Revisiting 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

*NTCIR-4 experiments of CLIR J-J and Patent tasks, focusing on comparative studies of two test-collections and two retrieval approaches in view of document length hypotheses are described. TF*IDF outperformed the language modeling approach in the CLIR J-J task while two approaches performed similarly in the Patent task. Two different document length hypotheses behind two tasks/collections are assumed by analyzing document length distributions of relevant / retrieved documents in NTCIR-3 and 4 collections. Given these hypotheses, TF*IDF is easily adapted to patent retrieval tasks. Document length priors are applied to the language modeling approach. For the patent task, task specific techniques such as IPC priors and position weighting are evaluated and reported. To facilitate the retrieval against large patent collections, a simple distributed search strategy is applied and found to be efficient despite the slight deterioration of effectiveness.*

Keywords: *Retrieval models, Test collections, Document length hypotheses, Language modeling approach to IR*

1. Introduction

Patent documentation 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 the use of the technologies described in documentation.

Patent documentation is characterized by its special stylistic features as well as highly structured and attributed bibliographic information. In some aspects, patent documentation is considered as techno-scientific writing describing technological 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 specialist of the domain who may possibly know disclosure, displays, publications or uses of the invention by products.

The term “invalidity search” is understood in its broader sense as an aspect of patent documentation. Information seeking efforts are similar to “technology survey” task i.e. subject topic search, but it retrieves documents describing inventions with high resemblance in their essence. Such broader definition of text retrieval aspects of invalidation investigation may be applicable to patentability, novelty, validity and infringement investigation when adapting to different search environments.

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

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

Is the clustering hypothesis applicable to such a task?

What types of hypothesis should be behind different document length?

As having been suggested by some experiences at the NTCIR-4 workshop experments, we focus on the second issue in this paper.

We examined comparatively two different search tasks: a traditional subject topic search against Japanese newspaper collections as monolingual runs at the NTCIR-4 CLIR J-J task and an invalidity search against a patent application collection as main task runs of the NTCIR-4 Patent task. Our examination also targeted to different retrieval models namely a traditional TF*IDF approach with BM25 TF and the Kullback-Leibler divergence (KL-divergence hereafter) approach that is one of the probabilistic language modeling approach recently introduced by some information retrieval researchers [16][29].

In our CLIR J-J runs, TF*IDF runs outperformed the KL-divergence runs while both perform similarly in patent runs and the difference in indexing range affects much more the retrieval effectiveness than the retrieval models.

The rest of the paper is organized as follows:

Section 2 briefly explains the NTCIR-3 and -4 Test Collections.

Section 3 describes our experimental system to NTCIR-4 evaluation experiments.

Section 4 explains the language modeling approach for information retrieval(IR hereafter).

In Section 5, we discuss the document length issues in IR and analyze the NTCIR Test Collections in view of document length hypotheses.

Section 6 describes the experiments of newspaper retrieval using NTCIR-4 CLIR-J-J test collection and Section 7 describes patent document retrieval experiments using the NTCIR-4 Patent Test Collection.

2. NTCIR Test Collections

The NTCIR(NACSIS-NII Test Collection for Information Retrieval Systems, for details see, for example, [13]) project group, organizing series of "distributed experiments/centralized evaluation" style workshops, provides workshop participants(and also other researchers) of reusable test collections for experimental research works of diverse information access technologies, consisting of diverse kinds of document collections including scientific paper abstracts, newspapers, web and patent documentation, mainly in Japanese but also some in English, Chinese and Korean.

In this paper, we report our experiments at the NTCIR-4 Workshop, where we participated in the CLIR J-J Task and Patent Task , together with some analyzing studies on the NTCIR-3 CLIR-J-J subtask and NTCIR-3 Patent task collections.

2.1 CLIR-J-J Test Collections

The NTCIR-3 CLIR-J-J Test Collection consists of 1998-99 Mainichi newspaper documents (220,078 documents) and 42 search topics with assessed document lists [3].

In the NTCIR-4 CLIR-J-J Test Collection, Yomiuri newspaper documents (375,980 documents) are added on top of the NTCIR-3 CLIR-J-J and 55 search topics with assessed document lists [14].

Each topic has four fields namely Title, Desc, Narr and Conc.

Task participants are asked to submit at least one run using only the TOPIC filed and another run using only the DESC field as mandatory runs.

2.2 Patent Test Collection

The NTCIR-3 Patent Test Collection consists of the documents of 1998-99 unexamined patent application (full text, SGML formatted 697,330 documents) released from Japanese Patent Office (JPO hereafter) and 31 topics cited from Mainichi newspaper articles [11].

The NTCIR-4 Patent Test Collection consists of the documents of 1993-1997 unexamined patent application (1,707,184 documents), 34 main topics for which pooling and relevance assessments were carried out, and 69 additional topics for which no relevance assessment was done but JPO citations are used as relevant documents [6]. JPO citations are patent document references to justify the rejection of the original patent application.

Both topic sets are independent claim sentences extracted from patent documents.

Task participants are asked to submit at least one run utilizing only the claim part of the topic patent application.

The organizers provided also participants of so-called the "Search Report Data; 2001, 2002, 2003", which consisted of 115 records of the search reports prepared by professional patent search intermediaries, and were used by patent examiners at the JPO as reference data for patent examination.

3. System description

Our evaluation environment: the PLLS system developed based on the Lemur toolkit 2.0.1 for indexing system [20]; the PostgreSQL RDB system is integrated for treating bibliographic information.

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

3.1 Indexing language

The Chasen version 2.2.9 Japanese morphological analyzer with the IPADIC dictionary version 2.5.1 are utilized for Japanese text segmentation and output single words are indexed excluding stop words.

