Revisiting the Document Length Hypotheses NTCIR-4 CLIR and Patent Experiments at Patolis

Sumio FUJITA
PATOLIS Corporation
2-4-29, Shiohama, Koto-ku,
Tokyo 135-0043, Japan
s_fujita@patolis.co.jp

Abstract

*This paper describes Patolis NTCIR-4 experiments of CLIR and Patent tasks, focusing on comparative studies of two test-collections and two retrieval strategies in view of document length hypotheses. TF*IDF outperformed language modeling approach in the CLIR task while two approaches performed similarly in the Patent task. We assumed two different document length hypotheses behind two collections. Some other task specific techniques are evaluated and reported.*

Keywords: *Information retrieval, Document length hypotheses, Language modeling approach to IR.*

1. Introduction

Patent document retrieval has different characteristics from quotidian document search tasks by subject topics, since it is related to legal activities to claim or to deny/invalidate rights to monopolize certain commercial activities involving uses of technologies described in patent documentation.

Patent documentation is characterized by its special stylistic features as well as highly structured and attributed information. In some aspects, patent documentation is considered as a technico-scientific writings describing technical inventions. NTCIR-3 patent task addressed such features of patent documentation: simulated information needs motivated by newspaper articles and simulated relevance assessments by a group of corporate intellectual property administrators from various industry domains.

On the other hand, an invalidation investigation is not limited to a traditional database retrieval against diverse kind of documentations looking for a prior art possibly invalidating the claim in question but it might be expanded to a sort of "know-who" search where looking for a specialists of the domain who may possibly know disclosure, displays, publications or uses of the invention by products.

Assuming that the claim in question is already granted the rights, applied patent documentations are exhaustively examined by a group of searchers, which leads us to the question such that an invalidation search against patent document collection really make sense.

We may understand the term "invalidation search" in its broader sense as an aspect of patent documentation in comparison with NTCIR-3 like "subject topic search". Such broader definition of text retrieval aspects of invalidation investigation may be applicable to patentability, novelty, validity and infringement investigation adapting different search environment.

Whatever to call, according to the functional roles in the information seeking situations, such types of search tasks require more rigid standards of relevance such that adequacy as an evidential material, than an ordinary subject topic search of technological documentation, that leads to a small number of relevant documents for each query.

From the viewpoints of traditional information retrieval studies, arise the questions as follows:

Is the clustering hypothesis applicable to such a task?

What types of models should be behind document length and its subject topics?

We examined comparatively two types of search tasks: traditional subject topic search against Japanese news paper databases as monolingual runs of the CLIR task and invalidation search against patent application databases as main task runs of the Patent task. We also examined two types of retrieval model namely traditional TF*IDF approach with BM25 TF and Kullback-Leibler divergence (KL-divergence hereafter) approach that is one of the probabilistic language modeling approach recently introduced by some information retrieval researchers[7][16].

In CLIR runs, traditional TF*IDF approach outperformed KL-divergence approach; there might be some technical problems in our KL-divergence runs but we have not yet find out the exact reasons.

It is found that the effectiveness difference caused by different search models is much smaller than an index range of target documents in the collection in our preliminary experiments using NTCIR-3 patent collections. This phenomenon is observed again in the NTCIR-4 patent task.

2. System description

Our evaluation environment: PLLS system developed based on Lemur toolkit 2.0.1 for indexing system[9],

which is being developed by the Lemur project and also based on PostgreSQL system.

The system is operated on a dual CPU PC server(Xeon 2.0GHz, 4GB RAM) running RedHat Linux operating system.

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 indexed as indexing units.

Stop word lists for patent documentation and for newspaper documentation are prepared.

2.2 Retrieval models

The following two retrieval models are examined in two tasks:

-TF*IDF with BM25 TF

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

2.3 Feedback strategies

Pseudo-relevance feedback is applied.

Rocchio feedback for TF*IDF and markov chain query update method for KL-divergence retrieval model[7], are adopted.

