DEWS2004 2-B-5

# 情報検索における分野情報を用いた語の重み付け

北内　啓†　小西　一也†　高木　徹†

† (株)NTT データ 技術開発本部　〒 104–0033 東京都中央区新川 1–21–2
E-mail: †{kitauchia,konishikzy,takakit}@nttdata.co.jp

**あらまし**　本稿では，情報検索において文書データに付与された分野情報（カテゴリ）を用いた語の重み付け手法を提案する．文書中の語と各カテゴリとの間の関係を考慮し，カテゴリごとに語の重み付けをすることによって，特定のカテゴリと関連の深い語に高い重みが付与されるようにする．また，特定のカテゴリと強い関連をもたない語に対してはすべてのカテゴリに共通な重み付けを行うことで，語の分布がカテゴリによって不均等であることの影響を解消する．特許データを対象とした評価実験を行った結果，提案手法が従来手法よりも高い性能を示すことを確認し，特に，重要な検索語を高い割合で含む短い検索要求において提案手法の有効性が高いことが分かった．

**キーワード**　情報検索, 特許検索, 分野情報, 語の重み付け

# Term Weighting Using Category Information for Information Retrieval

Akira KITAUCHI†, Kazuya KONISHI†, and Toru TAKAKI†

† Research and Development Headquarters, NTT DATA Corporation　Shinkawa 1–21–2, Chuo-ku, Tokyo, 104–0033 Japan
E-mail: †{kitauchia,konishikzy,takakit}@nttdata.co.jp

**Abstract**　In this paper, we propose a new term weighting scheme for information retrieval using the predefined category information labeled to the documents. Our scheme computes the term weight value for each category by considering the relationship between the term in the documents and the category, with the result that a high value is assigned when the term has high relevance to the specific categories. On the other hand, we set the same values for all categories to the terms which don't have high relevance to the specific categories, which is able to reduce the influence of the inequality of the term distribution for each category. The experimental results on patent data show the proposed term weighting scheme performs better than previous schemes, particularly for short queries involving many important terms.

**Key words**　information retrieval, patent retrieval, category information, term weigthing

## 1. はじめに

大量の文書から利用者が必要とする文書を探し出すための情報検索においては，ウェブ文書，電子メール，新聞記事など様々な種類の文書が検索対象となる．多くの場合，それらの文書には何らかの属性情報が付与されており，特に分野（カテゴリ）を示す情報が付与されている文書は多い．たとえば，Yahoo!のようなディレクトリ型検索エンジンではウェブ上のページが階層的なカテゴリに分類され，新聞記事には政治，経済，スポーツなどの分野情報が付与されている．

伝統的な情報検索の方法のひとつに語の統計情報を用いる手法があり，ベクトル空間モデル [5] や確率モデル [6] などが提案されている．この手法においては，文書を語の集合ととらえ，語が出現する文書数（DF）や文書中に出現する語の数（TF）など，語と文書の共起に基づく統計情報を用いて語の重みや文書スコアを算出し，文書のランク付けを行う．

これに対し，語と文書の関係だけでなく，文書に付与された分野情報も利用することによって情報検索の性能を向上させる手法が提案されている．Murata ら [2] は，確率モデルのひとつである 2-poisson モデルの式に分野情報を利用する項を追加している．具体的には，分野情報を利用しない一度目の検索結果の上位 100 件においてよく出現し，全記事ではそれほど出現しないカテゴリを検索要求との関連度が高いカテゴリとみなし，そのカテゴリの文書のスコアに一定の値を乗じることで，分野情報を利用した文書のスコア付けを行っている．この手法では，一つのカテゴリ内の文書間のランクは変化せず，語とカテゴリの関係を文書のスコア付けに利用していないため，カテゴリとの関連が強い重要な語を含む文書が必ずしも上位にランク付け

