社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (6)
2008－ＮＬ－183 (6)
2008／1／21

観点に着目した特許分類手法の高精度化の提案

水本浩司　　湯川高志
長岡技術科学大学

特許文献の分類を高速かつ高精度で行う自動分類技術の確立を目的として，χ^2 統計量を用いた単語重み付けによる手法として本研究室が提案した高速な分類手法と，特許分類の階層構造を利用した観点強調とを組み合わせた新たな分類手法を提案し,実装した．観点は階層構造の上位にあたり,分類が比較的容易である点に着目し，大量の分類基準をもつ特許文献を分類する際にそれを用いることで，精度の向上を図った．提案した分類手法に対して，NTCIR-6 のテストセットを使用して分類精度を評価し，既存手法の分類精度と比較した.その結果，8 割以上の課題テーマにおいて,提案手法による分類精度の向上が確認できた．

Improvement of Accuracy for an Automated Patent Classification Method using Viewpoints in the Category Hierarchy

Koji Mizumoto　Takashi Yukawa
Nagaoka University of Technology

This paper proposes a new method for patent document classification with combination of a term weighting technique using the chi-square statistics and viewpoint emphasis based on hierarchical structure of F-term categories. The viewpoints correspond to second-level of the F-term hierarchy and they are able to be accurately extracted compared with leaf categories of the hierarchy. Therefore, the use of them is expected to improve accuracy of the classification. The classification accuracy was evaluated using the test data set of the classification subtask at NTCIR6 patent retrieval task and compared with some conventional methods. The results show that the purpose method improved classification for more than 80 % of themes in the test data set.

1.はじめに

現在の文書分類技術においては，昨今の PC の性能の向上を受け，ベクトル空間モデル法を用いた分類手法が良く使用されている．NTCIR-6 Patent Classification Subtask で行われた，特許文書の F ターム分類においても，k-近傍決定則(k-NN)や，サポートベクターマシーン(SVM)等の，ベクトル空間法や複雑な学習を用いた分類手法が多く使われており，高い分類精度を得ている[1]．しかし,これらの手法は分類結果を得るまでに演算処理に時間がかかるという問題を持っている．特に，特許文書を扱った分類等，大量の文書データを扱う必要がある場合は，処理に掛かる時間は無視できるものではない．そこで，特許文書のような大規模のデータが対象でも処理時間が膨れ上がらない高速な手法が望まれる．ベクトル空間法等を用いずに文書の分類を行う方法

として TF-IDF 等を用いて単語毎の重みを付け，スコアを計算し分類する手法がある．この手法では，単語同士が独立としているという仮定を用いる為，文書中の単語間の関係を使用することは出来ないが，複雑な計算が必要ではなく，高速な処理が期待できる．しかし，その結果として，分類精度ではベクトル空間法を用いた手法に対して,一歩及ばない結果となっている．そこで,本稿では単語重み付けを用いた分類手法を元として，処理速度を大きく損なわずにベクトル空間手法に並ぶ分類精度の高さが得られる手法ができないかと考え，F タームの観点強調を用いた分類手法を提案し,評価を行った．

2.単語重み付けによる分類

2.1. 単語重み付け手法

単語重み付け手法の最もポピュラーな手法として，TF-IDF 法がある．TF-IDF 法では，文書中における索引語の出現頻度である TF (Term Frequency) と，全文書中で，索引語が存在する文書の出現頻度である DF (Document Frequency)の指標を用いる．TF 値は文書における索引語の重要性 DF 値は索引語の特定性をそれぞれ表す．TF 値と DF 値の逆数である IDF(Inverse Document Frequency)との積を重みとする．TF-IDF 法は単純な枠組みでありながら高い精度が得られ，検索等の用途で一般的に利用される手法である．しかし，TF-IDF 法 をはじめとした，多くの重み付け手法ではある単語がある文書，あるカテゴリーに出現する，重要であるという情報のみを計算式に用いる．しかし，ある単語がある文書,あるカテゴリーに出現しない，重要でないといった情報も利用できればより良い単語の重み付けを行えるのではないかと考えられる．そして，このような関係を考慮して重み付けを行うものとして χ^2 統計量単語重み付け手法がある．

2.2. χ^2 統計量単語重み付け手法

χ^2 統計量を用いた手法は，橋元らによって提案された特許分類手法である[2]．トレーニングデータの文書を元に文書中の単語と F タームの χ^2 統計量を以下のような 2×2 の分割表を用いて求める．

