特許検索の諸相
— 「NIIテストコレクション3特許」を用いて —

岩山 真[1]　藤井 敦[2]　神門 典子[3]　丸川 雄三[4]

1 東京工業大学/(株) 日立製作所 iwayama@pi.titech.ac.jp
2 筑波大学 fujii@slis.tsukuba.ac.jp
3 国立情報学研究所 kando@nii.ac.jp
4 東京工業大学 maru@pi.titech.ac.jp

1 はじめに

国立情報学研究所が主催した評価ワークショップNTCIR-3[1] (2001年6月〜2002年12月) において，特許情報を検索対象とする「特許検索タスク」が行われ，特許検索に特化した初めてのテストコレクション「NIIテストコレクション3特許」が構築された．

従来からTREC[2]，CLEF[3]，NTCIRでも大規模な文書検索用テストコレクションが構築されており，これらは文書検索の技術発展に大きく寄与している．ところが，ほとんどのテストコレクションでは，新聞など比較的短かい文書を検索対象としており，特許に関しては，TRECが検索対象に少数の特許を含むのみである[4]．このことは，多くの特許検索システム/サービスが商用で存在しているにも拘らず，特許検索の諸相が学術分野で深く議論されてこなかった一つの原因となっている．特許には，「文書が長い」「構造を持つ」「分類がある」等の様々な特徴があり，また検索目的も多様であるため，従来の文書検索の手法が必ずしも特許検索においても有効であるとは限らない．NTCIR-3ではこのような状況を鑑み，特許検索に関する技術交流，技術促進を目的に，特許検索タスクを立ち上げた．

本論文では，NTCIR-3特許検索タスクの概要を説明すると共に，「NIIテストコレクション3特許」を用いて行った検索実験の結果を報告する．実験の目的は，検索モデルの比較，検索対象の比較 (特許以外も含む)，索引付け手法の比較等を介して，特許検索の特徴を明らかにすることである．

[1] http://research.nii.ac.jp/ntcir/index-ja.html
[2] http://trec.nist.gov/
[3] http://clef.iei.pi.cnr.it/
[4] TREC-6の総文書数約175万件に対し特許は6,711件にしか満たない．

2 NTCIR-3特許検索タスク

一概に特許検索といっても，目的により検索戦略も異なる．NTCIR-3特許検索タスクでは，「技術動向調査」という一般の文献調査に近い検索に焦点を絞った．

タスクでは，非専門家のユーザが新聞記事から関連する特許を検索する状況を想定している．ユーザは新聞記事にメモ (補足情報) を付けて，特許検索の専門家に検索を依頼する．タスク参加者は，専門家に成り代わり，新聞記事とメモから関連特許を検索することになる．これは新聞から特許を検索するジャンル横断検索となっている．

タスクの詳細については [2] を参照してほしい．

文書コレクション

タスクでは以下の文書コレクションを配布した．

	公開公報	JAPIO抄録	PAJ
タイプ	全文	抄録	抄録
言語	日本語	日本語	英語
年	1998-1999	1995-1999	1995-1999
文書数	697,262	1,706,154	1,701,339
バイト数	18,139M	1,883M	2,711M

JAPIO抄録は，公開公報に含まれる要約をJAPIO(日本特許情報機構) の専門家が修正したものである．主に長さの統制，専門用語の統制を行っている．PAJはJAPIO抄録を専門家が英語に翻訳したものである．タスクでは公開公報の2年分 (98,99) を検索対象とした．

検索課題

JIPA(日本知的財産協会) の委員12名が，31個の検索課題を作成した．検索課題はSGMLで記述され，検索の一次情報である新聞記事<ARTICLE>とメモ<SUPPLEMENT>の他にも，検索課題の概要を2,3文で記述した<DESCRIPTION>，さらに詳しく説明した<NARRATIVE>なども含んでいる．検索課題は日本語版に加え，英語，韓国語，中国語版 (簡体・繁体) も用意した．

適合性評価

課題作成者 (JIPA) が検索結果の適合性を判定した．判定基準は，A(適合)，B(部分適合)，C(不適合:本文閲覧有り)，D(不適合:本文閲覧無し) の 4 段階である．検索目的が技術動向調査であるため，請求項とその他の部分とを区別せず，特許本文のどこに技術が開示されていても適合と判定した．

