教員データにおける高頻度語

関，隆宏
九州大学大学評価情報室

安元，裕司
九州大学大学院システム情報科学府

廣川，佐千男
九州大学情報基盤センター

出版情報：研究報告 ： 自然言語処理. 2005 (22), pp.1-8, 2005-03-10. 情報処理学会
バージョン：

教員データにおける高頻度語

関　隆宏* 安元 裕司† 廣川 佐千男‡

要旨. キーワードの重要度をどのように設定するかは，検索システムの実現において重要な課題である．本稿では各大学教員が書いた研究活動概要の文書を対象として，重要なキーワードがどのようなものであるかについて，単語の出現頻度に基づき分析を行った．Web 文書のように文書が多種多様な場合には，単純な出現頻度より tfidf のような値が標準的に用いられる．しかし，同種の文書群を対象とする場合には，共通に現れる高頻出の単語も特徴的な単語として考えなければならない．本稿では，筆者らが開発している九州大学研究者データベースに蓄積された約 2000 人の教員情報を具体的対象として，単語の出現頻度，使用者数，複数回使用者数の 3 つの尺度が分野特定性の識別に有効であることを示す.

Frequent Words in Research Activities of University Researchers

Takahiro Seki* Yuji Yasumoto† Sachio Hirokawa‡

Abstract. It is an important problem in the search engine how to set the degree of importance to each key word. In this report, we will analyze the frequent words that appear in the documents of outline for university researchers. We consider the DF (document frequency) and TF (term frequency) instead of other standard evaluation, e.g., tfidf. The evaluation tfidf is useful for a variety of documents like Web documents to eliminate frequent words that commonly appear in any sentences. Nevertheless, such frequent words are important to the documents in specific area as this report considers. We analyzed frequent words in documents of more than 2000 university researchers in Kyushu University.

1　はじめに

近年，教員の研究内容や研究業績，教育内容といった教員情報を印刷物の形ばかりでなく，インターネット上に公開する大学が増えている．そしてそこに教員検索機能を付加する大学も多い．この検索機能は例えば，ある教員に興味を持つ人がその教員について詳しく知りたい場合や，企業などが大学の教員と共同研究を行うためにはどの教員が適当なのか知りたい場合に利用されている．一方，教員データはただ単に社会に公開するばかりでなく，大学の経営戦略の策定等にも利用され (例えば [8])，その重要性は高くなっている．このように，教員を検索したり，大学の経営戦略を策定したりする際に，教員検索や教員クラスタリングといった情報検索技術が利用されると考えられるが，このなかでキーワード抽出は本質的に重要な役割を果たしている．そして，検索システムの実現においてキーワードの重要度の設定は重要な課題であり，さまざまな方法が提案されている．そのうちで最も基本的なものは語の出現頻度に基づくものである．

現在の情報検索技術では文書からキーワードを抽出し，その集合で文書の内容を近似することにより検索を行うのが一般的である．そうすると，文

*九州大学大学評価情報室・Office for Information of University Evaluation, Kyushu University
†九州大学大学院システム情報科学府・Graduate School of Information Science and Electrical Engineering, Kyushu University
‡九州大学情報基盤センター・Computing and Communications Center, Kyushu University

書中からその文書を特徴付けるキーワードをもれなく抽出することが重要になる．文書の特定性を高くするには，その文書には現れるが，他の文書には現れないようなキーワードを選択すればよい．しかし，文書にあまりに特化したキーワードだけを選ぶと，検索質問でそのキーワードが用いられる可能性も低くなり，その文書が検索されにくくなるという問題が発生する．一方，一般によく用いられる語 (「一般語」とも呼ばれる) をキーワードとして用いると，このキーワードは多くの文書のキーワードとなる可能性が高くなる．したがって，検索質問中でこのようなキーワードが用いられると，多くの文書が検索されることになる．しかし，検索されたすべての文書が必ずしもユーザが必要とするものとは限らない．

