03	0.4193	0.4250	1959	0.5277	0.4309	0.5028	0.4911	3844	0.6915	0.6128
%gain	+41.7	+29.2	+12.3	+29.8	+29.0	+32.9	+24.6	+11.6	+18.6	+19.8
YLMS-J-J-T-03 No FB	0.2960	0.3289	1745	0.4064	0.3340	0.3782	0.3940	3444	0.5830	0.5117
KL-Dir Mix FB	0.4134	0.4174	1902	0.5128	0.4277	0.4874	0.4811	3744	0.6702	0.5926
%gain	+40.4	+33.0	+11.3	+28.9	+25.6	+29.0	+21.8	+10.2	+15.0	+18.3
KL-Dir No FB	0.2944	0.3139	1709	0.3979	0.3404	0.3779	0.3951	3396	0.5830	0.5011

Table 2: Effectiveness of CLIR-J-J unofficial title only runs compared with YLMS-J-J-T-03

relevant documents are distilled by eliminating background noises using EM iteration as described by Zhai and Lafferty [21] and the feedback document model p(F|θ) is estimated given the fixed mixture parameter λ.

$$\log(p_\lambda(F \mid \theta)) = \sum_i \sum_w c(w; d_i) \log((1-\lambda)\, p(w \mid \theta) + \lambda p(w \mid C))$$

3. CLIR Experiments

3.1 CLIR official runs for J-J SLIR

We submitted two title only runs, two description only runs and one long query run using all the topic fields, of Japanese monolingual retrieval setting. All the official runs are using TF*IDF method with BM25 TF and a Rocchio feedback with a top k documents strategy.

YLMS-J-J-T-01 and YLMS-J-J-D-02 runs, which are title and description runs, perform poorer than our best performing runs because of some bugs in our program.

YLMS-J-J-T-03 and YLMS-J-J-D-04 runs are our best performing title run and description run respectively. The parameters are as follows:

YLMS-J-J-T-03: K1=1.4, b=0.35, k4=1, #feedback documents=9, Max feedback terms=70, feedback positive coefficient=0.5

YLMS-J-J-D-04: K1=1.7, b=0.5, k4=1, #feedback documents=16, Max feedback terms=180, feedback positive coefficient=0.9

YLMS-J-J-TDNC-05 is a fusion of a long query run (noted TDNC-run in Table 1) and YLMS-J-J-D-04 with a mixture parameter: 0.5.

$$\text{score} = (1-\alpha)\,\text{RunScore}\,1 + \alpha\,\text{RunScore}\,2$$

Test collection	Run description	MAP Rigid	PFB Gain %	MAP Relax	PFB Gain%
NTCIR-1 Adhoc DESC run	TF*IDF, pseudo feedback	0.3596	+11.4	-	-
	TF*IDF, no feedback	0.3227		-	
NTCIR-3 CLIR J-J TITLE query Rigid /Relax	BM25TFIDF, K1=1.0, k4=1.0, b=0.35, #feedbackDocs=7, #feedbackTerms = 100, PosCoeff = 0.1	0.3930	+19.4	0.4502	+17.5
	BM25TFIDF, K1 = 1.0, k4 = 1.0, b = 0.35, no pseudo feedback	0.3292		0.3830	
NTCIR-4 CLIR J-J TITLE query Rigid /Relax	BM25TFIDF, K1=1.0, k4=1.0, b=0.35, #feedbackDocs=7, #feedbackTerms = 100, PosCoeff = 0.1	0.3801	+23.0	0.4711	+19.1
	BM25TFIDF, K1 = 1.0, k4 = 1.0, b = 0.35, no pseudo feedback	0.3090		0.3956	
NTCIR-3 Patent Desc query A / AB	KL-Dir, μ =2000, #feedbackDocs=20, #feedbackTerms=120, Coeff = 0.5	0.3283	+15.4	0.3209	+14.2
	KL-Dir,μ = 2000, No pseudo feedback	0.2846		0.2811	
NTCIR-4 Patent Claim query A / AB	KL-Dir,μ = 900, #feedbackDocs = 8, #feedbackTerms = 100, Coeff = 0.22	0.2508	+9.5	0.1655	+6.8
	KL-Dir, μ = 900, No pseudo feedback	0.2290		0.1549	
TREC-9 Web Title run	BM25TF*IDF, pseudo feedback + reference collection feedback	-	-	0.2028	+15.8
	BM25*TF*IDF, no feedback	-		0.1751	
TREC-2001 Web Title run	BM25TF*IDF, pseudo feedback + reference collection feedback	-	-	0.2060	+20.9
	BM25*TF*IDF, no feedback	-		0.1704	
TREC 2004 MEDLINE Long query DR /DR+PR	BM25TFIDF, K1 = 0.4,k4 = 0.1,b = 0.8, #feedbackDocs = 7, #feedbackTerms = 30, PosCoeff = 0.1	0.3695	+4.8	0.4075	+4.1
	BM25TFIDF, K1 = 0.4, k4 = 0.1, b = 0.8, no pseudo feedback	0.3526		0.3915	