されないという問題がある．

また，Zhao ら [3] は，分野情報を利用して特定のカテゴリと強い関連をもつ語に高い重みを付与する手法を提案している．しかし，この手法は語とカテゴリの関連の強さにかかわらず語に対して一つの重みのみを付与しており，カテゴリごとに語の重みを付与していない．したがって，特定のカテゴリと強い関連をもつ語の場合，語との関連が弱いカテゴリにおいても高い重みが付与されてしまうという問題がある．

本稿では，分野情報が付与された文書データに対する情報検索において，各カテゴリと文書中の語の間の関係を考慮し，カテゴリごとに語の重み付けを行う手法を提案する．特定のカテゴリに関連の深い語に高い重みが付与されるようにするため，まず文書中の任意の語が特定のカテゴリとどの程度強く関連しているかという関連度を表す尺度を導入する．この関連度が高い語を専門語とみなし，その語との関連が強いカテゴリにおいて高い重みが付与されるよう，カテゴリごとに異なる語の重み付けを行う．一方，この関連度が低い語を一般語とみなし，すべてのカテゴリに共通な重み付けを行うことにより，カテゴリによって語の分布が不均等となっていることの影響を解消する．

公開特許データに付与されている国際特許分類（IPC, International Patent Classification）とよばれる分野情報を利用して，出願された特許に対して，その権利化拒絶の理由となる先願特許を 5 年分の公開特許データから検索する評価実験を行い，本手法の有効性について検証する．

## 2. 分野情報を用いた語の重み付け

本研究における，分野情報を利用した語の重み付けの基本的な考え方は，語と文書の関係を用いた重み付けである $tf \cdot idf$ と，語とカテゴリの関係を用いた重み付けを組み合わせるというものである．語とカテゴリの関係を用いた重み付けは，$tf \cdot idf$ を自然に拡張する形で定義する．ただし，このとき特定のカテゴリと強く関連する語に対してはカテゴリごとに異なる語の重み付けを行い，どのカテゴリとも強い関連をもたない語に対しては全カテゴリにおいて共通に語の重み付けを行う．なお本稿では，文書には必ず一つ以上のカテゴリが付与され，複数のカテゴリが付与されている場合があることも想定する．

本章では，まず $tf \cdot idf$ の拡張による，語とカテゴリの関係を用いた重み付けについて説明する．次に，カテゴリごとに異なる語の重み付けを行うかどうかを判別する手法を提案し，最後に語とカテゴリの関係を用いた重み付けに $tf \cdot idf$ を組み合わせた重み付け手法について述べる．

### 2.1 $tf \cdot idf$ の拡張

ベクトル空間モデルにおいて語の重みを表す伝統的な尺度である $tf \cdot idf$ は，語と文書の関係から語に重み付けを行うものであり，文書に付与されている分野情報を利用していない．そのため，一つのカテゴリ内における語の特徴や，ある特定のカテゴリにおいて重要な語の特徴といった，語とカテゴリの関係をうまく表すことができない．そこで本研究では，$tf \cdot idf$ の統計的な考えを語とカテゴリの関係に拡張することで，文書に付与された分野情報を利用した語の重み付けを行った．

$tf \cdot idf$ の考えに基づいて語の重みを表す式は数多く提案されている [4]．われわれは，文書 $d$ 中の語 $t$ の重み $tfidf(d,t)$ を表す式として次式を用いた．

$$tfidf(d,t) = tf(d,t) \cdot idf(t) \tag{1}$$

ここで，

$$tf(d,t) = \log(\frac{f_d^t}{f_d} + 1)$$

$$idf(t) = \log \frac{N}{N_t}$$

ただし，$f_d^t$ は文書 $d$ における語 $t$ の出現頻度，$f_d$ は文書 $d$ に含まれるすべての語の出現頻度の総和，すなわち $f_d = \sum_{t \in d} f_d^t$ である．$N$ は全文書数，$N_t$ は語 $t$ が出現する文書数である．