3 実験

以上で説明したテストコレクションを用いて検索実験を行った．実験の目的は，特許検索という観点から，検索モデルの有効性，索引付け手法の有効性などを調べ直すことである．従来，特許に関してこのような比較調査はあまり行われていない．実験では，特許と他の文書を比較するために，毎日新聞に関するテストコレクション (NTCIR-3 言語横断検索タスクで作成) も用いた．

いくつかの知見は，公式結果から得ることも可能であるが，異なるシステム間で要素技術を比較することは困難であるため，本研究では，同一のプラットホーム上で異る検索モデルを実装して比較を行った．検索エンジンには GETA[5] を用いた．

実験で用いたテストコレクション，検索エンジン，形態素解析プログラムは全て公開されているため，本実験は容易に追試することができる．補助的なプログラムも後日公開する予定である．

索引語

まず，ChaSen(version 2.2.9) と IPA 辞書 (version 2.4.4)[6] で形態素解析を行い，名詞，動詞，形容詞，未定義語を索引語とした (単語)．

次に，文字バイグラムを索引語とした．ただし，ひらがなを含むバイグラム，文書頻度 1 のバイグラム，全文書に出現するバイグラムは索引語から除いた (文字バイグラム)．

最後に，上記の二つを統合し，まずは単語のみ，文字バイグラムのみから成る索引でそれぞれ独立に検索を行い，二つのスコアの線型結合値を最終的なスコアとした (単語と文字バイグラムの併用)．

索引語に関する比較実験は，特許についてのみ行い，毎日新聞に対してはまだ行っていない．また，複合名詞を索引語とした場合も現在実験中である．

検索対象

以下の 6 種類を用意した．

「特許全文 (Full)」	公開特許公報全文
「出願人要約 (Abs)」	公開特許公報内の出願人要約
「請求項 (Claim)」	請求項全文
「出願人要約+請求項 (Abs+Claim)」	
「JAPIO 抄録 (Jsh)」	
「毎日新聞全文 (Full)」	毎日新聞の全文

[5] http://geta.ex.nii.jp/
[6] http://chasen.aist-nara.ac.jp/

検索課題

用いた検索課題数は，特許については 31 課題，毎日新聞については 42 課題である．使用したフィールドは，特許については，<DESCRIPTION> (D), <DESCRIPTION> + <NARRATIVE> (DN), <ARTICLE> + <SUPPLEMENT> (AS) の 3 種類，新聞については，<DESC> (D), <DESC> + <NARR> (DN) の 2 種類である．

検索モデル

以下の 9 個を比較した．ここで，"log(tf).idf.dl" とは，SMART[4, 5] で用いられている検索モデルである．SMART と BM25[3] は，文書長の正規化を行っている．図の 9 個の検索モデルは GETA のモジュールとして実装されているため，検索モデルの有効性が純粋に比較できる点に注意して欲しい．

検索モデル	w_t: $RSV_{q,d} = \sum_t w_t$
hits	$b_{q,t} \times b_{d,t}$
baseline	$f_{q,t} \times b_{d,t}$
tf	$f_{q,t} \times \frac{f_{d,t}}{dlf_d}$
idf	$f_{q,t} \times idf_t$
tf.idf	$f_{q,t} \times idf_t \times \frac{f_{d,t}}{dlf_d}$
log(tf)	$(1+\log(f_{q,t})) \times \frac{1+\log(f_{d,t})}{1+\log(avef_d)}$
log(tf).idf	$(1+\log(f_{q,t})) \times idf_t \times \frac{1+\log(f_{d,t})}{1+\log(avef_d)}$
log(tf).idf.dl	$(1+\log(f_{q,t})) \times idf_t \times \frac{1+\log(f_{d,t})}{1+\log(avef_d)} \times \frac{1}{avedlb+\mathrm{S}\times(dlb_d-avedlb)}$
BM25	$f_{q,t} \times \log(\frac{N-n_t+0.5}{n_t+0.5}) \times \frac{(\mathrm{K}+1)\times f_{d,t}}{\mathrm{K}\times\{(1-\mathrm{b})+\mathrm{b}\frac{dlf_d}{avedlf}\}+f_{d,t}}$

q: 検索課題, d: 検索対象文書, t: ターム, N: 総文書数,
n_t: t を含む文書数, $b_{x,t}$: x に t が存在する (1) かしない (0) か,
$f_{x,t}$: x 内での t の頻度, idf_t: $1+\log(\frac{N}{n_t})$,
dlb_x: x 内の異なりターム数, dlf_x: x 内のターム総頻度,
$avef_x$: x 内の平均ターム頻度,
$avedlb$: コレクション内での dlb_x の平均,
$S=0.2$, $K=2.0$, $b=0.8$ (定数値は予備実験により決定)