Table 3: Parameters, MAPs and % gains by pseudo feedbacks of two retrieval methods in 8 test collections (Official submission runs where available and their baseline no feedback runs)

Table 1 shows the effectiveness of official runs and a TDNC-run used in the fusion.

3.2 Automatic feedback strategy

Table 2 compares YLMS-J-J-T-03, our best performing title only run, with a no feedback baseline run as well as KL-Dir runs.

No feedback runs use terms extracted from the "title" fields of topics and the number of terms is 4.58 per topic, whereas YLMS-J-J-T-03 uses expanded queries as long as 70.44 terms per topic and the KL-Dir feedback run, 120.36 terms per topic. Improvements achieved by the feedback are as much as 41.7%, which we have never experienced in our past NTCIR and TREC experiments.

Table 3 shows the % gain achieved by automatic feedback strategies in our past experiments.

We compared our official results of the past participation in TRECs or NTCIRs where available, comparing with their no feedback baseline runs in post-submission experiments, otherwise unofficial comparative experiment results are presented [4][5][6][7][8].

In NTCIR-3 and 4 CLIR-J-J experiments, 19.4 to 23.0 % of improvements are observed by a feedback from top k documents of a pilot search strategy, whereas in TREC-9 and TREC-2001 Web tracks, 15.8 to 20.9% of improvements are achieved by combining two feedback strategies, one by top k documents from the target collection and the other from a reference collection, i.e. TREC CD1-3 collections. More than 40% of improvement is really exceptional.

Historically, an "automatic feedback from the top k documents strategy" dates back to the TREC-2 ad hoc track, where some groups tried to apply an "automatic query expansion without relevance information", and among them, the "CLARIT" group reported its effectiveness in both manual/automatic runs [3]. In the TREC-3 ad hoc track, where the

elimination of the "concepts" field in the topic description accelerates the techniques to automatically expand less rich queries [9]. Among groups adopting automatic feedbacks, the best performing runs by the "City/Okapi" group achieved "unexpectedly successful" improvement of as large as 19%, by extracting up to 40 terms from the top 30 documents [17]. The succeeding TREC-4 ad hoc track was memorized by the heavy use of such feedback techniques by several groups. For example the "Cornel/SMART" team achieved a 27% improvement from their baseline runs [1] whereas the "CUNY/PIRCS" team reported a 29% improvement by an expansion from best-ranked subdocuments [12].

Kwok et al. pointed out that the systems that heavily rely on such techniques outperform in the high recall region [12]. Figure 1 shows the recall-precision curves comparing feedback runs and their baseline runs. This illustrates the characteristics of these runs that the difference is larger when the recall is 30% to 60%. Such characteristics are more emphasized in inter-system comparison by recall-precision curves as seen in organizer's task overview papers for example in ad hoc short query runs in NTCIR-1 [11] where our submission "jscb1" [4] heavily relies on the feedback strategy is only the ninth position of inter-system ranking evaluated by "precision at 5 docs". Evaluated by the "precision at 15 docs", it jumped up to the first position and keeps the first position until the "precision at 1000 docs".

