Revisiting Document Length Hypotheses

NTCIR-4 CLIR and Patent Experiments at Patolis

4 June 2004
Sumio FUJITA
PATOLIS Corporation

Introduction

- Is patent search different from traditional document retrieval tasks?
- If the answer is yes,
  - How different?
  - And why different?
- Comparative study of CLIR J-J task and Patent main task may lead us to the answers.
- Emphasis on document length hypotheses

Why emphasis on document length?

- Because according to the retrieval methods, average number of passages of retrieved documents at NTCIR-4 Patent task are considerably different!
  - PLLS2(TF*IDF): 72
  - PLLS6(KL-Dir): 46
- Effectiveness in NTCIR-4 CLIR J-J(MAP)
  - TF*IDF: 0.3801 (PLLS-J-J-T-03)
  - KL-Dir: 0.3145
- Effectiveness in NTCIR-4 Patent(MAP)
  - KL-Dir: 0.2408 (PLLS6)
  - TF*IDF: 0.1703
- Different document length hypotheses to different tasks?

System description

- PLLS evaluation experiment system
- based on Lemur toolkit 2.0.1[Ogilvie et al. 02] for indexing system
- PostgreSQL integration for treating bibliographic information
- Distributed search against patent full-text collection partitioned by the published year
- Simulated centralized search as baseline

System description

- Indexing language:
  - Chasen version 2.2.9 as Japanese morphological analyzer with IPADIC dictionary version 2.5.1
- Retrieval models:
  - TF*IDF with BM25 TF
  - KL-divergence of probabilistic language models with Dirichlet prior smoothing[Zhai et al. 01]
- Rocchio feedback for TF*IDF and markov chain query update method for KL-divergence retrieval model [Lafferty et al. 01]

Language modeling for IR

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

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

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

Negative cross entropy between the query language model and a document language model

- retrieval version of a Naïve Bayes classifier

Smoothing methods

- Jelinek-Mercer method

Freq(w,d)/|d|

Background probability is not divided by doclen!

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

- Dirichlet-Prior method

Background probability is divided by doclen!