表 1. F ターム対単語 2×2 分割表

	単語 1	¬単語 1	小計
F タームA	y	x-y	x
¬F タームA	m-y	n-x-m+y	n-x
小計	m	n-m	n

「n」はあるテーマの全文書数,「 m」は単語 1 を含む文書数,「x」は F ターム A が付与されている文書数,「 y」は単語 1 を含み, F ターム A が付与されている文書数である.これらの値を元として，表 1 のような 2×2 分割表を作成する． 2×2 分割表のセルには「y」, F ターム A が付与されているが単語 1 が含まれない文書数「x-y」，単語 1 を含むが F ターム A が付与されていない文書数「m-y」, F ターム A が付与されず，単語 1 も含まない文書数「n-x-m+y」が，それぞれ入る．これらの値から，以下の式によって χ^2 統計量が求められる

$$\chi^2(x,y) = D(y,\alpha\gamma n) + D(x-y,\alpha(1-\gamma)n) + D(m-y,(1-\alpha)\gamma n) + D(n-x-(m-y),(1-\alpha)(1-\gamma)n)$$

$$\alpha = x/n,\ \gamma = m/n,\ D(o,e) = (o-e)^2/e$$

χ^2 統計量は「x-y」の値と「m-y」の値が小さいほど大きくなる．つまり, F ターム A に分類される文書中でのみ単語 1 が存在し，単語 1 が存在する文書でのみ F ターム A が付加されるという 1 対 1 の関係を持った時に最もスコアが高くなる．

特許文書の分類においては対象となる全ての F タームそれぞれに対して，各単語の χ^2 統

計量を求め，単語の重みとしている．この単語の重みを参照して未分類特許文書の分類を行う．ここで，それぞれの単語の重みはχ^2統計値を求める式によって一意的に決定する為，学習などの複雑で時間のかかる演算を行う必要がない．

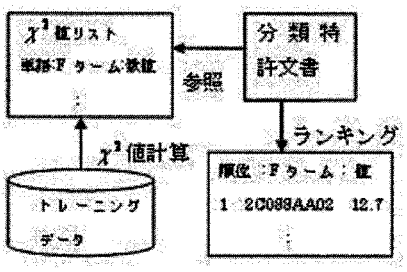

図 1. χ^2 統計量単語重み付け手法による特許文書分類

また，2 語から成り立つ複合語や，近くにある単語の関係の強さを考慮し，近傍共起する単語2つを1つにまとめ近傍共起単語とし，同様にFタームとのχ^2統計値リストを作成している．

本稿では高速ながら，良好な分類精度が得られるχ^2統計量単語重み付け手法をベースに使用し，高速性を保ちながらも，より高精度な分類を行える手法を提案する．

3. 観点強調による重み付け手法

3.1. Fタームの観点

NTCIR の特許分類サブタスクで対象となった F タームは非常に細かく分けられている．その数は 1 テーマに対して 100 を越える分類を持つものが殆どであり，多いものになると 500 を越える F タームが存在する．そこで，F タームがもつ「観点」に注目する．観点はFタームに比べると，分類の対象となる数が少なく，意味も F タームほど細かく分けられているわけではない．例えばNTCIR6の 特許分類サブタスクで課題となったテーマ「2C088」(弾球遊技機(パチンコ等))の F ターム数は 390 あり,対して観点の数は 9 と非常に少ない．この為，観点の分類はFタームと比べれば，容易であると考えられる．

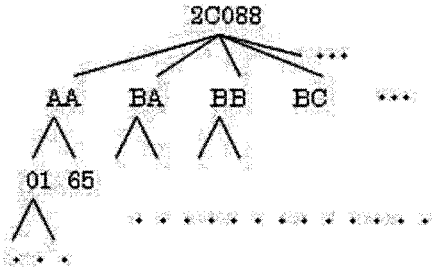

図 2. F タームの構造(テーマ 2C088)

F タームは図 2 で示すように階層構造を持っている．AA(AA00 ：ゲーム内容)や BA(BA00：パチンコ球処理)といった観点が階層の上位にありその下に細かく分けられた F ターム(BA01：球の検出，BA30:球の計量排出など)が存在している．例えば，図 2 と同様のテーマ 2C088 のとある特許文書において，観点がAAとBBだと解れば，他の観点以下にある F タームを分類の対象とする必要はなく，AAとBBの2つに属するFタームから選択するだけでよいことになる．その結果 F タームの分類において正解の順位を下げる要因となる雑音を減らすことが出来，代わりに正解となるFタームの順位を上げることが出来る．