3. Language modeling for IR

Uses of probabilistic language models in information retrieval intended to adopt a theoretically motivated retrieval model given that recent probabilistic models tend to use too many heuristics.

Ponte and Croft first applied a document unigram model to compute the probability of the given query to be generated from a document[10].

In TREC-7, Hiemestra and Kraaij[4] introduced linear interpolation of local and global probabilities while Miller et al.[8] used hidden Markov model to mixture two distributions. Berger and Lafferty[1] proposed a statistical translation to model user's distillation process of an information need into a succinct query.

3.1 Basic 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.

3.2 Smoothing methods

Zhai and Lafferty presented that the smoothing method plays a crucial role in language modeling IR [16].

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 proposed three types of smoothing strategies including Jelinek-Mercer method i.e. simple linear combination of an estimated document model and a background model p(w|C), Dirichlet-Prior method that computes maximum a posteriori parameter values with a Dirichlet prior (i.e. a kind of Laplace smoothing), and absolute discount method.

Jelinek-Mercer method is:

$$p_\lambda(w \mid d) = (1-\lambda)\, p_{ml}(w \mid d) + \lambda p(w \mid C)$$

Dirichlet-Prior method is:

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

Smoothing factor in the first case is λ while μ/|d|+ μ in the second case. Document length is taken into consideration in Dirichlet-Prior smoothing, the effects of which will be illustrated in the next sections.

3.3 Document dependent priors and mixture language models

Two language models, which normally represent textual characteristics of each document, can be combined by a parameter λ:

$$p(w \mid d) = (1-\lambda)\, p_{lm1}(w \mid d) + \lambda p_{lm2}(w \mid d)$$

On the other hand, any document dependent and typically query independent factors that may affect the relevance can be taken into consideration by the scoring process as document prior probabilities.

Document length is a good choice in TREC experiments since it is predictive of relevance against TREC test set [8][14].

4. Document length issues

4.1 Document length normalization

Document length normalization is a typical technique adopted by term weighting and query – document matching for document ranking of IR systems. A longer document has more sentences so that the terms have higher frequency than a shorter document as well as it has more likely to have more different terms. Document length normalization prevents the document ranking from matching longer documents penalizing matching scores of longer documents.

If the document length in the search target collection is uniform, no document length normalization is necessary. Since it is generally not true, one way to do this is to split a document into chunks of the same length and to search them. This idea leads to the use of subdocument retrieval in TREC 1[6] and 2[2] experiments.

Because cosine normalization adopted by vector space model[13] in early stage is found out inadequate for test collections of very long documents in TREC evaluation, many TREC systems tend to adopt a revised TF functions like log TF, maximum TF normalization, Okapi TF[12] and pivoted length normalization[14] in order to normalize term frequencies and also to penalize scores of longer documents with more matches.

4.2 Document length hypotheses

Robertson and Walker[11] postulated two hypotheses to model different length of documents namely the "Scope hypothesis" and the "Verbosity hypothesis".

The "Scope hypothesis" considers a long document as a concatenation of a number of unrelated short documents while the "Verbosity hypothesis" assumes that a long document covers the same scope as a short document but it uses more words. These two hypotheses model the extreme cases and real documents are always the mixture of the two cases.

The natural consequence of adopting the Scope hypothesis is that a long document is more likely to be relevant irrespective of search request since it covers more subject topics than a shorter one. Robertson and Walker assume that the Verbosity hypothesis implies that document properties such as relevance and eliteness are independent of document length.

Because longer documents are more informative than short ones even the subject coverage is the same, longer documents are more likely to be relevant even under the Verbosity hypothesis. From another view, the topic is denser in a short document so that it should be given higher score if other matching condition is the same.

4.3 Likelihood of relevance/retrieval in NTCIR-3

To validate the document length hypothesis of different types of document collections, NTCIR-3 CLIR and Patent test collections are examined by applying the analyses against TREC test collections by Singhal et al[14].