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

Since the system is not scalable enough to index entire textual contents of the whole Patent Collection at once, the collection is partitioned into five sub-collections according to the published year. Instead of full-text indexing for the Patent Collection, selected field indexing utilizing only the author abstract and claim fields are prepared as well as subdocument based indexing, which uses each passage in documents marked by the official tool as an independent retrieval unit so that document scores for ranking are decided on the basis of constituent passage scores.

3.2 Retrieval models

The following two retrieval models are examined in the two tasks:
-TF*IDF with Okapi BM25 TF(TF*IDF hereafter)
As will be explained in the section 6, BM25 TF is incorporated in the dot-product matching function between TF*IDF weighted vectors. Typical parameters like k1, b can be adjusted [7].
-KL-divergence of probabilistic language models with Dirichlet prior smoothing(KL-Dir hereafter)
For the KL-divergence model, the Jelinek-Mercer smoothing is also tried as an alternative smoothing method. As a document dependent prior probability described in the section 4, a document length prior probability can be used instead of a uniform probability.
Only for the Patent Test Collection, where claims of patent applications are used as search topics, IPC based document prior probability and a slope weighting over word positions are applied; both of them can be applied to these two retrieval methods.

3.3 Feedback strategies

Pseudo-relevance feedback is applied in both tasks.
Rocchio feedback[25] for TF*IDF and markov chain query update method for KL-divergence retrieval model [16], are adopted. The parameters such as the number of documents for the pseudo relevant set, the number of terms to feedback, some score cutoff threshold values and mixture coefficients of feedback terms against original terms are decided by pre-submission experiments using the NTCIR-3 Test Collections as well as “topic-relevant document” pairs extracted from “Search Report Data; 2001, 2002, 2003”.

4. Language modeling for IR

Uses of probabilistic language models in information retrieval intended to adopt a theoretically motivated retrieval model given that recent probabilistic approaches 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 [21].
In TREC-7, Hiemestra and Kraaij [9] introduced linear interpolation of local and global probabilities while Miller et al.[19] used hidden Markov model to mixture two distributions. Berger and Lafferty[1] proposed a statistical translation as a model of user’s distillation process from an information need into a succinct query.

4.1 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 for p(d|q), and eliminating document independent part, we have:

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

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 a 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 a query language model from a document language model in view of ranking documents against the given query.

4.2 Smoothing methods

Zhai and Lafferty presented that a smoothing method plays a crucial role in language modeling IR [29].
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 weighting. 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), Baysean smoothing using Dirichlet Priors 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 the Dirichlet-Prior smoothing: as p(w|C) is divided by the document length, scores of longer documents are more penalized than the Jelinek-Mercer smoothing.

4.3 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.
Some studies suggest that document length is a good choice in TREC experiments since it is predictive of relevance against the TREC test set [19][27].

5. Document length issues

5.1 Why emphasis on document length?

During the submission procedures of the Patent task, we found that the average number of passages of retrieved documents are considerably different, consequently document length as well, depending on adopted retrieval methods. For example, TF*IDF (PLLS2) returned the documents of average 72 passages while KL-Dir (PLLS6), average only 46 passages.

On the other hand, Table 1 compares the effectiveness of some runs of NTCIR-4 CLIR J-J and NTCIR-4 Patent tasks.

	TF*IDF	KL-Dir
NTCIR-4 CLIR J-J	0.3801 (PLLS-J-J-T-03)	0.3145
NTCIR-4 Patent	0.1703	0.2408 (PLLS6)

Table 1: Effectiveness of official runs and their baseline runs

TF*IDF outperforms in NTCIR-4 CLIR J-J while KL-Dir does so in NTCIR-4 Patent.
The following relation is observed:

Retrieved document length: TF*IDF >> KL-Dir

Effectiveness(Newspaper): TF*IDF >> KL-Dir

Effectiveness(Patent):TF*IDF << KL-Dir

All of these suggests that 1) the behavior of these retrieval methods against document length is different and 2) document length characteristics are different in these collections and 3) combining them makes such a reversed order of search effectiveness.
Iwayama et al.[12] compares the document length statistics from the NTCIR-3 Patent Retrieval Collection and CLIR J-J Collection and indicated that the average document length in words of the Patent Collection is 24 times that of the CLIR J-J Collection and the standard deviation of the Patent Collection is 20 times that of the CLIR J-J, consequently the length of patent documents is much more diverse than newspaper documents.
In this paper, we will investigate why the patent documents are different from newspaper documents in length distribution and how such difference affects retrieval effectiveness.

5.2 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 words so that the terms have higher frequency than a shorter document as well as it is 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 “fake” it is to split a document into chunks of the same length and to search them. This idea leads some

researchers to the use of subdocument retrieval in TREC 1 [15]and 2 [4] experiments. This approach is endorsed by the relevance "pact" at TREC, where a document is relevant if it mentions the subject topic of the information need in a portion of the whole document.

Because cosine normalization adopted by the vector space model [26] 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 [23] and pivoted length normalization [27] in order to normalize term frequencies and also to penalize scores of longer documents, which may have more matches.

5.3 Document length hypotheses

The question to be asked here is why longer documents are longer than shorter ones? Though this question may sound as a tautology, it is not. The problem is to know how each document differs in length.

If longer documents have more information, they may be more likely to be relevant against diverse queries, so that it is fair to get a higher matching score.

Robertson and Walker [22] postulated two hypotheses to explain 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 represent 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 requests 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 shorter 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.

A practical question is how much the score should be discounted depending on the document length.