3.3 KL-Dir runs

As seen in Figure 1, KL-Dir performs similar to BM25TF*IDF when no feedback is applied, whereas BM25TF*IDF is slightly better in feedback runs, though the both differences are statistically not significant.

As seen in Figure 2, the KL-Dir no feedback run converges to the best MAP when the Dirichlet prior is about 1500, whereas the feedback run peaks at 500.

Figure 3 shows the sensitivity of MAP to feedback mixture noise ratios and feedback coefficients. A feedback mixture noise ratio, which is the weight of background collection noises when linearly combining with the feedback document model, does not affect so much the effectiveness except the case of 1.0, i.e. the model consisting of only the background noises. When 0.0, i.e. the model of only the feedback documents, MAP is not dramatically worse than the best run, though the difference is statistically significant.

On the other hand, the feedback coefficient, i.e. the mixture parameter of the feedback document model by which it combined with the original query model, clearly affects the final results. In this case, it peaked at 0.9, whereas 0.8 in the NTCIR-3 and 4 CLIR-J-J collections, and 0.1 in the TREC-2004 MEDLINE collection.

3.4 Some feedback parameters

Figure 4 shows how the number of feedback documents affects the final results. Some TREC participants tended to have used much more documents than we did here. For example "Cornel/SMART" group reported 20 documents [1] whereas "CUNY/PIRCS" group used 40 subdocuments [12] and "City/OKAPI" used 30 documents [16]. Tao and Zhai reported that they achieved the best run when used 50 documents against the TREC AP collection [19].

It is a little bit surprising given that the average document length counted by terms in the TREC collections is 348, which is much longer than that of NTCIR CLIR-J-J collections, i.e. 189 in NTCIR-3, 193 in NTCIR-4 and 197 in NTCIR-5. As shown in Table 3, we rarely used more than 10 documents and typically used 6 to 10 documents.

Table 3 also shows that we typically used about 100 terms to feedback and Figure 5 shows that the best performance is observed at 140 terms.

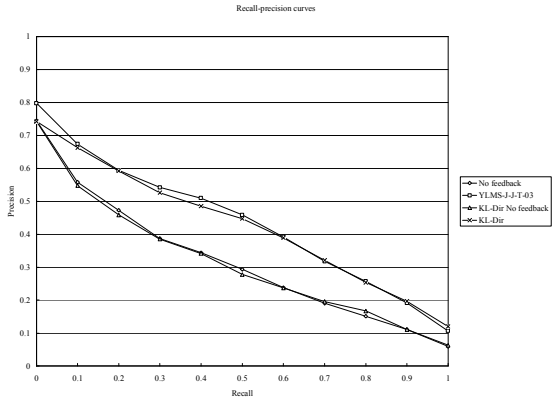
"Cornel/SMART" group reported 500 terms, "CUNY/PIRCS" group, 50 terms and "City/OKAPI" group, 40 terms.

Figure 6 shows the effect of feedback positive coefficient in BM25TF*IDF runs, i.e. coefficient of feedback terms. Surprisingly, the same weight as original query terms, i.e. 1.0, gives the best results. As have been shown in Table 3, this is also a rare case in our past experiences. "Cornel/SMART" group reported that they used the same coefficient to the original query terms and expanded terms in TREC-4.

3.5 Sensitivity of feedback effectiveness to test collection characteristics

From the observations made in the previous subsections, we assume that the top k documents feedback strategy is exceptionally successful in some test collections, such as NTCIR-4 and 5 CLIR-J-J, and the TREC-3 and 4 ad hoc track, where near to or more than 20% (even 40%!) of improvements are achieved by some groups.