NTCIR-3 CLIR Japanese document collection(Mainichi newspaper 1998,1999: 220078docs) and Patent document collection (Unexamined Patent Application 1998,1999: 697330 docs) are put into bins of 1000 documents in the order of the length of documents counted by the number of indexed terms. The last bins(221^{st} and 698^{th}) contain the longest 78 docs and 330 docs respectively.

We utilized 2538 "query-relevant document" pairs for 42 topics of CLIR test collection and 2311 "query-relevant document" pairs for 31 topics of Patent test collection. Partially relevant documents are included in these pairs in order to augment the data. From these pairs, p(d in Bin_i| d is relevant) for each i-th bin is computed.

From 42000 "query-retrieved document" pairs of CLIR collection and 31000 "query-retrieved document" pairs of Patent collection, p(d in Bin_i| d is retrieved) is computed.

Figure 1 shows p(Bin|Relevant) and p(Bin|Retrieved) by TF*IDF and Dir-Prior in NTCIR-3 Japanese CLIR collection, plotted against the median document length in the bin, and Figure 2 in NTCIR-3 Patent collection.

In the CLIR collection, approximation curves of plotted dots by a linear function indicate that TF*IDF retrieval ratio is almost overlapped on the ratio of relevance while no clear correlation is observed in Patent collection.

Different document length hypotheses might be assumed for these evaluation environments. Newspaper documents are typically the case of scope hypothesis while patent documents may be seen as a case of verbosity hypothesis. As required by the "Unity of Invention" principle, a patent document is about a single subject so that the document length may not affect relevance or elitness.

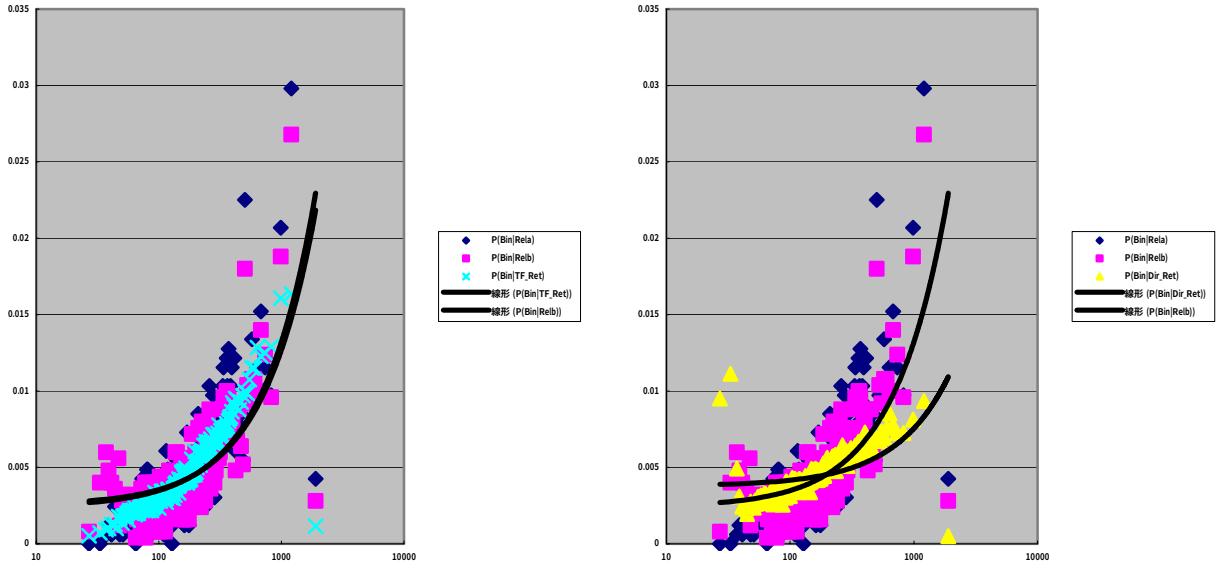

Figure 1: p(Bin|Relevant) and p(Bin|Retrieved) by TF*IDF(Left) and Dir-Prior(Right), plotted against the median bin length in NTCIR-3 Japanese CLIR Collection