キーワードを分析する際に，語の特徴量を使うことが考えられる．語の特徴を表す数量的尺度は，総頻度 (総出現回数) に代表される「網羅性」，idf や信号雑音比に代表される「特定性」，情報利得に代表される「識別性」，tfidf に代表される「代表性」に分類される ([3])．したがって，文書の特定性という点からは出現文書数を，一般語という点からは総頻度を調べることがまず考えられる．一般に 1 つの文書を固定したとき，高頻度のキーワード (以下「高頻度語」と呼ぶ) はその文書を特徴づける語と考えられる．また，文書群で考えると，高頻度語はその文書群を特徴づける語であると考えられる．ここで，ある文書群においてある語が高頻度語となる理由としては，(1) 多くの文書で用いられている，(2) 特定の文書で多く用いられている，の 2 つが考えられる．どちらの理由によるかを明らかにするためには，高頻度語の出現文書数がわかればよい．

また，多くの語は文書に繰り返し出現し，一度出現した語は再び出現する傾向にある．その度合を示す特徴量の一つに反復度 (adaptation) がある．反復度を求めるいくつかの手法が [1] で提案され，[2]，[7] では，[1] で述べられている df_2/df による推定に注目している．ここで，語を固定したとき，df はその語を含む文書の頻度，df_2 はその語を 2 回以上含む文書の頻度である．[1] では英語に関する反復度の分析を行い，反復度が語の内容量を表す特徴量であることを示している．また，[2]，[7] では，日本語論文抄録，中国語新聞記事，日本語新聞記事に関する反復度の分析を行っている．さらにこれらの論文では，日本語に代表される語分割の必要な言語においてキーワード境界を特定できるという点から反復度の有用性を指摘している．

本稿では，筆者らが開発している教員データベース ([5]) の研究紹介部分に現れる名詞に着目し，そこでの各種の頻度に関する分布状況と高頻度語に制限した場合の分析を行う．Web 文書のように多種多様な文書からなる場合には，重要キーワードを抽出する際に単純な出現頻度より，tfidf のような値が標準的に用いられる．しかし，同種の文書群を対象とする場合には，共通に現れる高頻出語も特徴語として考えなければならない．本稿では，その手がかりとして総頻度，使用者数，複数回使用者数，反復度に関する全体的な状況を調べ，特殊な語に共通する特色を明らかにする．この特殊な語の多くは分野特定性を持つ (あるいは持たない) という特性がある．教員検索場面を考えると，個人特定性はもちろん重要であるが，分野特定性も重要な要素と考えられる．それは，検索の粒度にかかわる問題であり，厳格な検索要求では個人特定性があれば十分であるが，おおらかな検索要求では分野特定性に意味があると考えられるからである．また，高頻度語は一般にストップワードとみなされるが，教員データにおける高頻度語には分野特定性を持つと考えられる語が含まれており，そのすべてをストップワードにできないと考えられる．したがって，高頻度語の分析から語の分野特定性の有無を明らかにする可能性があるといえる．本稿では，この分野特定性の識別に総頻度，使用者数，複数回使用者数の 3 つの尺度が有効であることを示す．

以下，2 章では，まず総頻度と使用者数の分布状況を調べ，3 章では，それらの間の関係ならびに教員データにおける反復度の性質について調査する．そして，4 章では高頻度語に注目してそれらを詳細に分析する．最後の 5 章でまとめを行う．

2　総頻度と使用者数

実験で用いたデータは 2005 年 1 月現在の「九州大学 研究者情報」([4]) から「研究・教育・社会活動概要」部分を抽出し，形態素解析器「茶筌」([10]) により名詞および不明語（これは茶筌が品