Typically many feedback terms with larger coefficients, i.e. aggressive feedback strategies are successful with such test collections. The common characteristics of such test collections include:



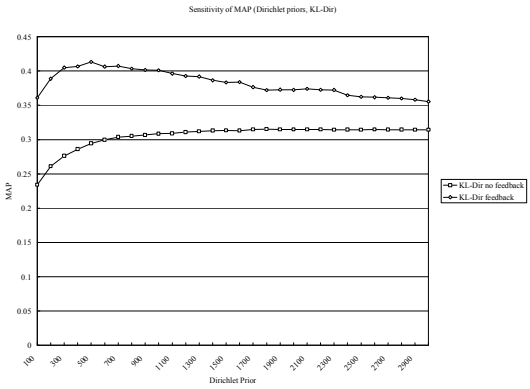
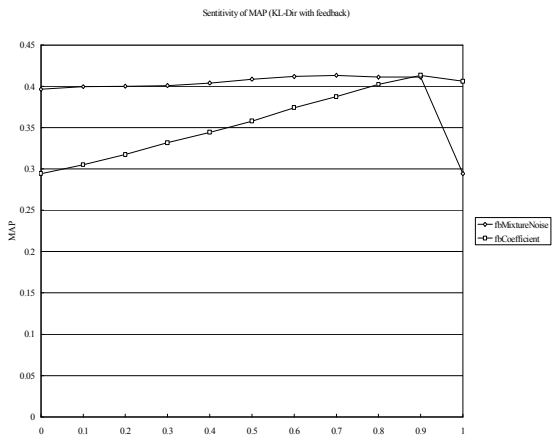


Figure 1 (Left): Recall-precision curves of YLMS-J-J-T-03, its no feedback baseline, KL-Dir run and its no feedback baseline

Figure 2 (Right): Sensitivity of MAP for KL-Dir runs and their no feedback baseline runs with different Dirichlet priors



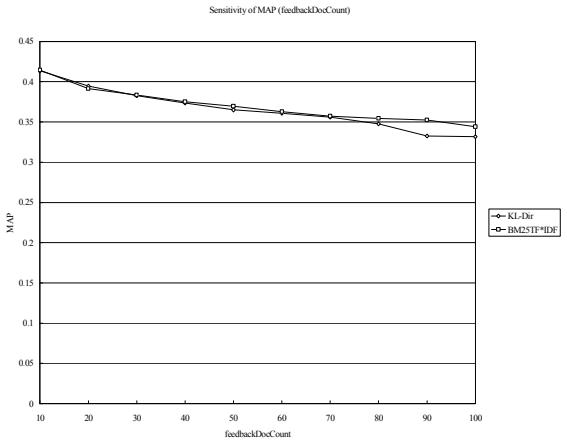
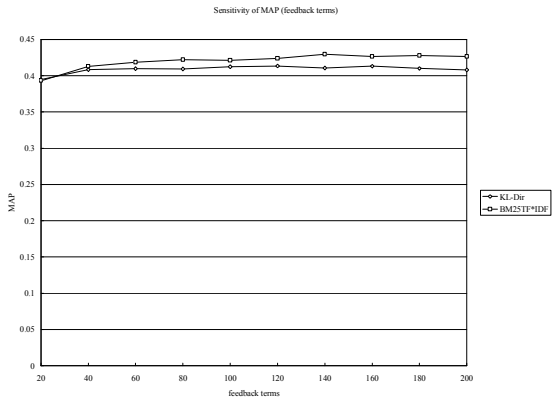


Figure 3 (Left): Sensitivity of MAP for KL-Dir runs with different feedback coefficient and different feedback mixture noise

Figure 4 (Right): Sensitivity of MAP to number of feedback documents (top k ranked)