3.2.観点強調を用いた分類手法

文書毎に観点を抽出し,その観点を持つ F タームのスコアに重み付けすることで，他観点をもつFタームと差別化する．

(1)F ターム対単語のχ^2統計値表で共通の観点をもつFターム同士まとめ，再計算し，観点対単語表を作成する．

スコアsは以下のような式で求める．

$$s = \sqrt{\frac{s_t}{r}} \cdot (0.5 + 0.5 f_v)$$

式において,それぞれの値の意味は，s_tは同観点を持つFターム対単語χ^2統計値の単純合計，rはFターム対単語χ^2統計値を参照した数，f_vはトレーニングデータの観点の頻度である．ルートで囲まれた式の左項はFターム数が多い観点のスコアを抑える為の補正，右項は観点の登場頻度を考慮する為の補正である．f_vの値は差が顕著である為，値に0.5の底上げを行い,極端な補正とならないようにしている．

(2)同様に，Fターム対近傍共起単語のχ^2統計値表から共通の観点を持つFターム同士をまとめ再計算し，観点対近傍共起単語表を作成する．

(3)観点対単語表及び観点対近傍共起単語表を参照し，特許文書毎の観点分類を行う．また，このときトレーニングデータより作成した，テーマ毎のストップワードを使用する．

(4)観点分類結果をスコアより降順にソートし，N位までの観点をFターム分類における強調の対象とする．

(5)χ^2統計値によるFターム分類で，観点強調の対象となるFターム全てにα倍のスコアの重みづけをし，計算する．

(6)計算結果についてスコアの高い200位までをFターム分類結果とする．

3.3.提案手法のパラメータ

提案手法においては，観点の分類精度の他に，強調する観点の数Nと重み付け時の係数αがFタームの分類精度に影響を強く与えると考えられる．以下,この2つのパラメータについて説明する．

・強調観点数N

観点の数はFタームと同じく特許毎に違う．しかし，個々の特許に対して正確な観点数を選択することは困難な為，テーマ毎に観点数を決定する．テーマのトレーニング文書を参照し，最も頻度が高い観点の数を強調観点数Nとする．

・観点強調係数α

スコアの重み付けで，正しい観点のみで構成される観点リストで強調を行えば，強調の倍率αが高い程，精度向上が望める．

表2.観点強調係数αとMAP値

強調(α倍)	MAP
1	0.4101
1.5	0.4523
2	0.4696
2.5	0.4791
3	0.4848
4	0.4907
5	0.4937

表2.は観点リストが正確に抽出されたと仮定した場合にα=1.0から次第に増加させたときのMAP値の変化である．αを増やすとMAP値は上昇していき，20%程度増加したところで飽和する．強調を極端化(α=∞)すれば，強調観点リストにない観点のFタームを全て取り除いた事と同じ意味になる．しかし，実際には，，観点の抽出が正確に行えない為，値を大きくしすぎると，正解のFタームがある観点を取り除いてしまう可能性もある．よって，提案手法ではαの値を過剰に上昇させず，スコア上昇率の高いα=2までを対象とする．

4. 評価

4.1. 評価方法

NTCIR-6 Patent Classification Subtaskと同じ条件で提案手法を評価する．

・トレーニングデータ

1993-1997年の特許文書

・テストデータ

1998-1999年の特許文書

・分類対象テーマ数：108

・分類対称特許文書数：21606

特許文書の分類を行い，スコアの高い順に200位までのFタームを結果とする．正解データと比較して，AP(Average Precision)を計算する．全ての分類対象特許文書についてAP値を求め，その平均であるMAP (Mean Average Precision)値を最終的な評価値とする．

4.2. 結果

提案手法の性能比較の為，提案手法，ベース手法，そして，ベクトル空間モデルのSVMを用いた手法のそれぞれのMAP値を示す．

表3.手法のMAP値

手法	MAP
χ^2 統計量単語重み付け	0.4101
提案手法	0.4223
SVM+NB (JSPAT1[3])	0.4381

提案手法のMAP値はχ^2統計量単語重み付け手法に対して、約3%の向上が得られている．しかし，ベクトル空間モデルを用いた分類手法には及ばない結果となった．

また，提案手法とベース手法をテーマ毎の比較を行った結果，2割程度のテーマについては観点の強調をすることによって逆にAPを下げているという問題が見られた．

5. 提案手法の改良

提案した分類手法により，ベースとなったχ^2統計量単語重み付けからある程度のMAP値の向上が得られた．しかし，その精度向上は，表2で示した理論値には及ばず、幾つかのテーマでは観点の強調で逆に精度を下げてしまっている．この問題を考慮して，提案手法の改良案を考える．