The Okapi probabilistic retrieval model, also known as BM25 [23], uses a document length correction factor as follows when assuming the "verbosity hypothesis".

$$BM25 = \sum_{t \in q} w \frac{(k1+1) freq(d,t)}{k1((1-b)+b\frac{dl}{avdl}) + freq(d,t)}$$

$$+ k2 \mid q \mid \frac{avdl - dl}{avdl + dl}$$

d : document

t : term

q : query

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

w : Robertson / Sparck Jones weight of t in q,

as follows when no relevance informatio n is available

$$\log(\frac{\mathrm{N - df(t) + 0.5}}{\mathrm{df(t) + 0.5}})$$

But in TREC experiments, they always ignore the correction factor by giving 0 to k2. Instead, they lean toward using passage based retrieval assuming the "scope hypothesis".

Fang et al. [5] proposed four formalized retrieval heuristics including two length normalization constraints as follows:

> LNC1: Let q be a query and d1,d2 be two documents. If for some word w $\notin$ q, c(w,d2)=c(w,d1)+1 but for all other word w,c(w,d2)=c(w,d1), then f(d1,q) ≥ f(d2,q)
>
> LNC2: Let q=w be a query. If $\forall$ k>1,|d1|=k|d2| and c(w,d1)=k c(w,d2), then f(d1,q) ≥ f(d2,q)

LNC1 stipulates that the score should decrease when the document has one more non-relevant word. LNC1 requires simply penalizing longer documents and the constraint is normally observed by scoring functions.

LNC2 is to prevent from over-penalizing saying that "if we copy a document k times to form a new document, then the score of the new document should not be lower than the original document". LNC2 is observed by Okapi, consequently by BM25TF but only conditionally observed by KL-Dir or by pivoted normalization.

This suggests that KL-Dir over-penalizes longer documents under some conditions where constraints are violated.

5.4 Likelihood of relevance/retrieved in NTCIR-3

To validate the document length hypotheses of different types of document collections, the NTCIR-3 CLIR J-J and Patent Test Collections are examined by re-applying the analyses against the TREC test collections described by Singhal et al. [27].

The NTCIR-3 CLIR Japanese document collection(Mainichi newspaper 1998,1999: 220078 documents) and Patent document collection (Unexamined Patent Application 1998,1999: 697330 documents) 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(221st and 698th) contain the longest 78 docs and 330 docs respectively.

We utilized 2538 "topic-relevant document" pairs for 42 topics of the CLIR test collection and 2311 "topic-relevant document" pairs for 31 topics of the 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 "topic-retrieved document" pairs of CLIR collection and 31000 "topic-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(Left) and KL-Dir(Right), plotted against the median document length in each bin, in the NTCIR-3 CLIR Japanese Collection, and Figure 2, in the NTCIR-3 Patent Collection.

In Figure 1, approximation curves of plotted dots by each linear function indicate that the ratio of "TF*IDF retrieved"-"document length" (P(Bin|d is Retrieved by TF*IDF)) is almost overlapped on the ratio of "relevance"-"document length" (P(Bin|Relevant)) while the curve of "KL-Dir retrieved"-"document length" (P(Bin|d is Retrieved by KL-Dir)) is below the P(Bin|Relevant). In Figure 2, no clear correlation between "relevance" (P(Bin|Relevant)) and median document length is observed in the Patent collection, although the curve of "relevance"-"document length" (P(Bin|Relevant)) slightly increases. The curve of "TF*IDF retrieved"-"document length" (P(Bin|d is Retrieved by TF*IDF)) increases linearly while the curve of KL-Dir decreases.

Different document length hypotheses might be assumed for these two evaluation tasks. Newspaper documents are typically the case of the "scope hypothesis", like TREC collections, where the longer documents necessarily mention more subject topics.

Patent documents may be seen as a case of the "verbosity hypothesis", where longer documents use more words to describe a specific subject topic. 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.

We tried another analysis using the number of claims in each patent document in the NTCIR-3 Patent Collection 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 "relevance" (P(Bin|Relevant)) and the number of claims suggests that a large number of claims do not necessarily signify many scopes of subject topics of the document, which may affect the "relevance" in the search task for "technology survey".

Finally the NTCIR-4 Patent Collection is examined. 1,707,184 documents are put into 1708 bins; 459 pairs of "topic-relevant document" pairs and 34000 "topic-retrieved document" pairs are used.

The curve of "TF*IDF retrieved"-"document length" (P(Bin|d is Retrieved by TF*IDF)) increases linearly while the curve of KL-Dir(PLLS6) is almost plain.

In summary, TF*IDF always tends to retrieve longer documents and this may be optimal against newspaper documents while KL-Dir(PLLS6) tends to retrieve much shorter documents. KL-Dir seems to be over-penalizing the matching scores of long documents since the approximation curves of P(Bin|d is Retrieved by KL-Dir) is almost plain or even decreasing against document length in the Figure 2.