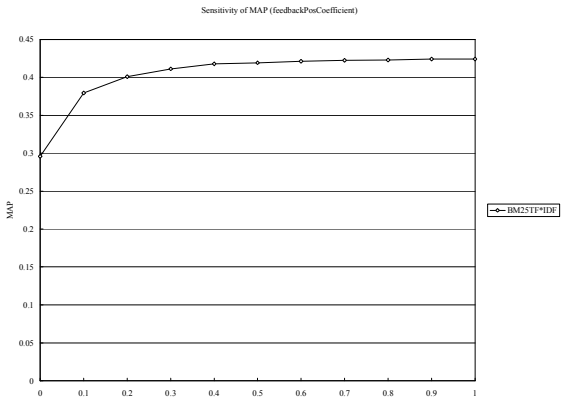


Figure 5 (Left): Sensitivity of MAP to number of feedback terms (Max n terms)

Figure 6 (Right): Sensitivity of MAP to feedback positive coefficient (coefficient of positive term weight)

1) Short query

Feedback gain is emphasized when the original queries are short and terminologically not so rich.

2) Sufficient number of relevant documents

In order to achieve improvements, there should be some relevant documents to be promoted, which have retrieved at lower ranks in the pilot search.

3) Terminologically controlled and "clean" document collections such as newspapers or newswires

The strategy is not straightforwardly applicable to web documents, where the gain is smaller.

4) The document collections are repeatedly used in the preceding workshops.

The repeated use of the document collections or similar collections uncovers the collection characteristics and the task practitioners can afford to take an aggressive strategy.

On the other hand, Eguchi et al. [2] categorized NTCIR-1 topics into three categories namely easy, middle and hard (in topic difficulties) according to the median average precision through 26 submitted results. They pointed out that the inter-system ranking is sensitive to the topic difficulties. The result we submitted to the track, "jscb1", which adopted a sort of aggressive feedback strategies, ranked at the second best place in the easy and middle categories but the first place in the hard category. Naturally feedback strategies can achieve improvements when original topic terms are not enough to retrieve all relevant documents and such topics presumably fall into the hard category.

The characteristic 3) suggests another issues about the feedback effectiveness. Terminological cohesiveness through relevant documents is assumed in order that the feedback is effective. Cohesiveness is understood as groups of shared terminology by documents relevant to the same search topics. The notion is studied as the "cluster hypothesis" [15] in the history of IR studies, and some measures indicating how well the "cluster hypothesis" holding true are proposed but no clear correlation between the measure and the retrieval effectiveness is observed [20]. The feedback effectiveness is possibly a measure that shows the cluster cohesiveness of topically related documents in the collection while it indicates how such collection characteristics affect the retrieval effectiveness at once. The fact that the feedback gains are always larger when evaluated in "rigid" relevance criteria as seen in Table 2 and 3, suggests that the "rigid" relevant documents are topically more cohesive than the "relax" relevant documents.

4. Conclusions

Our NTCIR-5 evaluation experiments of the CLIR-J-J task have been reported. A TF*IDF approach and a KL-divergence language modeling approach are applied to a Japanese newspaper test collection with aggressive feedback strategies.

We carried out some retrospective studies of the automatic feedback evaluations starting from literatures of TREC-2 to recent TREC and NTCIR collections. Comparative evaluation illustrates some characteristics of the test collections where the feedback is especially effective.

As the next stage, we will examine the feedback effectiveness comparing with the measures for "clustering hypothesis".

Acknowledgments

We thank NTCIR management/secretariat members and CLIR organizers for having admitted our additional submissions.

References

[1] Buckley, C., Singhal, A., Mitra, M. and (Salton, G.) New Retrieval Approach Using SMART:TREC 4. In NIST Special Publication 500-236:The Fourth Text REtrieval Conference (TREC-4), 25-48, 1995.

[2] Eguchi, K., Kuriyama, K. and Kando N. Sensitivity of IR Systems Evaluation to Topic Difficulty. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Vol.2, 585-589, 2002.