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

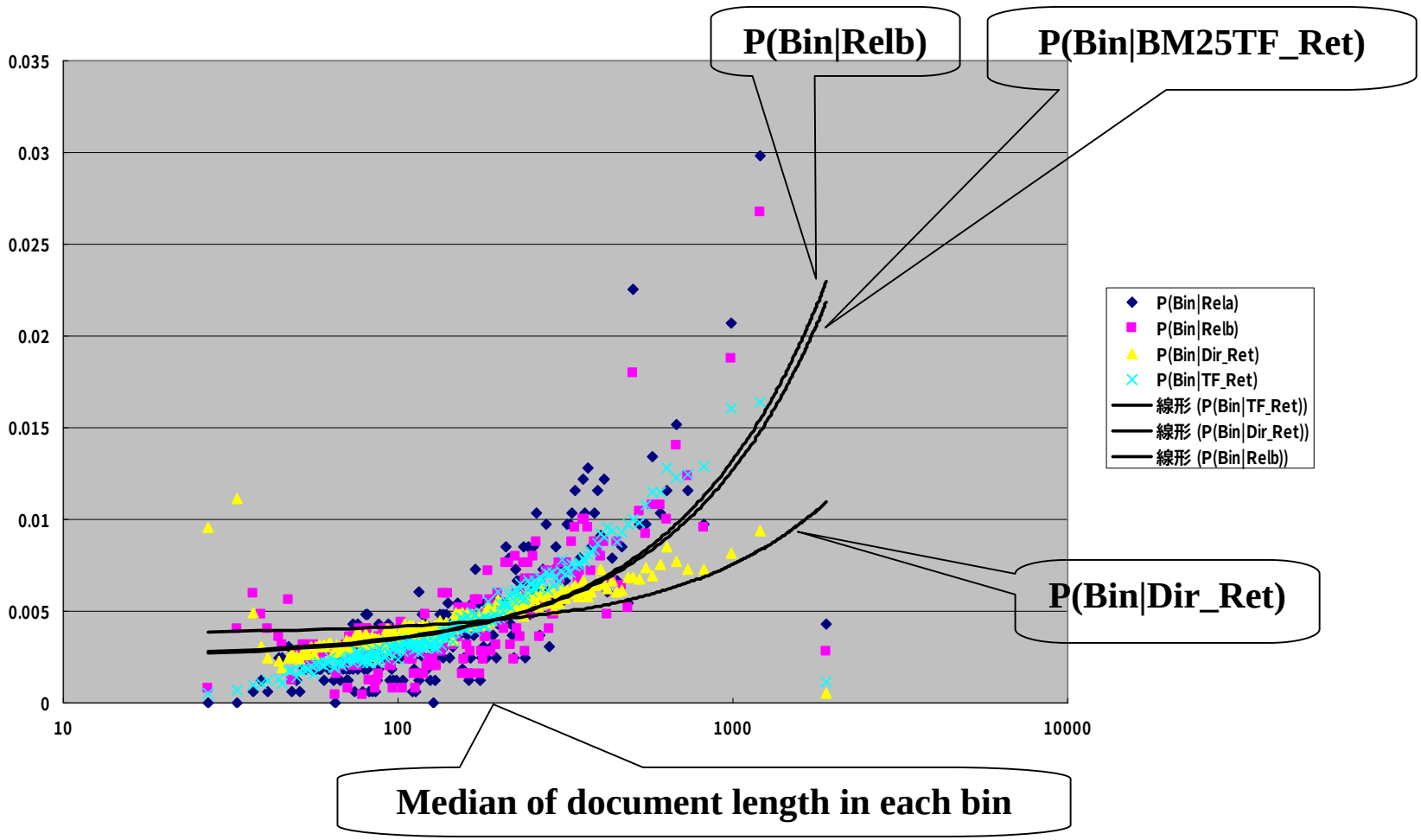
Document dependent priors

- Document length is a good choice in TREC experiments since it is predictive of relevance against TREC test set [Miller et al. 99][Singhal et al. 96].
- Hyper Link Information in Web search
- What are the good priors in Patent search?
  - IPC prior?