まず，$tf(d,t)$ を語とカテゴリの関係に拡張する．$tf(d,t)$ は，ある文書において多く出現する語ほど特徴的であることを表す尺度となっている．一方，語とカテゴリの関係においては，あるカテゴリにおいて多くの文書に出現する語ほど特徴的であるといえる．そこで，$tf(d,t)$ を語とカテゴリの関係に自然に拡張した式として次式を用いる．

$$cdf(c,t) = \log(\frac{N_c^t}{N_c} + 1) \tag{2}$$

ただし，$N_c^t$ はカテゴリ $c$ において語 $t$ を含む文書数，$N_c$ はカテゴリ $c$ に含まれる文書数である．

また，文書 $d$ 中の語 $t$ の重みを付与するにあたっては，文書 $d$ に付与されているカテゴリの集合 $C$ に対して，次式のように $cdf(C,t)$ を算出することとする．

$$cdf(C,t) = \log(\sum_{c \in C} \frac{N_c^t}{N_c} / \#C + 1) \tag{3}$$

ここで，$\#C$ はカテゴリ集合 $C$ の要素数，すなわち文書 $d$ に付与されているカテゴリの数である．

次に，$idf(t)$ を語とカテゴリの関係に拡張する．$idf(t)$ は，全文書数に対する語 $t$ の出現文書数の割合が小さい，すなわち少数の文書にのみ出現する語が重要であることを示している．そこで，語とカテゴリの関係において，全カテゴリ数に対する語 $t$ の出現カテゴリ数の割合が小さい，すなわち少数のカテゴリにのみ出現する語が重要であることを示す ICF [1] を，$idf(t)$ を拡張した式として定義する．

$$icf(t) = \log \frac{NC}{NC_t} \tag{4}$$

ただし，$NC$ は全カテゴリ数，$NC_t$ は語 $t$ を含むカテゴリ数である．

式（3），(4）より，$tf \cdot idf$ を語とカテゴリの関係に拡張した，語の重みを表す尺度は次式のようになる．

$$cdficf(C,t) = cdf(C,t)icf(t) \tag{5}$$

表 1　$N_t/NC_t$ の値が大きい語の例

| 順位 | 語 | 値 | $N_t$ | $NC_t$ |
|---|---|---|---|---|
| 1 | 発明 | 1384.5 | 1705739 | 1232 |
| 1001 | デバイス | 72.2 | 58998 | 817 |
| 2001 | 送風 | 36.9 | 36874 | 998 |
| 3000 | オーリックトリクロライド | 25.0 | 300 | 12 |
| 4001 | 疑似 | 19.4 | 16432 | 846 |
| 4836 | 島後 | 16.5 | 33 | 2 |
| 6001 | スクリーニング | 14.6 | 6683 | 459 |
| 6980 | ローラピボット | 13.0 | 52 | 4 |
| 8001 | クワガタソウ | 11.7 | 117 | 10 |
| 8995 | フェニルテルロ | 10.9 | 87 | 8 |
| 10001 | 結束 | 10.0 | 7874 | 785 |

表 2　$\log(N_t+1)/\log(NC_t+1)$ の値が大きい語の例

| 順位 | 語 | 値 | $N_t$ | $NC_t$ |
|---|---|---|---|---|
| 1 | イジェクトセンサー | 5.29 | 38 | 1 |
| 981 | ギヤシフトスケジューリングマップ | 3.17 | 8 | 1 |
| 1997 | ブチルアルミニウムプロポキシド | 2.93 | 24 | 2 |
| 2268 | マーポスゲージ | 2.81 | 6 | 1 |
| 3356 | チタニウムジハイドライド | 2.59 | 5 | 1 |
| 4985 | ブチルヘキシルテルロエステル | 2.48 | 30 | 3 |
| 5825 | スケルチツマミ | 2.32 | 4 | 1 |
| 5825 | グリシルシステイン | 2.32 | 4 | 1 |
| 5825 | シートアジャスタユニット | 2.32 | 4 | 1 |
| 8973 | ラダーホーン | 2.29 | 23 | 3 |
| 9656 | ラベルハンガマガジン | 2.18 | 10 | 2 |