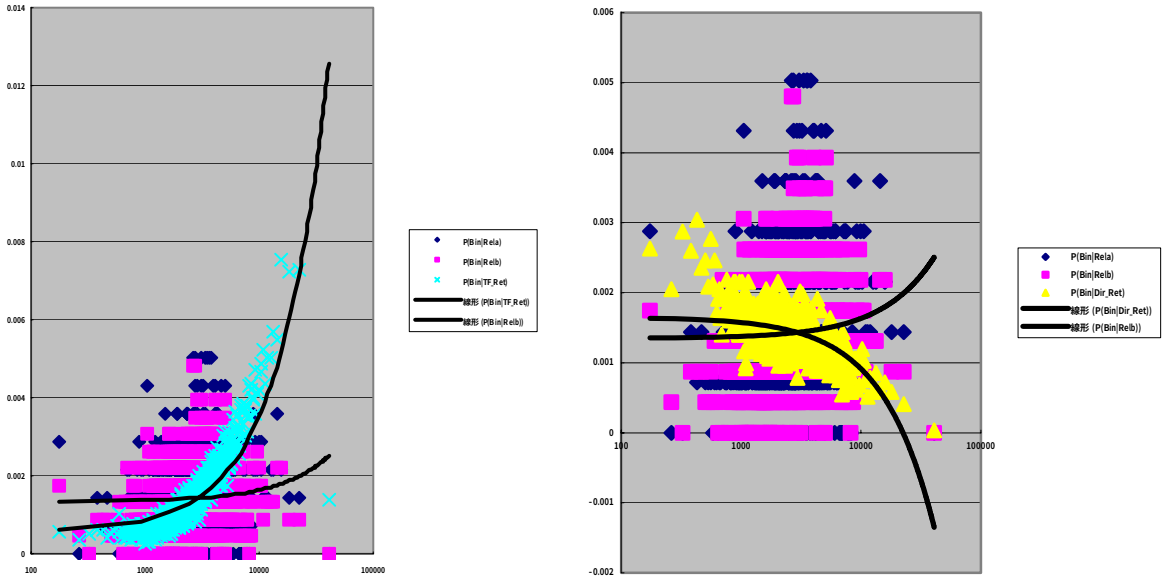

Figure 2: p(Bin|Relevant) and p(Bin|Retrieved) by TF*IDF(Left) and Dir-Prior(Right), plotted against the medial bin length in NTCIR-3 Patent Collection

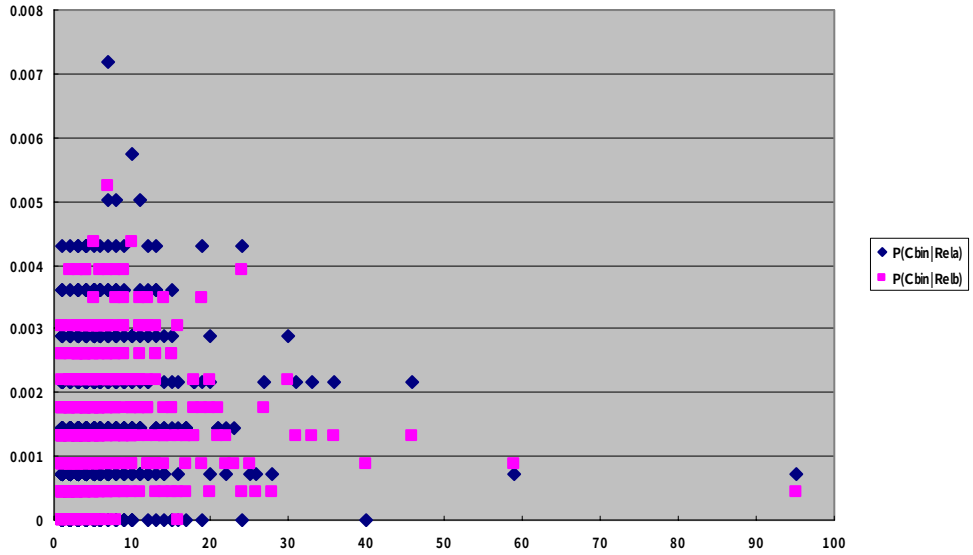

Figure 3: p(Bin|Relevant) plotted against the median number of claims in the bin, in NTCIR-3 Japanese Patent Collection