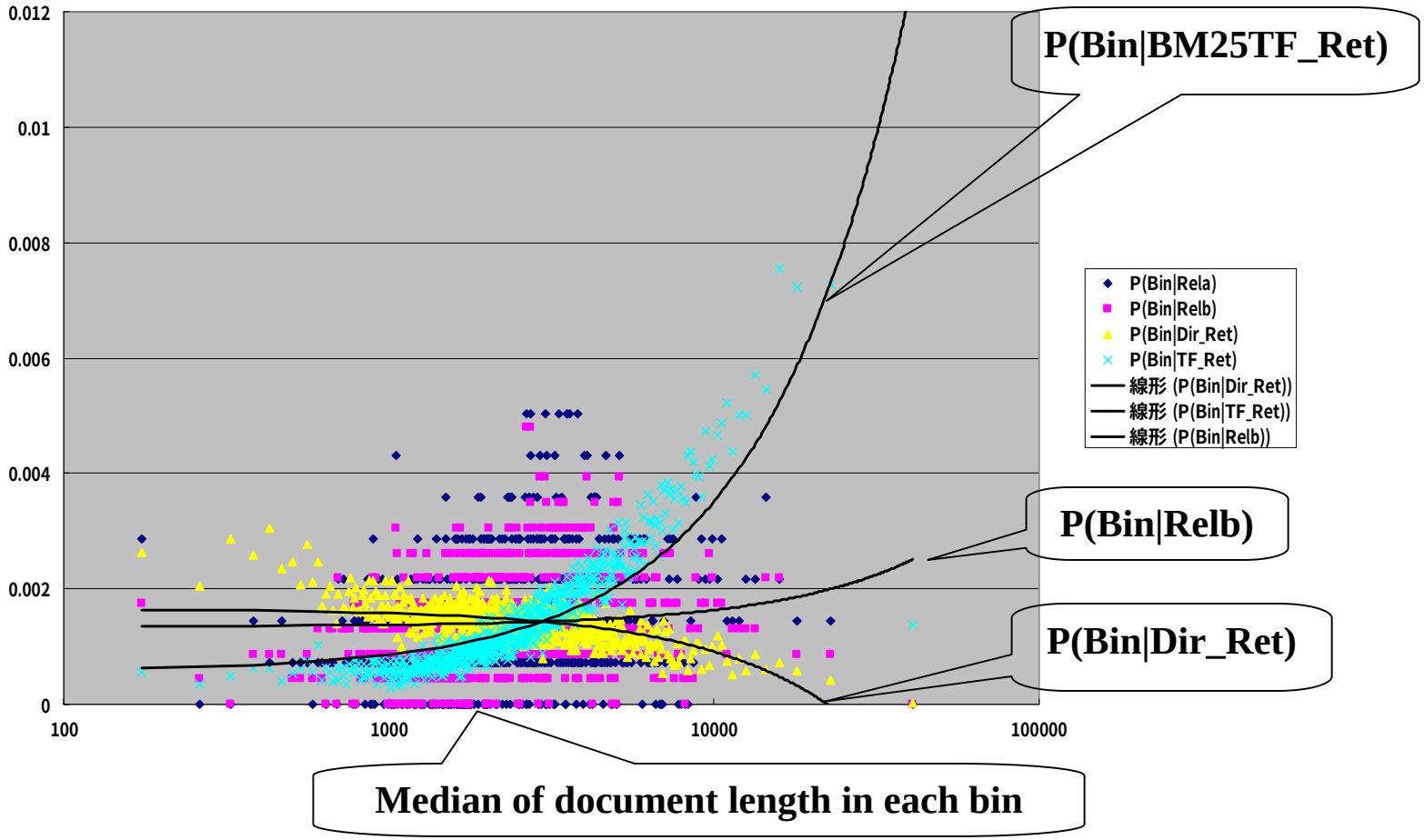
Document length hypotheses

- Why are longer documents longer than shorter ones?
- The “Scope hypothesis” considers a long document as a concatenation of a number of unrelated short documents.
- The “Verbosity hypothesis” assumes that a long document covers the same scope as a short document but it uses more words. [Robertson et al. 94]

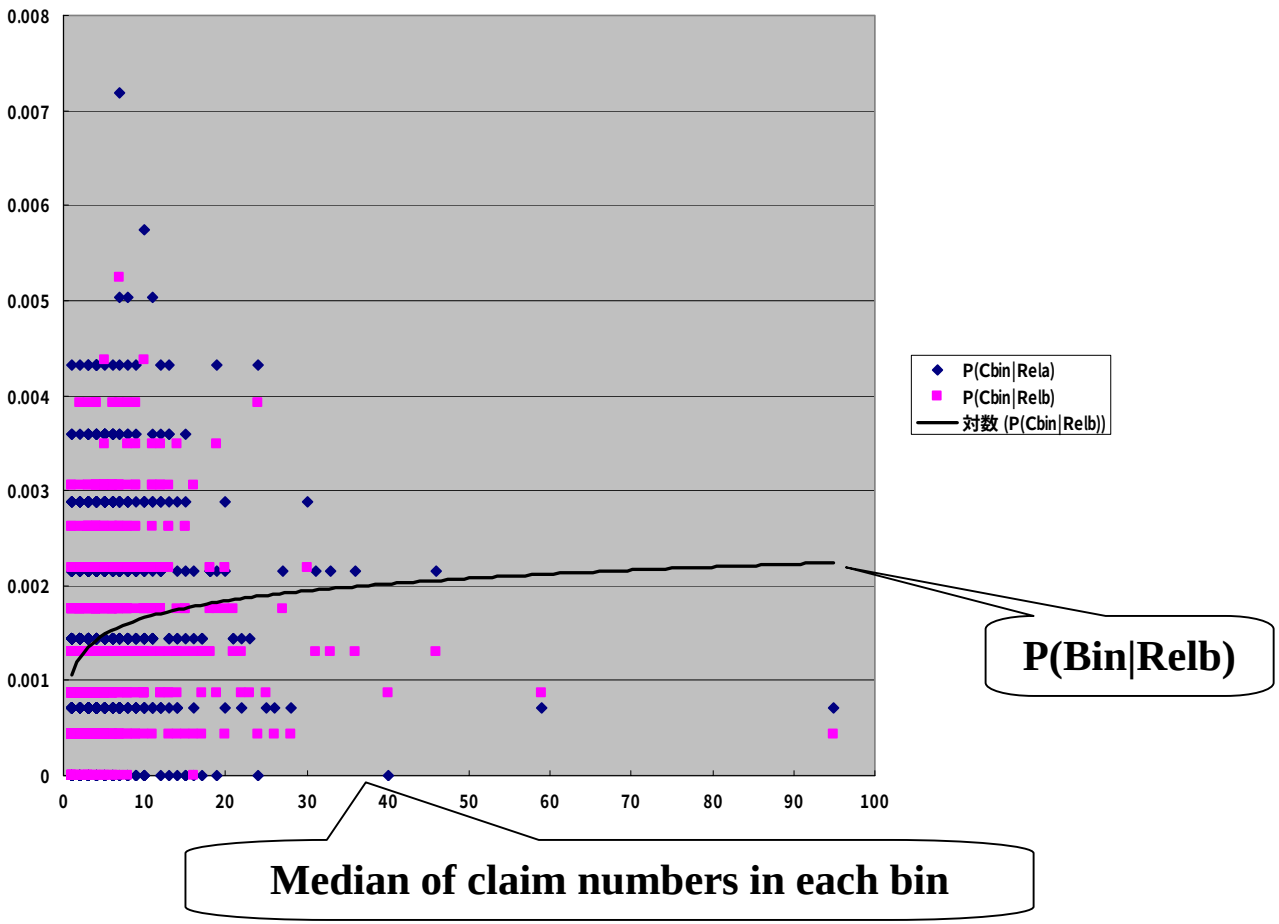
Scope hypothesis (NTCIR-3 CLIR-J-J)