5.1.提案手法に対する考察

観点強調で，抽出結果に大きく影響を与える要素は強調観点数Nと観点強調係数αであると考えられる．強調観点数Nはテーマ毎に決定しているが，実際の観点は特許文書毎に数が違う．Nの値を増やせば正解となる観点が増え，観点抽出のRecallが増加するが雑音が増えPrecisionが低下する．逆に減らせば正解となる観点が減り，観点抽出のRecallが低下するが雑音が減りPrecisionが向上する．Recallが低下した場合,正解となるFタームがもつ観点全てを抽出することができなくなる．この為，抽出されなかった正解の観点をもつFタームの順位が相対的に下がり，分類精度が低下する．Precisionを低下させた場合，正解以外の観点も強調することになる．これにより正解観点を持つFタームと，そうでないFタームとの差別化ができなくなり，観点強調の効果があまり得られなくなる．このように，強調観点の抽出ではRecall, Precisionのどちらか片方の値だけが極端に高くてもFターム分類で良い効果を得ることは出来ない．観点強調を効果的に生かすには，的確な強調観点数Nを選ばなければならない．

観点強調係数αは全テーマに対して同じ値にしている．しかし，観点の分類精度が低いテーマでαが高いとFタームの分類結果が逆に悪化し，観点の分類精度が高いテーマでαが低いと思うような精度の向上が得られない．この為，テーマ毎の傾向，分類精度にあった強調観点係数αを決定することが，良い分類結果を得るために必要であると考えられる．

5.2.パラメータの最適化

分類精度改善の為，観点強調係数αと強調観点数Nの2つのパラメータをテーマ毎に最適化を行う．トレーニングデータからテーマ毎に100通ずつ特許文書を抜き出し，それに対してパラメータを変化させながら分類，評価を行い最適な値を決定する．観点の抽出に

おける, Recall, Precision, F 値の高さは, F タ ーム分類精度の高さには直結しない. その為トレーニングデータでの, F タ ーム分類結果の MAP 値を評価の指標とした. パラメータの最適化は以下のような手順で行っている.

(1)観点強調係数 α を 1.0 から 0.25 刻みで変化させ, MAP 値が上昇をしなくなるまで α を増加させ, 最大の MAP 値をとった α を,そのテーマ個別の観点強調係数とする.

(2)手順(1)で α=1.0 が最大の MAP 値となったテーマに対して, 強調観点数 N を 1 減少させ α=1.25 での MAP 値を計算する.

(3) α=1.0 と手順(2)で計算した MAP 値を比較する. ここで, 値の上昇が見られたときは, その N を新しい強調観点数とする.

(4)上昇しない場合は減少させる前の N での α=1.25 の MAP 値と比較する. 比較した結果, 新しい N の値が大きければ N を更に減少させ(3)からの手順を繰り返す. 新しい方の値が小さい時は次の手順に移る.

(5)強調観点数 N を逆に 1 ずつ増やしていき(2)～(4)の手順を繰り返す.

(6)(2)～(5)の手順で強調観点数が更新されたテーマはその強調観点数で(1)と同じ手順によって最適な α を決定する.

5.3.改良手法の評価

改良手法で得た新たな分類結果より同様の評価を行った. MAP 値は 0.4263 となり更に 1%の精度の向上が出来た. また, 精度が向上したテーマ数が増え, 精度が下がってしまうテーマを 1 割以下に減らすことが出来, 全体的な精度の向上を行なえた. しかし, 最適化を行うことにより処理時間が大幅に増加してしまう問題点がこの手法にはある.

6.まとめ

本稿では観点抽出による特許分類手法の提案及び評価を行った. 提案手法では MAP 値で約 3%の向上が実現できた. しかし, 幾つかのテーマで分類精度が低下していること, 観点が正確に抽出されたと仮定した時に取る値と比べて分類精度が大きく劣ることから十分な効果が得られていないと考えられる. そこで, 観点抽出に強く影響を与えるパラメータに対して, 最適化を行った. 最適化で得たパラメータによる分類では, MAP 値はベース手法から 4%程向上し, 精度が向上したテーマ数も増え, 全体的な精度の改善を行うことが出来た. 結果的にベクトル空間手法の分類精度を上回ることが出来なかったものの, 高速性から得られる単語重み付け手法においてかなり良好な精度を達成したといえる.

謝辞

本手法の評価を行うにあたって, NTCIR-6 Patent Classification Task のデータセットを使用させていただきました. 開発者・作成諸氏に深く感謝いたします.

参考文献

[1]M. Iwayama, A. Fujii, N. Kando: Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task : Proceedings of the 6th NTCIR Workshop Meeting, pp366-372 (2007)

[2]K. Hashimoto and T. Yukawa: Term Weighting Classification System Using the Chi-square Statistic for the Classification Subtask at NTCIR-6 Patent Retrieval Task: Proceedings of the 6th NTCIR Workshop Meeting, pp385-389 (2007)

[3] M. Rikitoku: F-term classification Experiments at NTCIR-6 for Just System: Proceedings of the 6th NTCIR Workshop Meeting, pp420-427 (2007)