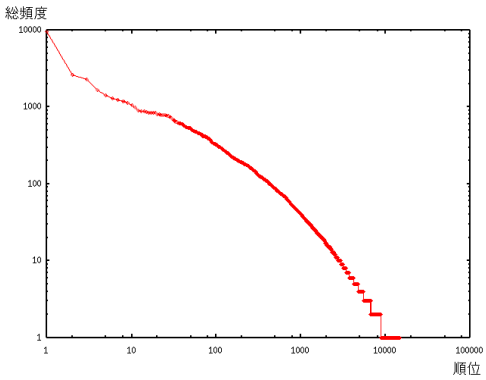

図 1: 順位と総頻度の関係

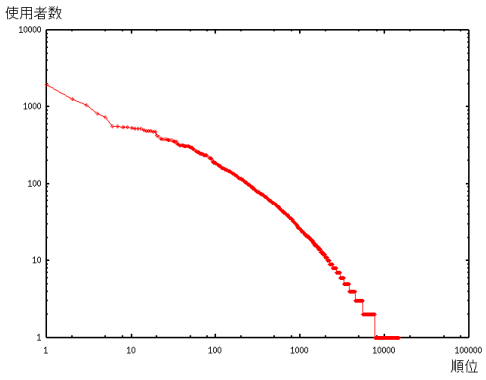

図 2: 順位と使用者数の関係

順位	総頻度	語
1	9623	研究
2	2621	教育
3	2318	活動
4	1662	開発
5	1432	社会
6	1299	細胞
7	1248	解析
8	1187	環境
9	1142	構造
10	1055	機能

表 1: 総頻度上位 10 語

順位	使用者数	語
1	1974	研究
2	1257	教育
3	1049	活動
4	828	社会
5	730	開発
6	567	解析
7	559	担当
8	548	現在
9	543	大学院
10	538	指導

表 2: 使用者数上位 10 語

詞を判別できなかった語であるが，多くは外国語あるいは専門用語である）のみを抽出した語についての総頻度および使用者数である．なお，教員数は 2147 人，出現語数は 14588 語である．本稿では，$tf(w)$ は語 w の総頻度，$df(w)$ は語 w の使用者数，$df_2(w)$ は語 w を 2 回以上使った人数とする．さらに，$tf(w)$，$df(w)$，$df_2(w)$ をそれぞれ tf，df，df_2 と略記し，それぞれ「総頻度」,「使用者数」,「複数回使用者数」と呼ぶ．本章では，総頻度，使用者数，複数回使用者数と順位の関係について述べる．

まず，総頻度 (tf) について，縦軸に総頻度，横軸に順位をとった両対数グラフを図 1 に示す．総頻度についてジップの法則が成り立つことがわかる．参考として，表 1 に総頻度が高かった上位 10 語を記す．

次に，使用者数 (df) について，横軸に順位，縦軸に使用者数をとった両対数グラフを図 2 に示す．使用者数についてもジップの法則が成り立つことがいえる．参考として，表 2 に使用者数の多かった上位 10 語を記す．

最後に，複数回使用者数 (df_2) について，横軸に順位，縦軸に複数回使用者数をとった両対数グラフを図 3 に示す．複数回使用者数に関してもジップの法則が成り立つことがいえる．参考として，表 3 に複数回使用者数が多かった上位 10 語を記す．

3 各種の値の関係

本章では，総頻度 (tf)，使用者数 (df)，複数回使用者数 (df_2) の間の関係について見ていく．2 つの指標の比をとることで特異な語を明らかにすることができるが，それはグラフにしたときに全体

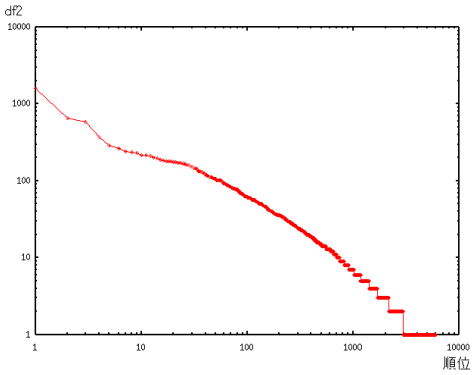

図 3: 順位と複数回使用者数の関係

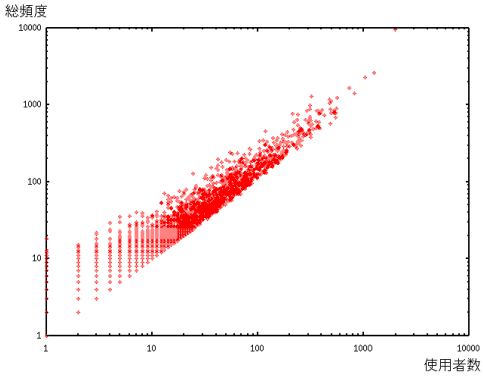

図 4: 総頻度と使用者数の関係

順位	複数回使用者数	語
1	1600	研究
2	662	活動
3	581	教育
4	375	開発
5	292	解析
6	264	社会
7	244	機能
8	237	構造
9	229	環境
10	216	講義