[3] Evans,D. and Lefferts,R. Design and Evaluation of the CLARIT-TREC-2 System. In NIST Special Publication 500-215:The Second Text REtrieval Conference (TREC 2), 137-150, 1993.

[4] Fujita, S. Notes on Phrasal Indexing—JSCB Evaluation Experiments at NTCIR AD HOC. In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, 101-108, 1999.

[5] Fujita, S. Reflections on "Aboutness"—TREC-9 Evaluaton Experiments at Justsystem. In NIST Special Publication 500-249:The Ninth Text REtrieval Conference (TREC 9), 281-288, 2000.

[6] Fujita, S. More reflections on "Aboutness"—TREC-2001 Evaluaton Experiments at Justsystem. In NIST Special Publication 500-250:The Tenth Text REtrieval Conference (TREC 2001), 331-338, 2001.

[7] Fujita, S. Revisiting the Document Length Hypotheses --NTCIR-4 CLIR and Patent Experiments at Patolis. In Working notes of the fourth NTCIR workshop meeting, 238-245, 2004.

[8] Fujita, S. Revisiting Again Document Length Hypotheses TREC 2004 Genomics Track Experiments at Patolis. In NIST Special Publication: SP500-

261:The Thirteenth Text REtrieval Conference (TREC 2004), 2004, available at http://trec.nist.gov/pubs/trec13/t13_proceedings.html .

[9] Harman, D. Overview of the Third Text Retrieval Conference (TREC-3). In NIST Special Publication 500-226:The Third Text Retrieval Conference (TREC-3), 1-19, 1994.

[10] Harman, D. Overview of the Fourth Text Retrieval Conference (TREC-4). In NIST Special Publication 500-236:The Fourth Text Retrieval Conference (TREC-4), 1-23, 1995.

[11] Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H. and Hidaka, S. Overview of IR Tasks at the First NTCIR Workshop. In Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, 11-44, 1999.

[12] Kwok,K.L. and Grunfeld, L. TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS. In NIST Special Publication 500-236:The Fourth Text REtrieval Conference (TREC-4), 145-152, 1995.

[13] Lafferty, J. and Zhai, C. Document language models,query models, and risk minimization for information retrieval. In Proceedings of the 2001 ACM SIGIR Conference on Research and Development in Information Retrieval, 111-119, 2001.

[14] Ogilvie,O. and Callan,J. Experiments Using the Lemur Toolkit, In NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001),103-108, 2002.

[15] van Rijsbergen, C.J. Information retrieval. Butterworths, London, 2nd edition, 1979, also available at http://www.dcs.gla.ac.uk/Keith/Chapter.3/Ch.3.html .

[16] Robertson, S. and Walker S. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In Proceedings of the 1994 ACM SIGIR Conference on Research and Development in Information Retrieval, 232-241, 1994.

[17] Robertson, S. E., Walker, S., Jones, S., M.Hancock-Beaulieu, M., and Gatford, M. Okapi at TREC-3. In NIST Special Publication 500-226:Overview of the Third Text REtrieval Conference (TREC-3), 109-126, 1995.

[18] Salton, G. Automatic Text Processing –The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley publishing company, Massachusetts ,1988.

[19] Tao, T. and Zhai, C. A Mixture Clustering Model for Pseudo Feedback in Information Retrieval. In Proceedings of the 2004 Meeting of the International Federation of Classification Societies, 2004.

[20] Voorhees, E.M. The cluster hypothesis revisited. In Proceedings of the 1985 ACM SIGIR Conference on Research and Development in Information Retrieval, 188-196, 1985.

[21] Zhai, C. and Lafferty, J. Model-based feedback in the KL-divergence retrieval model. In Proceedings of the Tenth International Conference on Information and Knowledge Management(CIKM 2001), 403-410, 2001.

[22] Zhai, C. and Lafferty, J. A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of the 2001 ACM SIGIR Conference on Research and Development in Information Retrieval, 334-342, 2001.