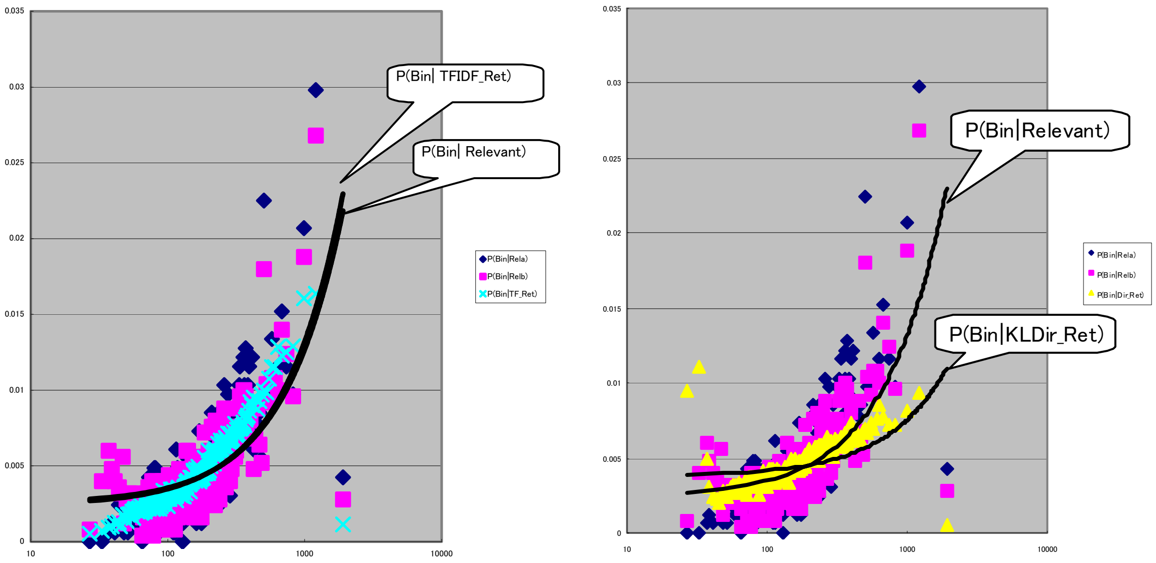

Figure 1: p(Bin|Relevant) and p(Bin|Retrieved) by TF*IDF(Left) and KL-Dir(Right), plotted against the median bin length in the NTCIR-3 CLIR Japanese Newspaper Collection

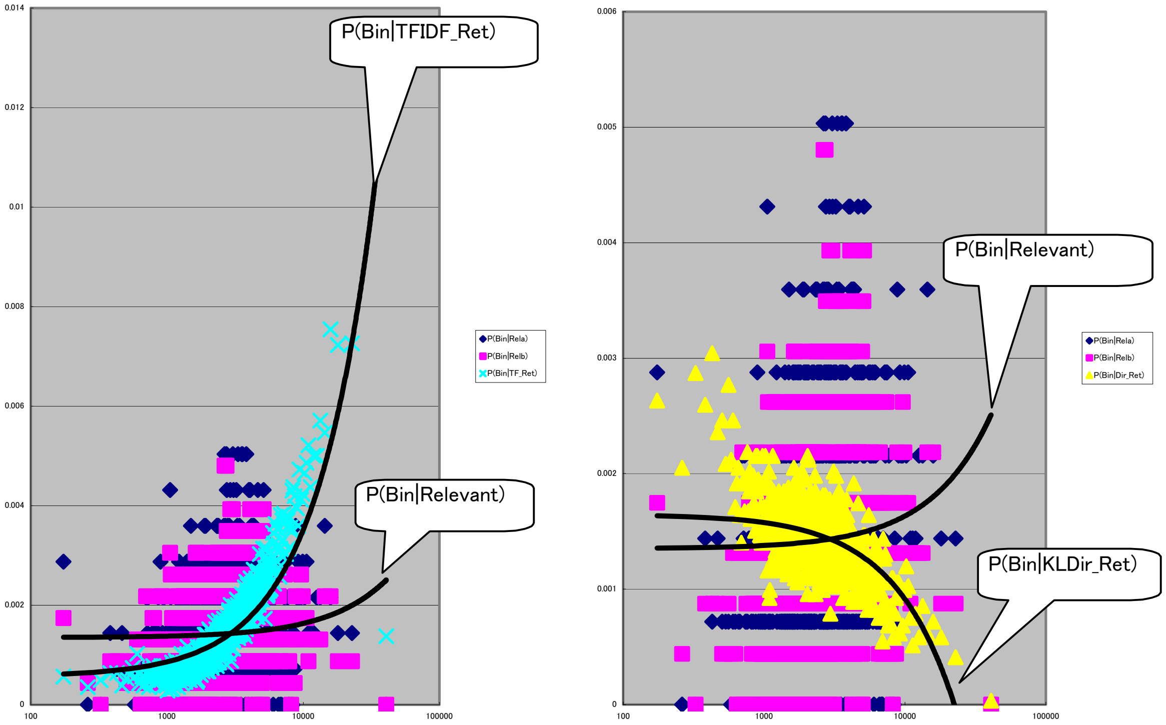

Figure 2: p(Bin|Relevant) and p(Bin|Retrieved) by TF*IDF(Left) and KL-Dir(Right), plotted against the median bin length in the NTCIR-3 Patent Collection

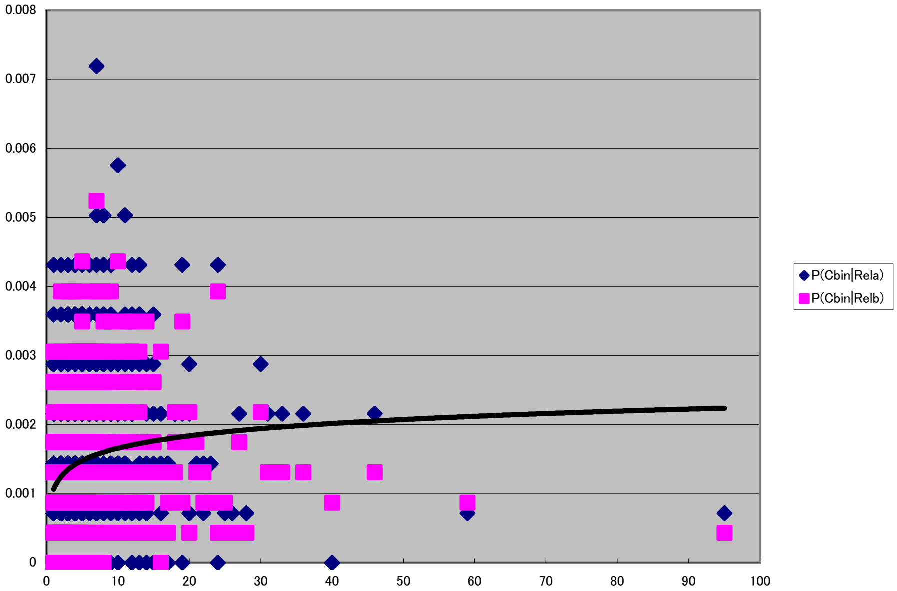

Figure 3: p(Bin|Relevant) plotted against the median number of claims in each bin, in the NTCIR-3 Patent Collection