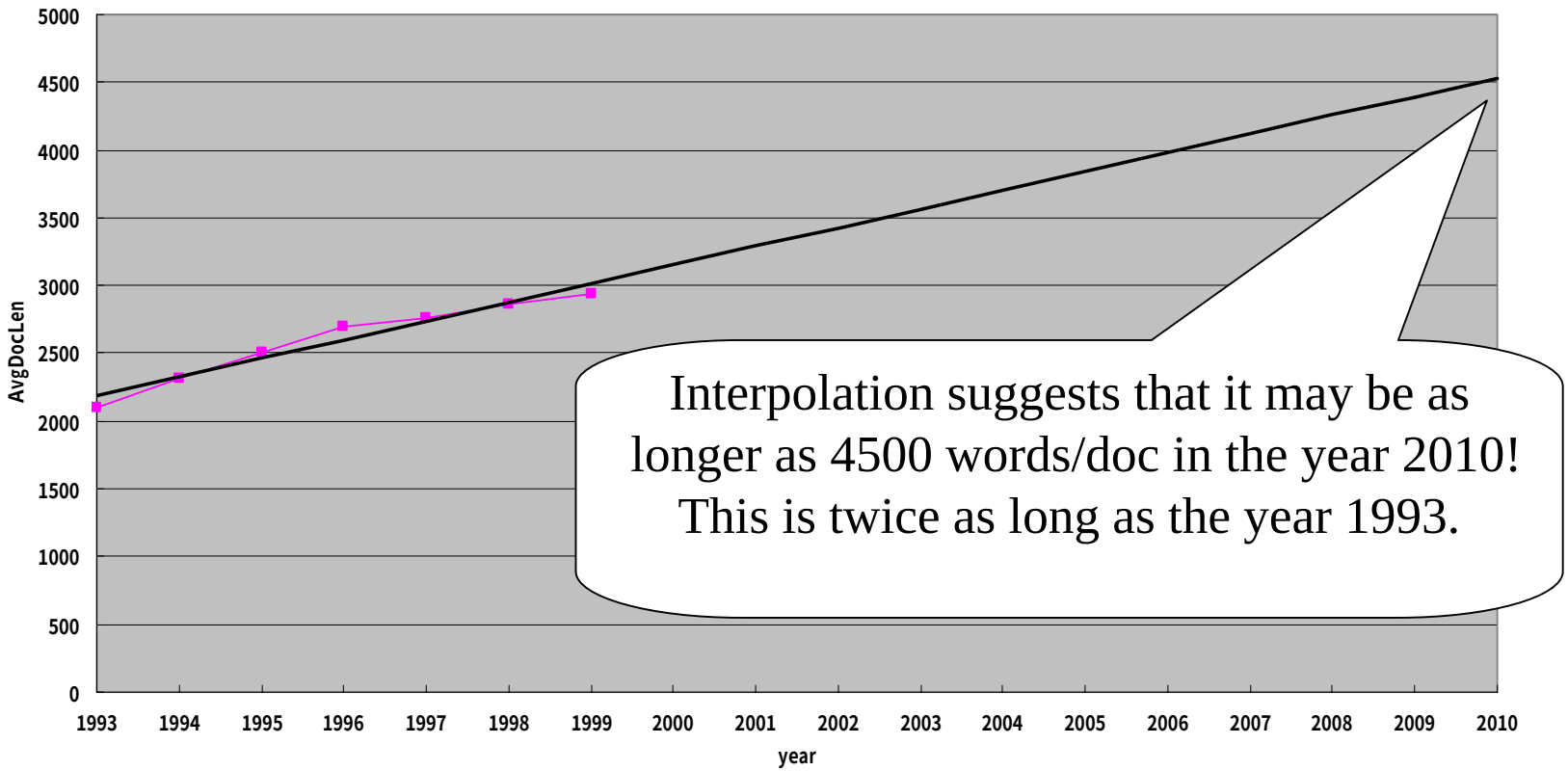
Verbosity hypothesis (NTCIR-3 Patent)