We tried another analysis using the number of claims in a patent document instead of the number of terms. Figure 3 shows p(Bin|Relevant) for 698 bins plotted against the median claim numbers in each bin. Observing no clear correlation between the number of claims and relevance suggests that a large number of claims do not necessarily signify many scopes of subject topics of the document that may affect relevance.

5. CLIR experiments

5.1 CLIR official runs for J-J SLIR

We submitted a title only run, a description only run and two title & description runs of Japanese monolingual retrieval setting.

The title only run and the description only run are using TF*IDF method with BM25 TF[12] and Rocchio pseudo-relevance feedback.

Since one of our aims is to compare retrieval effectiveness across different ad hoc search tasks, the strategies are very orthodox.

$$w(d,t) = (k4 + \log\frac{N}{df(t)})\frac{(k1+1)\,freq(d,t)}{k1((1-b)+b\frac{dl_d}{avdl}) + freq(d,t)}$$

d : document

t : term

N : total number of documents in the collection

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

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

Such BM25 TF weighting is applied successfully to TREC web ad hoc search task characterized by very short queries and various lengths of documents[3].

TD-01 and TD-02 runs, which are title and description runs, are fusion of T-03 and D-04 with different mixture parameters.

$$score = (1-\alpha)TitleRunScore + \alpha DescRunScore$$

α is either 0.43(TD-01) or 0.5(TD-02) respectively.

Table 1 shows the effectiveness of official runs.

5.2 Post submission experiments

Table 2 compares 4 experimental runs with Jelinek-Mercer smoothing / Dirichlet Prior smoothing. Their MAPs are all far below those of the baseline TF*IDF runs(T-03 and D-04).

Pseudo feedback is performed by interpolating pseudo relevant document models with the original query models. Pseudo relevant document models are distillated by eliminating background noises using EM iteration as described by Zhai and Lafferty [15].

We suspect one of the reasons of the failure as long document preference of relevance judgment observed in NTCIR-3 CLIR test collection. In order to validate this hypothesis, we will apply document length priors and promote matching scores of longer documents. We will report if any improvement is achieved until the workshop.

	AP-Rigid	RP-Rigid	AP-Relax	RP-relax
PLLS-J-J-TD-01	0.3915	0.4100	0.4870	0.4975
PLLS-J-J-TD-02	0.3913	0.4098	0.4878	0.4986
PLLS-J-J-T-03	0.3801	0.3922	0.4711	0.4783
PLLS-J-J-D-04	0.3804	0.3978	0.4838	0.4931

Table 1: Effectiveness of CLIR official runs

	AP-Rigid	RP-Rigid	AP-Relax	RP-relax
JMSmooth λ=0.45 TITLE	0.2696	0.3025	0.3756	0.4077
JMSmooth λ=0.55 DESC	0.2683	0.3110	0.3703	0.4146
DirSmooth μ=1000 TITLE	0.3145	0.3445	0.3990	0.4313
DirSmooth μ=2000 DESC	0.3006	0.3311	0.3907	0.4226

Table 2: Effectiveness of CLIR unofficial runs with JM Smoothing and Dirichlet Prior Smoothing

6. Patent official runs

We submitted six mandatory runs of full-auto query construction.

TF*IDF runs utilize the same scoring as CLIR runs. KL-divergence runs utilize the scoring method described in the early in this paper. Pseudo-relevance feedback is applied in all official runs.

6.1 Evaluation measures

Three topic sets(main, additional and all), three different relevance judgment set(relevant/partially relevant by JIPA assessors, JPO citation set) and two measures (the mean average precision and the average search length based measure) lead to a combinatorial explosion of evaluation results such that as many as 20 scores (consequently different ranks amongst submitted runs) are given for each run.