表 3: 複数回使用者数上位 10 語

総頻度	使用者数	語
9623	1974	研究
995	311	分子
889	310	委員
767	215	平成
758	239	遺伝子
456	118	経済
243	54	プラズマ
196	42	顎
127	24	地震
18	1	スペイン

表 4: tf と df からの特異な語

から外れた点として出現することになる．

3.1 総頻度と使用者数の関係

単純に考えれば，総頻度と使用者数は比例関係にあるように思われる．図 4 は横軸に使用者数，縦軸に総頻度をとった両対数グラフである．総頻度と使用者数の相関係数は 0.88 であるから，強い相関があるといえ，総頻度と使用者数はほぼ比例関係にあるといってよい．

図 4 の概形は三角形になっているが，そこから少々外れた語は特異な語と考えられる．図 4 に則していえば，グラフの上方に外れたものが特異な語である．この特異な語は df を固定したときに tf/df が相対的に大きい語，すなわち，平均すれば使用者が高い頻度で使っている語である．目視により抽出した特異な語についての総頻度ならびに使用者数を表 4 に示すが，ここに出現する語の多くは分野特定性が高いことが指摘できる．

3.2 反復度

本節では，反復度に関して述べる．本稿では，[2], [7] にならい，$df_2(w)/df(w)$ を語 w の反復度とする．また，これを df_2/df と略記する．

まず，反復度の基本的データとなる使用者数 df と複数回使用者数 df_2 の関係について見ていく．単純に考えれば，df と df_2 は比例関係にあるように思われる．図 5 は横軸に df，縦軸に df_2 をとった両対数グラフである．ここで，df と df_2 の相関係数は 0.91 であるから，強い相関があるといえ，df と df_2 はほぼ比例関係にあるといえる．

図 5 の概形は三角形になっているが，そこから少々外れた語は特異な語と考えられる．図 5 に則

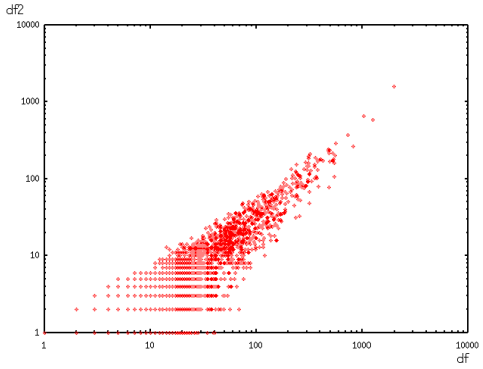

図 5: df_2 と df の関係

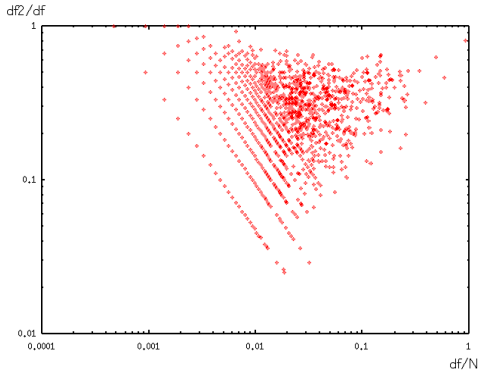

図 6: df_2/df と df/N の関係

df	df_2	語
543	107	大学院
484	78	学部
316	48	主
257	33	大学院生
235	31	最近
156	16	業績
153	16	観点
120	10	卒業
68	2	将来
56	2	啓蒙
40	1	最新
39	1	いくつか