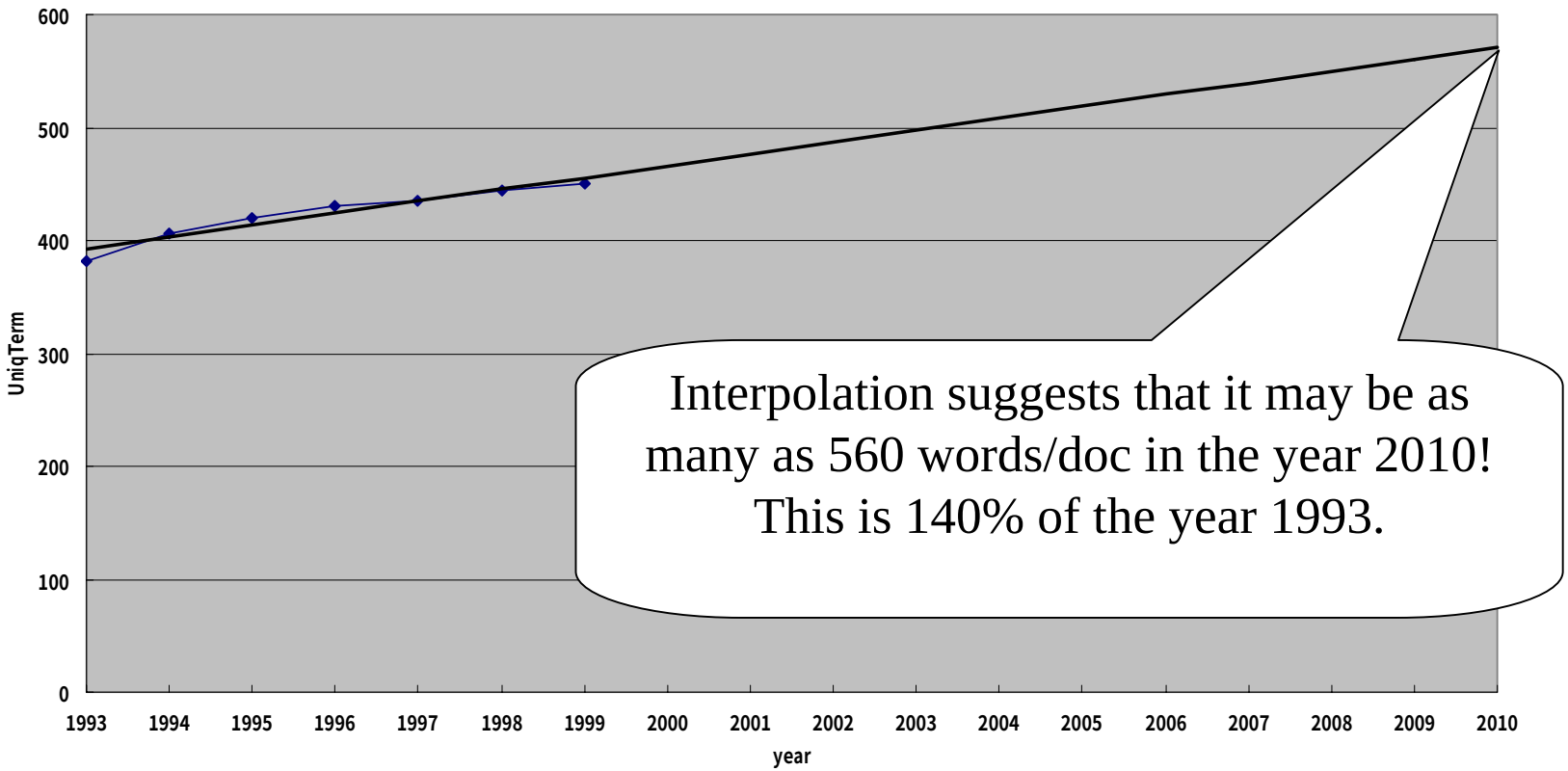
Verbosity hypothesis (NTCIR-3 Patent)