### 2.2　カテゴリごとに異なる語の重み付けの判別

式（2）の $cdf(c,t)$ と式（4）の $icf(t)$ はともに，語とカテゴリの関係を利用して語の特徴度を表す尺度である．しかし，$cdf(c,t)$ は語 $t$ に対しカテゴリごとに異なる重みが付与されるのに対し，式 (4) の $icf(t)$ は語 $t$ に対しすべてのカテゴリに共通する重みが付与されるという違いがある．$cdf(c,t)$ は，あるカテゴリ $c$ において語 $t$ がどの程度特徴的であるかを表す尺度であり，語が特定のカテゴリと強く関連するときに特に有効な尺度であるといえる．逆に，特定のカテゴリに関連しない語の場合はカテゴリごとに $cdf(c,t)$ を付与するのではなく，すべてのカテゴリにおいて同じ値を付与することにより，カテゴリごとの語の出現文書頻度の揺れを吸収することができる．そこで，特定のカテゴリとの関連度を表す尺度を考え，関連度が高い語に対しては $cdf(c,t)$ を用いてカテゴリごとに語を重み付けし，関連度が低い語に対してはすべてのカテゴリに共通な重み付けを行う．

特定のカテゴリとの関連度を表す尺度としては，まず語の出現カテゴリ数 $NC_t$ を考えることができる．出現カテゴリ数が少ない語ほど，その語が出現するカテゴリと強く関連しているといえるからである．しかし，出現カテゴリ数 $NC_t$ が少ない語であっても，出現文書数 $N_t$ も少ない場合は，各カテゴリ内における語の出現文書数 $N_c^t$ が少なくなるため，特定のカテゴリと強く関連しているとはいえない．逆に，語が少数のカテゴリに集中して出現していれば，特定のカテゴリと強く関連しているといえる．そこで，語が少数のカテゴリに集中して出現するほど大きな値をもつような尺度を，特定のカテゴリとの関連度を表す尺度として考える．

そのような尺度としては，出現カテゴリ数が少なく出現文書数が多いほど大きな値をもつ $N_t/NC_t$ を考えることができる．しかし，$N_t$ が極端に大きければ，$NC_t$ がある程度大きい場合でも $N_t/NC_t$ は大きな値となってしまう．そこで，$N_t$ が大きい場合の悪影響を軽減するため，分母と分子双方の対数を取った次式を特定のカテゴリとの関連度を表す尺度として用いる．

$$rel(t) = \frac{\log(N_t+1)}{\log(NC_t+1)} \tag{6}$$

公開特許データ約 170 万件に含まれる約 49 万語のうち，$N_t/NC_t$，$\frac{\log(N_t+1)}{\log(NC_t+1)}$ それぞれの値が大きい語を表 1, 表 2 に示す．ここで，特許データに付与されている国際特許分類のうち，1233 種類の分類を分野情報として利用した．上から順に上位 1 位, 1001 位, $\cdots$, 10001 位の語とその値を出現文書数 $N_t$ と出現カテゴリ数 $NC_t$ とともに列挙した．なお，同じ順位の語が複数ある場合は無作為に語を選択した．表 1 では，「発明」「疑似」「結束」など，出現カテゴリ数と出現カテゴリ数の両方が多く，必ずしも特定のカテゴリと強く関連しているとはいえない一般的な語が上位に出現している．一方，表 2 では，出現カテゴリ数が少なくかつ出現文書数の多い語が上位に列挙されており，少数のカテゴリに集中して出現するほど大きな値となっていることが分かる．

次に，特定のカテゴリとの関連度 $rel(t)$ がある閾値 $th_r$ よりも大きい場合はカテゴリ（またはカテゴリ集合）ごとに語の重み $cdficf(C,t)$ を付与し，そうでない場合は全カテゴリに共通する重みを付与する．$cdficf(C,t)$ のうち，カテゴリに依存する部分 $\frac{N_c^t}{N_c}$ に対し，$N_c^t$, $N_c$ それぞれにおける，すべてのカテゴリの合計値 $\frac{\sum_c N_c^t}{\sum_c N_c}$ をカテゴリに依存しない値として与える．すると，$cdficf(C,t)$ に対するカテゴリに依存しない値は次式によって与えられる．

$$\begin{aligned}
&\log(\sum_{c\in C}\frac{\sum_c N_c^t}{\sum_c N_c}/\#C+1)\cdot icf(t)\\
&=\log(\sum_{c\in C}\frac{N_t}{N}/\#C+1)\cdot icf(t)\\
&=\log(\frac{N_t}{N}+1)\cdot icf(t)
\end{aligned}$$