表 5: df_2 と df からの特異な語

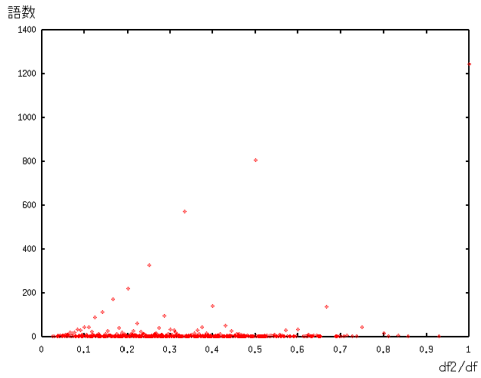

図 7: df_2/df と語数の関係 (1)

していえば，グラフの下方に外れたものが特異な語である．この特異な語は df を固定したときに df_2/df が相対的に小さい語，すなわち，複数回使用されていてもその可能性は低い語である．目視により抽出した特異な語についての使用者数 (df) ならびに複数回使用者数 (df_2) を表 5 に示すが，ここに出現する語は分野特定性が低いことが指摘できる．

次に，反復度 df_2/df と使用者数 df の関係について見ていく．これは [1]，[2]，[7] で論じられてきた df_2/df と df/N の関係と本質的に同じであるから，本稿ではこれを踏襲する．ここで N は全教員数である．図 6 は横軸に df/N，縦軸に df_2/df をとった両対数グラフである．ここで，左上から右下方向に伸びる直線が何本か見てとれるが，これらは左から順に $df_2 = 1, 2, \cdots$ を表す直線であることに注意する．図 6 から，経験的に知られている df/N と df_2/df が 1 桁か 2 桁か違うこと，df_2/df が df/N に対して無相関であるといった，反復度の持つ特徴が教員データにおいても成り立つことがわかる．

最後に，df_2/df の値ごとの語数について見ていく．図 7 は横軸に df_2/df，縦軸にその値の語数をとったものである．ただし，$df_2/df = 0$ となる点 (8608 語) は除いてある．図 7 において突出している点の df_2/df の値は 0.25，0.33，0.5，1 である．これらの箇所の df を観察すると，$df \leq 4$ である語が極めて多い．実際の状況を表 6 に示すが，表中の $n(df \leq 4)$，$p(df \leq 4)$ は $df \leq 4$ であるような語数，割合 (単位は%) をそれぞれ表す．ここで，$df \leq 4$ の語は使用者が 4 人以下であるので，個人特定性が高い語であることに注意する．また，表 6 において，$df \leq 4$ であるような $df_2/df = 0.25$，$df_2/df = 0.33$

df_2/df	語数	$n(df \leq 4)$	$p(df \leq 4)$
0.25	326	230	70.6
0.33	572	400	69.9
0.5	807	699	86.6
1	1245	1244	99.9

表 6: 突出した df_2/df の値

df_2/df の区間	語数	df の平均	df_2 の平均
$(0, 0.25)$	1522	27.49	4.49
$[0.25, 0.33)$	768	32.37	9.32
$[0.33, 0.5)$	1239	31.18	12.49
$[0.5, 1)$	1206	13.45	7.95
$[1, 1]$	1245	1.10	1.10

表 7: df_2/df の区間の平均値

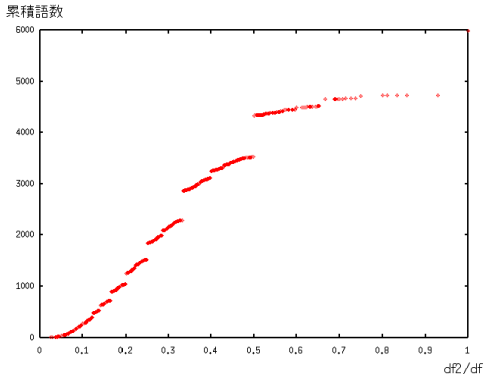

図 8: df_2/df と語数の関係 (2)