The sources of unstable inter-system ranking seem to be co-existence of a small number of relevant documents and unstable judgment. This may make the test collection inadequate for evaluating retrieval techniques such as term weighting typically and intensively studied in past test collection based evaluations. Though some technical points that are found to have made big differences are analyzed in the next sub-sections.

6.2 Indexing range: full text vs selected fields indexing

PLLS1 to PLLS5 use abstract and claim fields indexing while PLLS6 uses full text indexing.

Indexing range seems to be a crucial factor in patent document search as well as in more traditional retrieval tasks.

NTCIR-3 Patent task revealed the predominacy of the full-text indexing over the selective indexing. This seems to be the case in NTCIR-4 as well, which was a big misleading for us. We spent most of preparation time for tuning the system to perform best against selected indexing databases but these runs are outperformed by full text indexing runs: PLLS6 in our submission and also many runs submitted by other groups.

6.3 Distributed retrieval strategy for grid computing vs centralized retrieval

PLLS6 used a simple score merge strategy of 5 document collections partitioned by published year of documents. This strategy enables the search process to be decomposed into retrieval against each small sub-set of the collection, and finally result lists from many small sub-set retrieval are merged into a combined list and cut off at a certain number of documents.

Each retrieval process can be completely independent and no statistics information should be propagated through the network. This simplicity makes a big advantage when applied to a grid style highly distributed computing environment not only the search time but also separately managing a large volume of collections.

In TF*IDF approach, IDF and document length normalization use global collection statistics and these make difficult to decompose retrieval process into sub collection search. RSV is not comparable through different collections. In KL-divergence language modeling approach, background language models p(w|C), which are global statistics, affect the score comparability across different collections. Even though, the KL-divergence approach seems to be robust in view of score merging.

TF*IDF baseline of PLLS6 achieved MAP of 0.1703, which is almost same as the runs against selected index (PLLS1-5).

Because of technical problems in indexer program, the baseline centralized retrieval is not yet evaluated (hopefully we will report it until the workshop).

It is also worth trying to use a shared background model p(w|C) for different partitioned collections, making the score more comparable to each other.

6.4 KL-Divergence vs TF*IDF

Comparing MAPs of PLLS1, best performed TF*IDF, with PLLS3, KL-divergence both against selected indexing,

PLLS1 is slightly better in 3 evaluation points(main_rel.a, all_rel.a and main_rel.b) and PLLS3 is also slightly better in other 3 points(add_rel.a, add_rel.b and all_rel.b).

Comparing them by other evaluation measures also gives an impression that there is no big difference in effectiveness between them. As seen in the analyses of probabilities of relevance/retrieved made in the previous sections, there seems to be no specific advantage of TF*IDF against KL-divergence in the Patent collection.

6.5 Pseudo-feedback vs no feedback

Pseudo relevance feedback is performed by so-called "markov chain method" proposed by Lafferty and Zhai[7], which consists of computing p(w|q,R(q)) given a set of relevant or pseudo-relevant documents R(q) as follows:

$$p(w \mid q, R(q)) \propto p(w) \sum_{d \in R(q)} p(d \mid w) p(q \mid d)$$

The baseline MAP(of main_rel.a set) is 0.2094 and PLLS6 is 0.2408(+15.0%).

6.6 IPC priors vs uniform priors

As the document dependent prior probability to compute p(q|d), International Patent Classification(IPC hereafter)[5] code attributed to the documents are used.

First, in order to estimate a IPC of given search topic, top n documents in the results list are examined and for each IPC c, P(c|q,R(q)) is estimated.

The documents attributed IPC c in the result list are promoted according to this estimation.

For PLLS6, where IPC priors are applied, MAP of baseline runs are 0.2347(main_rel.a) and 0.1702(main_rel.b). PLLS6 gets +2.5% gain in A judgment and -1.0% in B judgment.

One of the reasons why the method was not very successful is supposed that a significant change of IPC system had been effectuated at 1995, just middle of the duration of document collections.