4 実験結果

表 1 に索引語による MAP(Mean Average Precision) の比較結果を示す．MAP とは，検索課題毎の平均精度を検索課題に渡って更に平均した値である．平均精度を計算する際は，"A" のみを適合とした．表 2 と表 3 に，検索対象，検索モデル，検索課題による MAP の比較結果を示す．なお，これらは単語を索引語とした場合である．

5 考察

5.1 索引語

表 1 より，単語と文字バイグラムとの差は微小である．t-検定でも，いずれも有意差にはならない．

ところが，両者を併用すると，単語，文字バイグラムを単独で利用するよりも MAP 値が大きい．表には，併

表 1: 索引語による MAP: 索引対象は特許全文

	SMART		BM25	
	D	DN	D	DN
単語	.2272	.2660	.2224	.2418
文字バイグラム	.2131	.2600	.2311	.2599
併用 (最大値)	.2488	.2906	.2780	.2904
比 (単:バ)	10:8	10:6	10:9	7:10

用の比率を変化させたとき[7] の MAP 最大値とその時の併用比率を示してあるが，他の比率でも，安定して併用の方が上回っている．併用 (最大値) と単独使用 (各々) との差も約半数において，危険率 5%以下で有意となっている．

以上，単語，文字バイグラムには優劣付け難く，両者を併用することが効果的であることがわかった．以降では，索引語を単語に限り詳しく考察を加える．

5.2 tf(term frequency)

表 2 および表 3 を見ると，文書内単語頻度のみを考慮した検索モデル (tf) は他の検索モデルに劣っていることがわかる．hits や baseline といった単純なモデルにも劣る．更に，tf を idf と組みあわせた場合 (tf.idf)，tf は idf 単独での効果を軽減するよう負に働いてしまう点にも注目してほしい[8]．

idf と tf.idf との差を t-検定すると，特許，新聞共に，検索条件が異なる全ての場合において，危険率 5%で tf.idf < idf となる．つまり，単純な tf の使用は，特許，新聞共に検索効率に寄与していない．

一方，tf を対数化して非線型にすると (log(tf))，特に対象が全文の場合，idf と組みあわせる効果が出てくる．特許全文 (Full)，毎日新聞において，危険率 1%で log(tf).idf > idf となる．

5.3 idf(inverse document frequency)

idf の効果は大きい．抄録系の Abs や Jsh を検索対象とした場合，idf のみを使うだけで，SMART や BM25 といった最新の検索モデルをも上回る．

特に JAPIO 抄録 (Jsh) を検索した場合，どの検索フィールドから検索しても idf がトップとなる．2 位以下との差は危険率 1%で有意である．

idf が抄録系で有効な理由として，抄録は長さがそろっているという特徴を挙げることができる．つまり，SMART や BM25 において文書長正規化の働く余地が無いということである．

5.4 出願人抄録 対 JAPIO 抄録

表 2 を見ると，全ての場合において JAPIO 抄録を検索する方が勝っている．t-検定の結果も，ほとんど全てにおいて，危険率 1%で有意差がある．出願人要約を専門家が修正したことの効果があらわれていると言えよう．

5.5 SMART 対 BM25

TREC 等でも良く使われ好成績を上げている，SMART(ベクトル空間モデル) と BM25(確率モデル) とを比較する．表 2, 3 を見ると差は微少であり，ほとんどの場合において統計的な有意差もない．

唯一，新聞記事 (AS) から特許全文 (Full) を検索するジャンル横断検索において，SMART が BM25 を大きく上回る．現在この原因を調査中である．

5.6 ジャンル横断検索

ジャンル横断検索 (AS) と，<DESCRIPTION>や<NARRATIVE>から検索を行う場合 (D,DN) とを比較すると，ジャンル横断検索が押しなべて劣っている．

新聞と特許における単語分布は大きく異なるため，ジャンル横断検索ではジャンル間での単語分布の隔たりを考慮した検索モデルが必要となるのかもしれない．公式結果でも，このような検索モデルを提案したチーム [1] が好成績を上げている．ジャンル横断検索は，従来あまり注目されておらず，今後の研究成果が期待できる分野である．