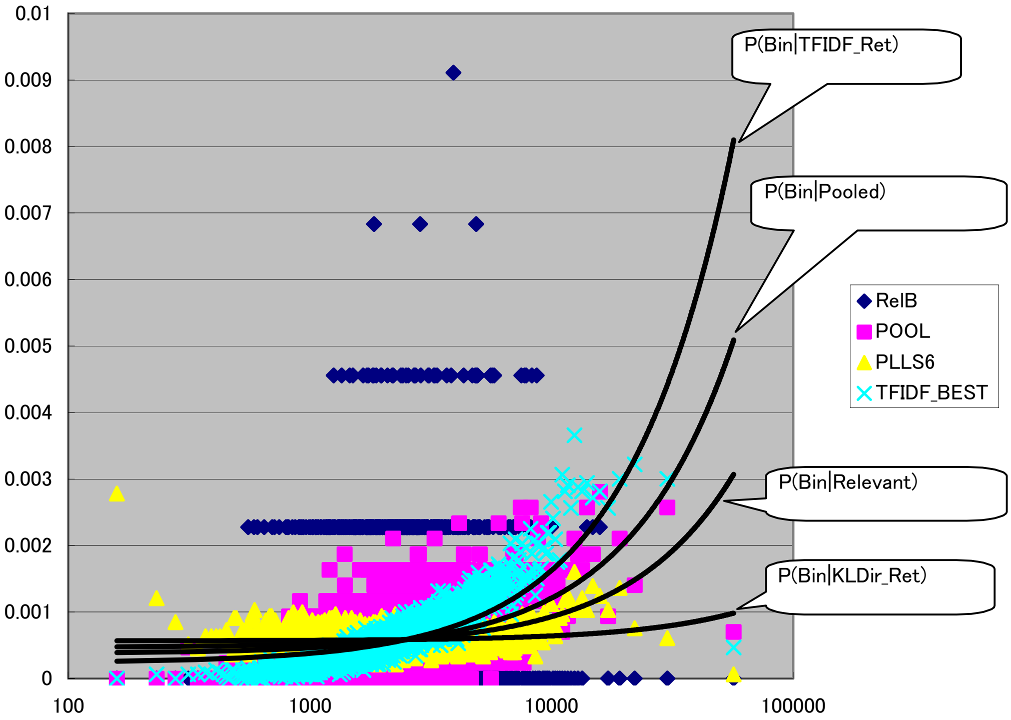

Figure 4: p(Bin|Relevant) , p(Bin|Pooled) and p(Bin|Retrieved) by TF*IDF and KL-Dir, plotted against the median bin length in NTCIR-4 Patent Collection

5.5 Are long patent documents simply verbose?

The question arises here is: if document length does not affect the relevance, why are they long? if they are simply verbose in view of subject topic coverage / topical relevance, how about in view of "invalidating posterior applications"?

Length of newspaper documents is controlled by editorial policies while no control is imposed on patent documents. Statistics from NTCIR-3 and NTCIR-4 patent document collections show that patent documents are getting longer every year as shown in Table 2. Even no restriction on document length, stylistic control is applied, therefore much efforts are required to write a longer patent application. Therefore, there should be some reasons to pay for human efforts to write longer documents with reasonable stylistic quality.

The reasons seem to concern the motivation to write a patent document i.e. to claim rights.

Longer patent documents are stronger because :

1) they can broaden the extensions of the rights covered by the claim by describing more possible application of the invention.

2) They can cover and to describe augmenting complexities of technological domains.

	Average Doclen	Average unique terms
1993	2104	382.3
1994	2320	406.8
1995	2508	419.2
1996	2700	430.5
1997	2759	435.1
1998	2866	444.6
1999	2945	449.8
NTCIR-4 (93-97) average	2478.2	414.8
NTCIR-3 (98-99) average	2905.5	444.8

Table2: Average document length and average unique term numbers in patent documents by year

5.6 Document length hypotheses in summary

Average length of documents in four categories namely relevant(A), relevant or partially relevant(AB), pooled(ABCD), all the documents in the collection(All docs) are listed in Table 3.

The NTCIR-3 CLIR-J-J Test Collection is a typical example of the "scope hypothesis" where relevant documents are 67 points longer than the whole collection in average document length. The NTCIR-3 Patent Test Collection is a case of the "verbosity hypothesis", where relevant documents are only 9 points longer than the whole collection. The NTCIR-4 Patent Collection seems to be the middle of these two typical cases, where relevant documents are 27 points longer than the whole collection. Because longer patent documents possibly contain the description of more prior arts, which may invalidate the application in question.

	NTCIR-3 CLIR-J-J	NTCIR-3 Patent	NTCIR-4 Patent
A docs	315(167%)	3164(109%)	3137(127%)
AB docs	290(153%)	3075(106%)	2946(119%)
ABCD docs	232(123%)	3123(107%)	3321(134%)
All docs	189(100%)	2906(100%)	2478(100%)

Table 3: Average document length of relevant(A), partially relevant(AB), pooled documents(ABCD) and the whole collection(All docs)

6. Newspaper retrieval experiments

Similar to TREC ad hoc retrieval collections, the NTCIR-3 and 4 Japanese newspaper collections are the case of "scope hypothesis" where longer documents preferred relevance is observed.