以上より，語とカテゴリの関係を用いた語の重み $weight_{cat}(C,t)$ は次式のようになる．

$$\begin{aligned}
&weight_{cat}(C,t)\\
&=\begin{cases} cdficf(C,t) & (rel(t)>th_r)\\ \log(\frac{N_t}{N}+1)\cdot icf(t) & (rel(t)\leqq th_r)\end{cases}
\end{aligned} \tag{7}$$

表 3　国際特許分類の階層の例

| 分類記号 | 階層 | タイトル |
|---|---|---|
| G | セクション | 物理学 |
| G02 | クラス | 光学 |
| G02B | サブクラス | 光学要素，光学系または光学装置 |
| G02B 21 | メイングループ | 顕微鏡 |
| G02B 21/22 | サブグループ | 立体視装置 |

表 4　各検索手法による平均精度の比較

| 手法 | 入力 A | 入力 B |
|---|---|---|
| TFIDF | 0.1184 | 0.0824 |
| SameCat | 0.1395 | 0.0846 |
| ICF*IDF | 0.0930 | 0.0770 |
| CDFICF*TFIDF（判別なし） | 0.0934 | 0.0941 |
| CDFICF*TFIDF（判別あり） | 0.1514 | 0.0947 |

### 2.3　$tf \cdot idf$ との組み合わせ

式 (7) において定義した，語とカテゴリの関係を用いた重み付け手法は，特定のカテゴリにおいて特徴的な語に高い重みが付与される．したがって，検索要求に関連するカテゴリ内の文書を見つけ出すことが期待できる．しかし，そのカテゴリ内の文書群の中からさらに検索要求に強く関連する文書を見つけ出すには，語とカテゴリの関係だけでは不十分であり，語と文書の関係を利用することも必要である．そこで，語とカテゴリの関係を用いた重み付けの式 (7) を，語と文書の関係を利用した重み付けの式 (1) の $tf \cdot idf$ と組み合わせ，両者の効果を等しい比重で得ることができるようにするため，次式のように相乗平均を用いることによって重み付けする．

$$weight_{comb}(d, C, t) = \sqrt{weight_{cat}(C, t) \cdot tfidf(d, t)} \quad (8)$$

## 3. 評価実験

提案手法の有効性を検証するため，特許データに付与されている国際特許分類を分野情報として利用して情報検索を行う評価実験を行った．本章では，実験データや評価尺度などの実験条件を説明したのち，各手法による実験結果について述べる．

### 3.1　実験条件

評価のための文書データとして，特許出願された発明のうち拒絶査定を受けた発明と，それに対する先願特許が含まれる公開特許データを用いた．先願特許とは，特許出願された発明に対して，特許として登録すべきでないと判断した根拠となる特許のことであり，拒絶査定を受けた発明と内容が類似している．そこで，拒絶査定を受けた発明 19 件を入力文書とし，それに対する先願特許を正解データとして 5 年分の公開特許データ約 170 万件を対象に検索する実験を行った．正解データの件数は平均約 2.6 件，合計 50 件であった．なお，出願された発明に関する書類と審査の経緯の記録は包袋（ほうたい）と呼ばれる書類に収められており，拒絶査定を受けた発明とその先願特許の情報は，包袋に含まれる拒絶理由通知書を参照することによって得ることができる．

公開特許データに付与されている国際特許分類とは，出願された発明に対して技術的な観点から付与される国際的な分類体系のことである．国際特許分類は，表 3 に示すように階層的構造をもち，5 つの階層に細分化されている．本実験では，上から 3 階層目のサブクラスを分野情報として利用した．文書データに付与されている分野情報は全部で 1233 種類であり，一つの文書に付与されている分野情報の数は平均 1.6 個であった．