Augmenting average document length year by year



Average unique terms in a document as well



Are long patent documents simply verbose?

- Presumably verbose in view of subject topic coverage / topical relevance?
- How about in view of “Invalidation”?
- Why patent documents are getting longer every year?
- Longer patent documents are stronger because of their document characteristics.
  - They can broaden the extension of the rights covered by the claim.
  - Needs to cover and to describe augmenting complexities of technological domains.

Average document length of relevant and non-relevant documents

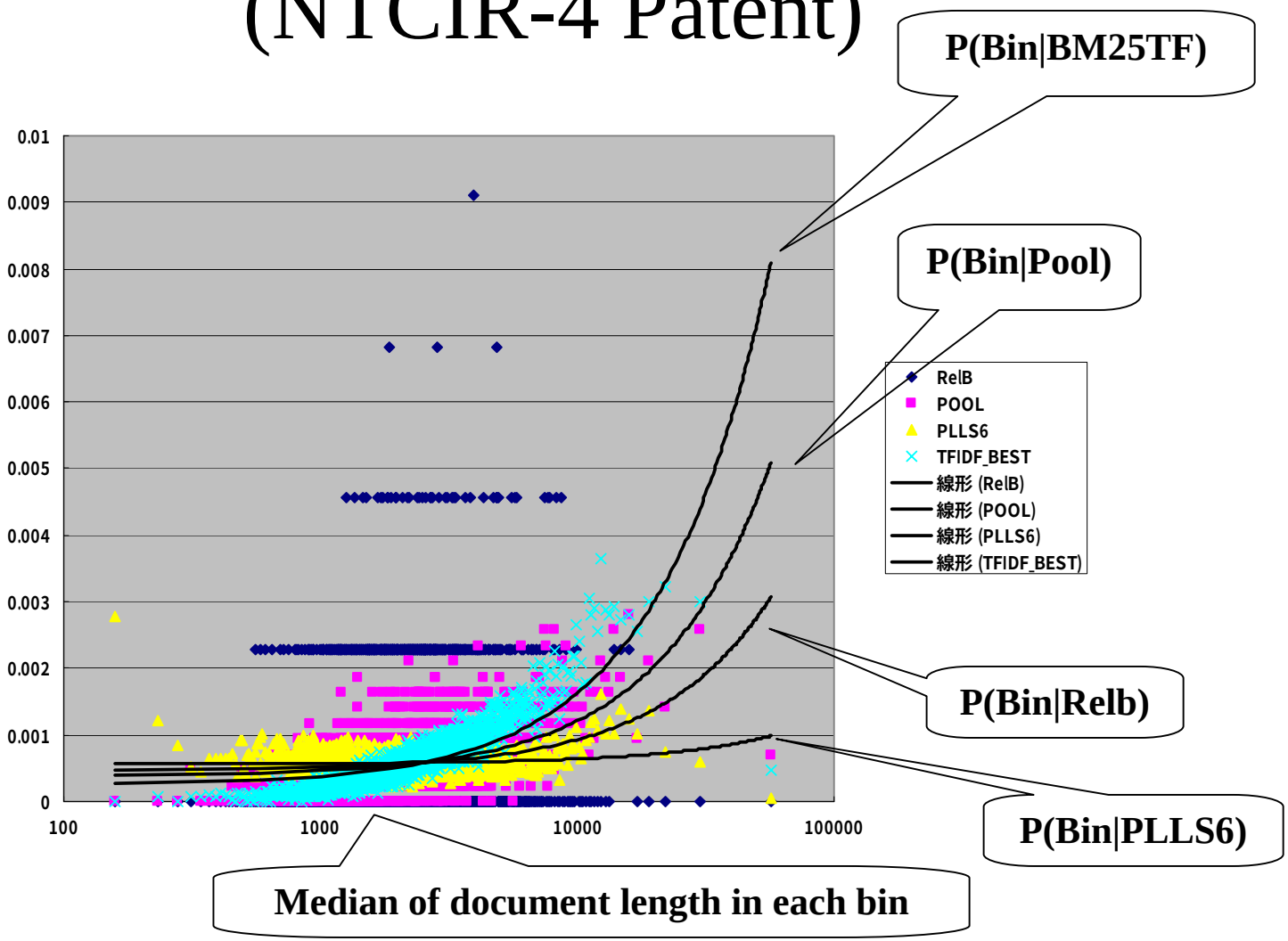
Document length clearly affects the relevance.