6.1 NTCIR-4 CLIR-J-J Japanese newspaper retrieval task

At the NTCIR-4 workshop, 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 [23] and Rocchio pseudo-relevance feedback.

Since one of our aims is to compare retrieval effectiveness across different ad hoc search tasks/collections, 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

Instead of the Okapi IDF:
log(N-df(t)+0.5/df(t)+0.5) that gets a negative value when df(t) is very large, we adopted a standard IDF adjusted by the k4 parameter as described in Robertson and Walker [24]. The same weighting is applied for the query part but with a different value for k1 and without length normalization i.e. b=0.
Such a dot-product matching between BM25 TF weighted vectors is applied successfully to TREC web ad hoc search task characterized by very short queries and various lengths of documents where subdocument based retrieval is applied [7]. Web documents seem to be heterogeneous in the sense that they are a mixture of the "scope hypothesis" and the "verbosity hypothesis".
Average document length is 192. Other parameters to be set are as follows:
T-03: k1=1.0, k4=1.0, b=0.35, number of feedback docs=7, number of feedback terms=100
D-04:k1=1.2, k4=1.0, b=0.5, number of feedback docs=10, number of feedback terms=100
TD-01 and TD-02 runs, which are title and description runs, are fusion of T-03 and D-04 with different mixture parameters.

$$score = (1-\alpha)T\_RunScore + \alpha D\_RunScore$$

α is either 0.43(TD-01) or 0.5(TD-02) respectively.
Table 4 shows the effectiveness of official runs. AP indicates mean average precision(also MAP) and RP, R-precision. "Rigid" evaluations utilize only "relevant" documents while "Relax" utilize "relevant" and "partially relevant" documents.

6.2 Post submission experiments

Table 5 compares 4 experimental runs with Jelinek-Mercer smoothing / Dirichlet Prior smoothing. As described in the section 4, the Jelinek-Mercer smoothing is a traditional version of smoothing, which is adopted by some TREC participating groups [9][19], and it tends to retrieve shorter documents than the Dirichlet Prior smoothing. Pseudo feedback is performed by interpolating pseudo relevant document models with the original query models for the sake of the comparison with TF*IDF runs. Language models for pseudo relevant documents are distillated by eliminating background noises using EM iteration as described by Zhai and Lafferty [29].
Their MAPs are all far below those of the baseline TF*IDF runs (T-03 and D-04).
We suspected that one of the reasons of the failure is long document preferred relevance judgment observed in the NTCIR-3 CLIR J-J Test Collection. In order to validate this hypothesis, we will apply document length priors and promote matching scores of longer documents.

6.3 Document length priors

In order to promote the score of longer documents, document prior probability of document length is

	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 4: 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 5: Effectiveness of CLIR unofficial runs with JM Smoothing and Dirichlet Prior Smoothing

incorporated.

$$p(d) = \frac{|d| + \mu}{\sum_{d \in C} (|d| + \mu)}$$

MAPs against the NTCIR-4 CLIR-J-J test collection is as follows:

PLLS-J-J-T-03(TF*IDF):0.3801
Dirichlet :0.3145
Dirichlet with a doc length prior:0.2908(when μ is 1000)
Document length priors did not improve the result at all. Although we have not yet known the exact reasons, a simple promotion by document length does not seem to help even in such an evaluation environment.

7. Patent document retrieval experiments

The NTCIR-3 Patent document collection is a case of the "verbosity hypothesis" while the NTCIR-4 Patent task is the middle of two hypotheses.

7.1 NTCIR-4 Patent task

We submitted six mandatory runs, which uses only the "CLAIM" field, of full-auto query construction.
TF*IDF runs use the same scoring as CLIR J-J runs. KL-divergence runs use the scoring method described in the early in this paper. Pseudo-relevance feedback is applied in all official runs.
Instead of submitting ranked document lists, patent main task participants are asked to rank all the passages on top of each ranked document. Since we focus on the evaluation by document basis, passage ranking and passage based evaluation are ignored in this paper.
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 an average search length based measure) lead to a combinatorial explosion of evaluation results such that as many as 20 evaluation scores (consequently different ranks amongst submitted runs) are assigned 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.
It is controversial to utilize MAPs as the evaluation measure of Patent document retrieval especially of invalidation search where the number of relevant documents is so small that evaluation results may be unstable. Despite such a controversy, we adopt MAPs here for the sake of comparison with newspaper retrieval, and the observations are carefully examined if they are stable enough across different settings. Especially some technical points that made considerable differences in effectiveness are analyzed in the next sub-sections.
Table 6 and 7 show the effectiveness of our patent official runs.

7.2 Indexing range: full text / selected fields / subdocument based indexing

PLLS1 to PLLS5 use abstract and claim fields indexing, called selected field indexing hereafter,

	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 6: 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 7: Effectiveness(MAP) of Patent official runs(AB)

	λ	main_rel.a	main_rel.b
Tfidf,subdoc	0.2	0.1566	0.1618
Tfidf,subdoc	0.5	0.1577	0.1561
Tfidf,subdoc	0.8	0.1655	0.1482
KL-Dir,subdoc	0.2	0.1730	0.1544
KL-Dir,subdoc	0.5	0.1640	0.1547
KL-Dir,subdoc	0.8	0.1603	0.1494

Table 8: Effectiveness(MAP) of unofficial subdocument based runs(MAIN,A,AB)

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.

Although NTCIR-3 Patent task revealed the predominacy of the full-text indexing over the selective indexing [12], it seemed reasonable that the author abstract fields contain essential keywords of the document and claim fields describe the essence of the invention so that these fields can act as a surrogate index of the whole document. In fact some commercial patent full text search services index only these fields. Nevertheless, in order to avoid term miss-matching problems caused by the characteristics of patent terminologies i.e. for example intentional uses of non-standard terminology, idiosyncracy depending on each patent documentalist, full-text indexing is crucial factor in view of high average precision.