特許出願された発明には，その内容が項目別に記述されている．本実験では，発明全体の概要が記述されている「要約」と特許の権利範囲が記述されている「特許請求の範囲」を検索要求の入力とした場合（入力 A とよぶ）と,「要約」,「特許請求の範囲」に加えて，発明の内容が詳細に記述されている「発明の詳細な説明」を検索要求の入力とした場合（入力 B とよぶ）の 2 種類の検索を実施した．19 件の検索要求において，入力 A, B から抽出された検索語数の平均はそれぞれ約 65 語，約 378 語であった．入力 A には発明の概要や要点が記述されているため，検索語数が少なく重要な語を含む割合が高い．一方，入力 B には詳細な内容が記述されているため，検索語数が多く重要でない語の割合が高いという特徴がある．この 2 種類の入力に対する検索精度を評価することにより，それぞれの語の重み付け手法が，重要な語と重要でない語それぞれの特徴をどの程度的確にとらえることができるかを検証した．

検索の手順としては，まず形態素解析によって検索要求から名詞等の語を抽出し，検索語集合 $T_q$ とする．次に，検索対象文書の各文書に対して同様に形態素解析を行って抽出された語の集合を $T_d$ とし，文書に付与されたカテゴリ（国際特許分類）の集合を $C$ として，次式によって文書 $d$ のスコアを付与した．

$$score(q, d, C) = \sum_{t \in T_q \cap T_d} weight(d, C, t) \quad (9)$$

ここで，$weight(d, C, t)$ は語の重みであり，語の重み付け手法によってそれぞれ異なる値が付与される．なお，検索要求中の検索語の頻度情報は利用せず，すべての検索語の重みを 1 とした．上記の文書スコアを用いて検索対象文書をランク付けし，平均精度 (average precision) によって語の重み付け手法の検索精度を評価した．

### 3.2　実験結果

各検索手法の平均精度を表 4 に示す．検索手法の内容はそれぞれの以下の通りである．

- TFIDF: 式（1）による，$tf \cdot idf$ に基づく重み付け手法.
- SameCat: 手法 TFIDF でランク付けしたのち，正解文書に付与されているいずれかのカテゴリに含まれる文書のスコアに一定の値を加点し，それ以外の文書よりも上位にランク付けされるようにする手法．実際の検索においては正解文書のカテゴリを参照することは不可能であり，Murata ら [2] による，カテゴリ単位で文書のスコアを修正するというアプローチの特徴を分析するために，理想的な条件下でランク付けする手法である．文書のスコアに加点する値は正解文書に付与されているカテゴリに関わらず一定であるが，正解文書に付与されているカテゴリの文書以外のランクを確実に下げることができるため，高い精度を得ることができる．

表 5　手法 TFIDF と手法 SameCat の平均精度および正解文書の順位の比較

| 検索要求 | 正解 | TFIDF | | SameCat | |
|---|---|---|---|---|---|
| | 文書数 | 平均精度 | 順位 | 平均精度 | 順位 |
| 1 | 1 | 0.0021 | 467.0 | 0.0033 | 301.0 |
| 2 | 1 | 0.0024 | 421.0 | 0.0044 | 229.0 |
| 3 | 2 | 0.0670 | 22.0 | 0.0670 | 22.0 |
| 4 | 2 | 0.0675 | 104.0 | 0.0765 | 102.5 |
| 5 | 3 | 0.0287 | 245.0 | 0.0289 | 234.0 |
| 6 | 2 | 0.0185 | 136.5 | 0.0189 | 135.0 |
| 7 | 2 | 0.1026 | 14.5 | 0.1270 | 11.5 |
| 8 | 1 | 0.0110 | 91.0 | 0.0122 | 82.0 |
| 9 | 3 | 0.0536 | 208.7 | 0.0536 | 208.0 |
| 10 | 4 | 0.2721 | 53.8 | 0.2734 | 52.0 |
| 11 | 5 | 0.0369 | 191.2 | 0.0370 | 188.6 |
| 12 | 2 | 0.1706 | 128.5 | 0.5048 | 105.5 |
| 13 | 4 | 0.5098 | 89.8 | 0.5098 | 89.5 |
| 14 | 4 | 0.1184 | 60.3 | 0.1191 | 58.0 |
| 15 | 1 | 0.0370 | 27.0 | 0.0526 | 19.0 |
| 16 | 3 | 0.0142 | 135.3 | 0.0197 | 97.7 |
| 17 | 2 | 0.5263 | 19.5 | 0.5294 | 17.5 |
| 18 | 7 | 0.2091 | 64.1 | 0.2095 | 63.6 |
| 19 | 1 | 0.0017 | 573.0 | 0.0041 | 246.0 |
| 平均 | 2.6 | 0.1184 | 160.1 | 0.1395 | 119.1 |