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

Document length merely affects the relevance.

Document length fairly affects the relevance.

Verbose but strong? (NTCIR-4 Patent)



CLIR experiments

- Title or Description Only runs: simple TF*IDF with PFB
- Title and Description runs: Fusion of Title run and Description run
- Post submission: KL-divergence runs(Dirichlet smoothing, KL-Dir) with/without document length priors

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

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

CLIR runs for J-J SLIR

	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

	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

KL-JM/KL-dir runs perform poorly.

CLIR J-J with doc length priors

- PLLS-J-J-T-03(TF*IDF):0.3801
- Dirichlet :0.3145
- Dirichlet with a doc length prior:0.2908
- Simple penalization or promotion by document length does not help.
- More work is needed for document length normalization in Language modeling IR.

Patent main task experiments

- Invalidation search by claim-document matching(claim-to-be-invalidated-as-query)
- Indexing range:
  full text vs selected fields indexing
- KL-Dir vs TF*IDF
- Distributed retrieval strategy vs centralized retrieval

Indexing range: full text vs selected fields indexing

- Full text is much better(statistically significant, p=0.05) than selected fields(Abs+Claims) indexing.
- KL-Dir, Selected fields, (PLLS3):0.1548
- KL-Dir,Fulltext,(PLLS6):0.2408

KL-Dir vs TF*IDF

- TF*IDF, Selected, (PLLS1):0.1734
- KL-Dir, Selected, (PLLS3):0.1548
- But with additional topic set:
- TF*IDF, Selected, (PLLS1):0.0499
- KL-Dir, Selected, (PLLS3):0.0557
- No big difference(not statistically significant)!

Distributed retrieval vs centralized retrieval

No statistically significant difference between KL-Dir and TF*IDF

	KL-Dir	TF*IDF
Distributed base	0.2408	0.1703
Distributed BEST	0.2488	0.2516
Centralized base	0.2274	0.1712
Centralized BEST	0.2508	0.2625

Centralized search is not necessarily must!

Patent with doc length penalization

- TF*IDF Best(Centralized): 0.2625
- Best while B=0.9-1.0
  - Doc length penalization helps!
  - NTCIR-4 CLIR J-J: 0.35 – 0.5
  - Usually 0.2-0.3 while document length is controlled
  - Theoretically 0.0 while document length is uniform
- Best while k1 is about 0.9
  - NTCIR-4 CLIR J-J: 1 – 1.2
- Better while query TF is constant

Conclusions

- According to the different document length hypotheses of the retrieval tasks, different retrieval methods are examined with various parameters.
- In news paper search, BM25 TF, which tends to retrieve longer documents outperforms KL-Dir method while no big difference in patent retrieval.
- Simple penalization or promotion by document length prior does not help i.e. cosine normalization or document length priors.