This seems to be the case in NTCIR-4 as well, although we underestimated this, and it 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.

Another possibility is to split the whole patent document into chunks of roughly the same length, to retrieve against such chunks and to decide the document score according to the scores from constituent chunks; this strategy, so-called subdocument based retrieval, was successfully adopted by TREC participants [15][4]. This strategy worked as well for the Web document search task at TREC[7], where document length is varied considerably such that some extremely long documents get high matching scores against practically any queries submitted.

In view of document length, a Patent collection is similar to the Web collections, but subdocument based retrieval does not work well.

The document is split into author's abstract field and each passage delimited by the organizers' tool for passage marking. Documents are ranked by the combination of the score of the abstract chunk and the maximum score among all other chunks.

$$score = (1-\lambda) AbstScore + \lambda \, \mathrm{MAX}(ChunkScore)$$

As Table 8 shows, the results are almost the same as the selected field runs. Although subdocument based retrieval assumes the "scope hypothesis" i.e. a document consists of many subject scopes and each chunk split from the whole document falls into one of the scopes, it seems to work under the "verbosity hypothesis" as well.

The reasons for the unsuccessful subdocument based retrieval against the Patent collection seem to be caused by splitting the document into chunks:

1)small chunk based matching is more severely affected by the term miss-match problem.

2)While subject topical relevance is typically represented in local parts of the document, patent search relevance especially invalidation relevance can be scattered through the whole document.

7.3 Distributed retrieval strategy for grid computing vs centralized retrieval

Distributed Selective Search is one way to seek trade-offs between efficiency and effectiveness when retrieving documents from very large collections: the whole collection is partitioned by some criteria like publication date order, author's name order, original document location or content basis classification etc. and stored into separate sub-collections. The search process consists of 1)selecting sub-collections to be searched, 2)distributed searching from all the sub-collections selected, 3)merging the result lists from the selected sub-collections, 4) exhaustive searching against all sub-collections, if the user requests it.

Many studies on distributed retrieval carried out by researchers of the IR society so far tend to more focus on the sub-collection selection (also called database selection) Callan et al. 1995; Larkey 1999; Larkey 2000; Fujita 2001]; failing to properly select the target sub-collections causes severe degradation in effectiveness. Recently, it becomes very important to split an expensive retrieval task into small parts and to compute them on grid style highly distributed computing environment since large collections can be exhaustively searched by a divide-and-conquer strategy on inexpensive PC networks.

PLLS6 used a simple score merge strategy of five sub-collections partitioned by the 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 the retrieval processes against small sub-collections are merged into a combined list and cut off at a certain number of documents. From the data organization viewpoint, partitioning the collection by the published year is preferable because search constraints using the published year are very common in commercial patent retrieval work.

Each retrieval process can be completely independent and no statistics information should be propagated through 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 the collection-wide 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, and cause the score comparability problem across different collections. Even though, the KL-divergence approach seems to be robust in view of score merging. Because of technical problems in the indexer program, we submitted a distributed run; the baseline centralized retrieval is implemented after the submission by merging each sub-collection statistics into collection-wide statistics at the run time. It is also worth trying to use a shared background model p(w|C) estimated somehow to all sub-collections for making the matching score from a sub-collection more comparable to each other.

7.4 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 in view of document length issues.

After submission, we implemented simulated centralized search functionality and carried out comparative evaluation focused on different retrieval models(TF*IDF/KL-Dir) and distributed/centralized search as shown in Table 9.

Comparing by baseline runs i.e. no parameter adjustment after the official submission, KL*Dir outperforms TF*IDF while no significant difference between distributed and centralized search.

Paying special attention on document length normalization factors, we adjusted some parameters as described in the next sub-section and finally achieved the best performance of our system as seen in the row marked "BEST" in Table 9.

Comparing by the best MAPs, there is no statistically significant difference ($p<0.05$) between TF*IDF and KL-Dir, and between distributed and centralized runs. This result of comparing between distributed / centralized search confirms the result reported by Larky[18] where the USPTO patent collections are partitioned into 401 sub-collections according to the chronological order and retrieval results from each sub-collection are merged by some normalization methods, though normalization methods did not affect the effectiveness measured by high-precisions.

	TF*IDF	KL-Dir
Distributed baseline	0.1703	0.2408 (PLLS6)
Distributed BEST	0.2516	0.2488
Centralized baseline	0.1712	0.2274
Centralized BEST	0.2625	0.2508

Table 9: MAPs(main_rel.a) of Distributed / centralized and KL-Dir / TF*IDF runs

7.5 Patent task with document length penalization

In order to achieve our best TF*IDF performance: 0.2625(MAP, centralized), we assigned 0.9 to 1.0 to the parameter b of BM25 TF, which means maximizing the penalization against long documents. The b can be theoretically 0.0 when all the documents in the collection is the same length. If the document length is controlled as under the subdocument based retrieval, 0.2 to 0.3 is assigned to b, as subdocument based Web retrieval described in [7]. Our NTCIR-4 CLIR J-J runs use 0.35 to 0.4 for b. Therefore document length penalization helps very much in patent document retrieval. In other words, the KL-Dir retrieval method, which performs similarly to the best TF*IDF run against the NTCIR-4 Patent Collection, seems to incorporate very strong document length penalization.

We assigned 0.9 to k1 while 1 to 1.2 in NTCIR-4 CLIR J-J; this means a slightly flat TF curve performs better.

A constant query TF performs better than the query part of BM25 TF i.e. typically similar to raw query TF.

7.7 Pseudo-feedback vs no feedback

Pseudo relevance feedback is performed by so-called "markov chain method" proposed by Lafferty and Zhai[16], 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%).