6.7 Slope weighting over positions in a claim

Regarding the stylistic features of claim sentences especially such as Jepson style where novelty elements appear after the introductory statements preceded by transition words, terms are re-weighted according to the first position they appeared in the claim such that the term appearing later gets more weight.

This heuristics seemed to give a slight improvement in pre-submission experiments but it is not the case in official runs.

The baseline MAPs without the heuristics against PLLS6 are 0.2410(main_rel.a) and 0.1618(main_rel.b). The improvement of MAPs are –0.1%(main_rel.a) and +4.1%(main_rel.b).

	Main_rel.a	Add_rel.a	ALL_rel.a
PLLS1(tfidf,sel)	0.1734	0.0499	0.0907
PLLS2(tfidf,sel)	0.1628	0.0355	0.0775
PLLS3(kl,sel)	0.1548	0.0557	0.0884
PLLS4(tfidf,sel)	0.1661	0.0492	0.0877
PLLS5(kl,sel)	0.1537	0.0553	0.0878
PLLS6(kl,full)	0.2408	0.0971	0.1445

Table 3: Effectiveness(MAP) of Patent official runs(A)

	Main_rel.b	Add_rel.b	ALL_rel.b
PLLS1(tfidf,sel)	0.1625	0.0537	0.0904
PLLS2(tfidf,sel)	0.1625	0.0396	0.0809
PLLS3(kl,sel)	0.1565	0.0574	0.0908
PLLS4(tfidf,sel)	0.1597	0.0531	0.089
PLLS5(kl,sel)	0.1526	0.057	0.0892
PLLS6(kl,full)	0.1685	0.0988	0.1223

Table 4: Effectiveness(MAP) of Patent official runs(AB)

7. Conclusions

Patolis NTCIR-4 evaluation experiments of CLIR and Patent tasks have been reported.

Document length issues of different collections are examined using NTCIR-3 CLIR and Patent collections and different document length hypotheses are assumed.

A TF*IDF approach and a KL-divergence language modeling approach are applied to two test collections with different characteristics and different search tasks.

Comparative evaluation suggests that we have not yet achieved successful application of the language modeling approach to these tasks, especially finding a good document priors and applying a background language model of the whole collections to make scores comparable are the must.

As the next stage, we will examine two test collections and two retrieval tasks in view of clustering hypothesis. Document characteristic such as an adequacy as a citation invalidating a claim is supposed to have much narrower extension than topical relevance, result sets may not be suitable for clustering analysis by topical aboutness. This may lead us to another hypothesis behind the retrieval approaches.

8. Acknowledgments

Our thanks to NII-NTCIR projects for providing us of NTCIR-3 CLIR and Patent test collections.

9. References

[1] Berger,A. and Lafferty,J. Information retrieval as statistical translation. In *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, 222-229, 1999.

[2] 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.

[3] 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.

[4] Hiemstra, D. and Kraaij, W. Twenty-one at TREC-7: Ad-hoc and cross-language track. In *NIST Special Publication 500-242:The Seventh Text REtrieval Conference (TREC 7)*, 227-238, 1998.

[5] International Patent Classification (IPC). http://www.wipo.int/classifications/fulltext/new_ipc/T

[6] Kwok,K.L. , Papadopoulos,L. and Kwan,K.Y.Y. . Retrieval Experiments with a Large Collection using PIRCS, *NIST Special Publication 500-207:The First Text REtrieval Conference (TREC-1)*, 153-172, 1992.

[7] 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.

[8] Miller, D. H., Leek, T., and Schwartz, R. A hidden Markov model information retrieval system. In *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*, 214–221, 1999.

[9] 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.

[10] Ponte, J. and Croft, W. B. A language modeling approach to information retrieval, In *Proceedings of the 1998 ACM SIGIR Conference on Research and Development in Information Retrieval*, 275–281, 1998.

[11] 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.

[12] 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.

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

[14] Singhal, A., Buckley, C., and Mitra, M. Pivoted document length normalization. In *Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval*, 21–29, 1996.

[15] 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.

[16] 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.