- ICF*IDF: Zhao ら [3] と同様に，特定のカテゴリと強い関連をもつ語に高い重みを付与するというアプローチに基いて，語 $t$ に対し全カテゴリに共通な重みとして $tf(d,t)\cdot\sqrt{icf(t)\cdot idf(t)}$ を付与する手法.
- CDFICF*TFIDF（判別なし，判別あり）：提案手法．式（8）により，手法 TFIDF と手法 CDFICF を組み合わせた重み付けをする．特定のカテゴリと強く関連する語を判別しなかった場合，すなわち式（7）において $th_r = 0$ とし，すべての語に対して式 (5) を適用した場合と，判別した場合（$th_r$ の値は 1.8 とした）の 2 通りを評価した．

表 4 より，入力 A，入力 B とも提案手法である CDFICF*TFIDF（判別あり）がもっとも高い精度を得ていることが分かる．また，全体的に入力 A の方が入力 B に比べて高い精度を示している．

## 4. 考　　察

評価実験の結果をもとに，分野情報を利用した従来手法 SameCat および ICF*IDF の特徴，カテゴリごとに異なる語の重み付けの効果，特定のカテゴリと強く関連する語を判別した場合の効果について考察する．

### 4.1　従来手法の特徴

手法 SameCat では，正解文書と同じカテゴリの文書を上位にランク付けするため，元となる手法 TFIDF よりも精度が低下することはないにもかかわらず，入力 A,B ともに提案手法より精度が低かった．入力 A における，TFIDF，SameCat それぞれの検索要求ごとの平均精度と正解文書の順位の平均を表 5 に示す．すべての検索要求において，SameCat の精度は TFIDF よりも向上しているか同じであるが，精度が大きく向上した検索要求は少ない．また，正解文書の順位の平均が大きく向上したも検索要求も少なかった．これは，最初の TFIDF によるランク付けの結果において，正解文書よりも上位にランク付けされた文書の多くに正解文書と同じカテゴリが付与されていたことを示している．このような状況においては，語とカテゴリの関係を文書のスコア付けに利用せず，カテゴリ単位で文書のスコアを修正するのみというアプローチでは高い効果が得られないことが分かる．

また，手法 ICF*IDF は入力 A,B とも TFIDF と比較して低い精度であり，特に入力 A では TFIDF との差が大きかった．TFIDF と ICF*IDF の語の重み付けにおける違いは ICF の有無であり，入力 A は重要な語を高い割合で含むことから，特に重要な語に対する語の重み付けの効果は，IDF と比べて ICF では低いことが分かる．これは，カテゴリ数が文書数に比べて圧倒的に少なく，ICF では語の細かい特徴をとらえることができないためであると考える．