5.7 全文 対 抄録 対 請求項

全文 (Full) と抄録 (Abs,Jsh) とを比べると，特に SMART, BM25 において，全文を検索するほうが優れていることがわかる．前述したように文書長正規化の効果が原因の一つである．また，請求項のみを検索対象とした場合 (Claim)，全てにおいて最低値となっている．これは，技術動向調査という目的で，請求項から情報を汲みとることが難しいことを示唆している．

5.8 特許 対 新聞

検索モデルと MAP という観点で比べると，予想に反して，特許と新聞間に傾向の違いは無かった．一般に，「特許では専門用語，新語が多い」「請求項では発明の範囲を広げるために一般語が頻出する」等言われるが，今

[7] 併用の比率を一度決めた後は，全課題にこの比率を用いている．
[8] この問題は ACM SIGIR 2000 で開催された特許検索に関するワークショップでも報告されている．

表 2: 検索モデル/検索課題/検索対象による MAP(特許)

	Full			Abs			Claim			Abs+Claim			Jsh		
検索モデル	D	DN	AS	D	DN	AS	D	DN	AS	D	DN	AS	D	DN	AS
hits	.1050	.0534	.0166	.0727	.0429	.0120	.0516	.0128	.0025	.0840	.0242	.0045	.1171	.0547	.0373
baseline	.0931	.0725	.0292	.0732	.0813	.0304	.0566	.0572	.0168	.0854	.0741	.0274	.1066	.1138	.0538
tf	.0156	.0227	.0046	.0132	.0158	.0036	.0136	.0172	.0047	.0151	.0183	.0042	.0113	.0166	.0025
idf	.1515	.1577	.0744	**.1197**	**.1272**	.0755	.0941	.0935	.0367	.1278	.1265	.0665	**.1730**	**.1682**	**.1271**
tf.idf	.0277	.0390	.0231	.0239	.0284	.0197	.0279	.0337	.0155	.0298	.0353	.0227	.0222	.0258	.0166
log(tf)	.1642	.1255	.0337	.0579	.0723	.0237	.0669	.0527	.0099	.0917	.0787	.0186	.0821	.0899	.0457
log(tf).idf	.2230	.2132	.1082	.0884	.1151	.0781	.0978	.1029	.0380	.1237	.1306	.0725	.1226	.1465	.1223
log(tf).idf+dl	.2272	**.2660**	**.1790**	.0887	.1169	**.0844**	.1028	**.1182**	**.0752**	.1215	.1419	**.1062**	.1184	.1501	**.1271**
BM25	**.2280**	.2503	.0875	.0838	.0997	.0707	**.1039**	.1129	.0557	**.1302**	**.1426**	.0786	.1356	.1474	.1015

D: <DESCRIPTION>, N: <NARRATIVE>, A: <ARTICLE>, S: <SUPPLEMENT>

表 3: 検索モデル/検索課題による MAP(毎日新聞)

検索モデル	D	DN
hits	.1397	.1063
baseline	.1436	.1865
tf	.0755	.1054
idf	.1914	.2443
tf.idf	.1041	.1279
log(tf)	.2266	.2124
log(tf).idf	.2940	.2853
log(tf).idf+dl	.2746	.3212
BM25	**.2759**	**.3346**

D: <DESC>, N: <NARR>

回の技術動向調査に関する限り，従来の検索モデルが特許にも十分適用可能であることがわかった．

これは今回の技術動向調査が，適合性判定において請求項を特別扱いしていないことが原因かもしれない．請求項は発明の範囲を定める重要な項目であるにもかかわらず，前述したように，請求項のみを検索したのでは良い結果が得られていない．このことから，請求項に対しても従来の検索モデルが有効であるかどうかについては，今回の結果からはまだ結論付けることはできない．次回の NTCIR-4 特許検索タスクでは，無効資料調査を対象にして，請求項を焦点に特許特有の現象を追及していく予定である．

6 おわりに

本論文では，NTCIR-3 で行われた「特許検索タスク」の概要を説明すると共に，そこで作成された「NII テストコレクション 3 特許」を用いて行った検索実験の結果を報告した．

最後に，タスクに参加してテストコレクションの構築に協力していただいた方々，および日本知的財産協会の方々に感謝したい．

参考文献

[1] H. Itoh, H. Mano, and Y. Ogawa. Term distillation for cross-DB retrieval. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[2] M. Iwayama, A. Fujii, N. Kando, and A. Takano. Overview of patent retrieval task at ntcir-3. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, 2003.

[3] S. Robertson and S. Walker. Some simple effective approximation to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[4] G. Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall, 1971.

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