7.8 IPC priors vs uniform priors

As the document dependent prior probability to compute p(q|d), International Patent Classification(IPC hereafter)[10] code attributed to documents in the collection are used.

First, in order to estimate a IPC of the given search topic, top n documents in the result 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. Strictly speaking, this is a heuristic promoter rather than a prior probability of the document but it should work just like a prior probability.

For PLLS6, where IPC priors are applied, MAPs 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.

The method does not achieve not a successful result; presumably because a significant change of IPC system had been effectuated at 1995 i.e. just middle of the duration of document collections.

7.9 Slope weighting over positions in the 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).

As described in the sub-section 7.5, a constant TF, which possibly discounts the weight of repeating words appearing in the introductory part, performs better in TF*IDF runs.

8. Conclusions

A comparative study of Japanese newspaper and patent retrieval using the NTCIR-4 CLIR J-J and Patent collections has been reported with the focus on the document length normalization of retrieval functions.

Document length issues of different collections are examined using NTCIR-3 / -4 CLIR J-J and Patent collections and two document length hypotheses i.e. the reason for the document length variation, are assumed namely the “scope hypothesis” and the “verbosity hypothesis”.

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

In the newspaper retrieval task, TF*IDF with a BM25 TF, which tends to retrieve longer documents, outperforms KL-Dir method while no statistically significant difference is observed in the patent retrieval task. Simple penalization or promotion by document length prior does [10]not improve the performance of KL-Dir.

In the patent document retrieval, a retrieval function with a strong document length penalization i.e. TF*IDF BM25TF with a higher value for the parameter b or KL-Dir, which intrinsically has strong penalization against longer documents, performs well.

Comparative evaluation results suggest that we have not yet achieved a successful application of the language modeling approach to these tasks, especially in newspaper retrieval, an adjustable document length normalization factor intrinsic to the smoothing method are preferably to be incorporated. In patent retrieval, a good document prior probability estimated by, for example, using IPC information, may help. In view of invalidation search, a method using an analyzed query structure to make a better query language model is worth trying.

9. Acknowledgments

Our thanks to the NII-NTCIR projects for providing us of the NTCIR-3/-4 CLIR and Patent Test Collections.

We are grateful to CMU/Umass Lemur Project for making the Lemur toolkit available and also to Dr. Zhai for his helpful comments and advise.

10. References

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

[2] Callan, J.P., Lu, Z. and Croft, W.B. 1995. Searching Distributed Collections With Inference Networks. In Proceedings of *the 1995 ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, 21-28.

[3] Chen, K.H., Chen, H.H., Kishida, K., Kuriyama, K., Kanodo, N., Lee, S., Myaeng, S.H., Eguchi, K. and Kim, H. 2002. Overview of CLIR Task at the Third NTCIR Workshop, In Working notes of the third NTCIR workshop meeting Part I Overview, 23-60.

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

[5] Fang, H. Tao, T. and Zhai, C. 2003. An Exploration of Formalized Information Retrieval Heuristics. In Proceedings of the ACM SIGIR 2003 Workshop on Mathematical/Formal Methods in IR,Toronto, Canada.

[6] Fujii, A. , Iwayama, M. and Kando, N. 2004. Overview of Patent Retrieval Task at NTCIR-4, In Working notes of the fourth NTCIR workshop meeting, 225-232.

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

[8] Fujita, S. 2001. 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.

[9] Hiemstra, D. and Kraaij, W. 1998. 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.

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

[11] Iwayama, M. , Fujii, A. , Kando, N. and Takano, A. 2002. Overview of Patent Retrieval Task at NTCIR-3, In Working notes of the third NTCIR workshop meeting Part I Overview, 67-76.

[12] Iwayama, M. , Fujii, A. , Kando, N. and Marukawa, Y. 2003. An empirical study on retrieval models for different document genres: patents and newspaper articles, In *Proceedings of the 2003 ACM SIGIR Conference on Research and Development in Information Retrieval*, Toronto, Canada, 251-258.

[13] Kanodo, N. 2004. Overview of the Fourth NTCIR Workshop, In Working notes of the fourth NTCIR workshop meeting, i-viii.

[14] Kishida, K., Chen, K.H., Lee, S., Kuriyama, K., Kanodo, N., Chen, H.H., Myaeng, S.H. and Eguchi, K. 2004. Overview of CLIR Task at the Fourth NTCIR Workshop, In Working notes of the fourth NTCIR workshop meeting, 1-59.

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

[16] Lafferty, J. and Zhai, C. 2001. 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*, New Orleans, LA, 111-119.

[17] Larkey, L. S. 1999. A Patent Search and Classification System. In *Digital Libraries 99 - The Fourth ACM Conference on Digital Libraries*, Berkeley, CA, Aug 1999, 79-87.

[18] Larkey, L. S., Connell, M. and Callan, J. 2000. Collection Selection and Results Merging with Topically Organized U.S. Patents and TREC Data. In *Proceedings of Ninth International Conference on Information Knowledge and Management*, Washington D.C., Nov 2000, 282-289.

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

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

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

[22] Robertson, S.E. and Walker S. 1994. 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*, Dublin, Ireland, 232-241.

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

[24] Robertson, S.E. and Walker S. 1997. On relevance weights with little relevance information, In *Proceedings of the 1997 ACM*

SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, 16-24.

[25] Rocchio, J.J. 1971. Relevance feedback in information retrieval, In The SMART Retrieval System: Experiments in Automatic Document Processing, G. Salton ed. Prentice-Hall, Englewood Cliffs, NJ, 313-323.

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

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

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

[29] Zhai, C. and Lafferty, J. 2001. 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*, New Orleans, LA, 334-342.