### 4.2　カテゴリごとに異なる語の重み付け

入力 A と入力 B の精度の違いに注目すると，手法 TFIDF, ICF*IDF では入力 A の方が精度が高かったが，提案手法（判別なし）ではほぼ同じ精度であった．従来手法である前者と比べて，提案手法では式（3）のカテゴリ内出現文書数（CDF）を用いてカテゴリ（またはカテゴリ集合）ごとに異なる語の重み付けをしている点が違っており，CDF を用いたかどうかが精度の差となって現れている．入力 B は入力 A の検索語に加えて重要でない語を多く含むため，CDF を用いない従来手法ではそれらの語がノイズとなって精度が低下するのに対し，CDF を用いる提案手法ではそのような語のノイズを軽減することができたと考える．たとえば，語の出現文書数や出現カテゴリ数が少ない場合，$idf(t)$ や $icf(t)$ の値が大きくなるため従来手法では高い重みが付与される．これに対し，提案手法ではそのような語のうち，CDF が小さい，つまりあるカテゴリに集中して出現していない語はそのカテゴリにおける重要度が低いとみなし，低い重みを付与することができる．このように，提案手法では語の重み付けに CDF を追加することによって，重要でない語の重みを低くすることができたと考える．

### 4.3　特定のカテゴリと強く関連する語の判別

提案手法である，CDFICF*TFIDF（判別なし）と CDFICF*TFIDF（判別あり）の精度を比較すると,「判別あり」の方が高い精度を示しており，特定のカテゴリとの関連が強い語とそうでない語の判別が有効であったことが分かる．また，特に入力 A において「判別あり」の精度が大きく向上していた．重要な語を高い割合で含む入力 A において,「判別なし」では特定のカテゴリとの関連が強くない語に対してもカテゴリによって異なる重みを付与しており，特に重要な語の場合，カテゴリによって語の重みが大きく変化するため，精度が低下する.「判別あり」では，このような語が与える影響を軽減することができ，高い精度を得ることができたと考える．

## 5. おわりに

本研究では，分野情報を利用した情報検索手法において，特定のカテゴリとの関連度によってカテゴリごとに異なる語の重み付けを行うかどうかを判別する手法を提案した．国際特許分類が分野情報として付与されている公開特許データを対象とした評価実験を行い，提案手法が従来手法よりも高い性能を示すことが分かった．特に，検索要求が短く重要な語を高い割合で含む場合，分類情報を利用する従来手法は分類情報を利用しない $tf \cdot idf$ による重み付け手法と比較して精度が低下するのに対し，カテゴリごとに異なる語の重み付けを行うかどうかを判別する提案手法では精度が大きく向上することが分かった．本研究では，公開特許データに付与された階層的なカテゴリのうち，特定の階層のカテゴリのみを利用した場合の検索性能を検証した．今後の課題として，分類階層による検索性能の違いの検証や，複数の分類階層を同時に利用した語の重み付け手法の検討を考えている．

## 文　献


[1] K. Cho and J. Kim, “Automatic Text Categorization on Hierarchical Category Structure by using ICF (Inverted Category Frequency) Weighting,” Proceedings of KISS conference, pp.507-510, 1997.

[2] Masaki Murata, Kiyotaka Uchimoto, Hiromi Ozaku, Qing Ma, Masao Utiyama and Hitoshi Isahara, “Japanese Probabilistic Information Retrieval Using Location and Category Information,” the Fifth International Workshop on Information Retrieval with Asian Languages (IRAL2000), pp.81–88, 2000.

[3] Ying Zhao and George Karypis, “Improve Precategorized Collection Retrieval by Using Supervised Term Weighting Schemes,” Proceedings of the IEEE International Conference on Information Tehnology: Coding and Computing (ITCC 2002), pp.16–21, 2002.

[4] Akiko Aizawa, “The Feature Quantity: An Information Theoretic Perspective of Tfidf-like Measures,” Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000), pp.104–111, 2000.

[5] G. Salton and C. Buckley, “Weighting Approaches in Automatic Text Retrieval,” Infrmation Processing and Management, Vol 24, No. 5, pp. 513–523, 1988.

[6] S. E. Robertson, S. Walker, S. Jones, and M. M. Hancock-Beaulieu, “Okapi at TREC-3,” Proceedings of the third Text REtrieval Conference (TREC-3), pp.109–126, 1994.