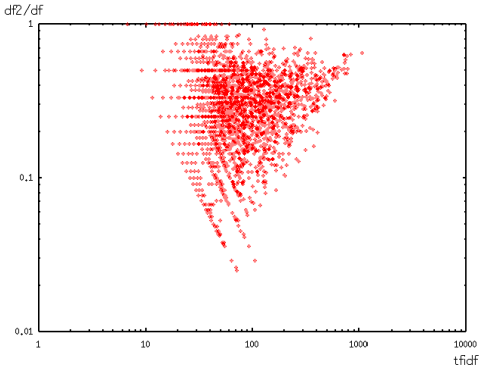

図 9: df_2/df と tfidf の関係

については，それぞれ $(df_2, df) = (1, 4), (1, 3)$ の場合しかなく，複数回利用されることが少ないことも指摘できる．したがって，このような語を複数回使っている教員にとって，その語はその教員の専門を特定する上で重要な語である可能性が高くなると考えられる．

他の値の状況を調べるため，横軸に df_2/df，縦軸にその値以下の語数をとったグラフを図 8 に示す．図 8 を見ると，図 7 で突出した df_2/df の値の箇所でその値以下の語数が急激に増加していることに気づく．参考に df_2/df の値の区間を，$(0, 0.25)$，$[0.25, 0.33)$，$[0.33, 0.5)$，$[0.5, 1)$，$[1, 1]$ に分け，その 5 区間の語数ならびに df，df_2 の値の平均を表 7 に示す．

3.3 反復度と tfidf

本節では，前節で議論した反復度と一般によく知られている tfidf の関連について見ていく．本稿では，語 w の tfidf 値 $tfidf(w)$ の定義として次を採用した．

$$tfidf(w) = tf(w) \cdot idf(w),$$

ここで $idf(w) = \log_{10} \dfrac{N}{df(w)}$，$N$ は教員数である．

図 9 は横軸に tfidf，縦軸に反復度をとった両対数グラフである．図 9 は図 6 に似た形状をしている．これは 3.1 節で述べたように，総頻度と使用者数の間に強い相関があるため，df/N と tfidf にも強い相関があると考えられる．実際，df/N と tfidf の相関係数は 0.81 である．また，3.2 節で述べたように反復度 df_2/df と df/N の間の相関が低いことから，結果として反復度と tfidf の相関が低くなる．したがって，反復度と tfidf は別の指標であると考えられる．

4 高頻度語についての分析

高頻度語は個人特定性が低いが，分野特定性からみると必ずしもそうとはいえない部分がある．実際，分野特定性があると考えられる「細胞」,「分子」,「化学」といった語は総頻度順でそれぞれ 6 位，11 位，16 位である．前章まではすべての語を対象に各種指標から分析してきたが，本章では高頻度語に限定した場合に分野特定性がわかるかどうか検討する．ここでは，総頻度が高かった上位

tf	研究，経済，癌，膜，細胞，平成，水，免疫，空間，合成
df	教育，活動，社会，大学院，現在，開発，指導，担当，学部，基礎

表 8: tf と df からの特徴語

df_2	研究，細胞，遺伝子，平成，分子，癌，膜，経済，委員，エネルギー
df	社会，大学院，教育，学部，現在，学生，指導，中心，基礎，担当

表 9: df_2 と df からの特徴語

100 語に限定して分析を進める．本章の以下の部分では「高頻度語」はこのような語を指すものとする．

4.1 回帰直線からの分析

総頻度 (tf) と使用者数 (df)，使用者数と複数回使用者数 (df_2) の間に強い相関があることをそれぞれ 3.1 節ならびに 3.2 節で指摘した．このとき，回帰直線から離れた語は何らかの特徴を持っていると考えられるが，本節では高頻度語に限定してその特徴を明らかにする．

まず，総頻度と使用者数について，回帰直線と離れている語を表 8 に示した．表 8 の tf の欄，df の欄は回帰直線からそれぞれ総頻度側，使用者数側に離れた上位 10 語を示している．表 8 で tf の欄にある語は 1 人が利用する回数が比較的多い語であると考えられるが，分野特定性の高い語が多いことに気づく．それに対して，df の欄にある語は 1 人が利用する回数が比較的少ない語であると考えられるが，分野特定性の低い語が多いのが特徴である．

次に，使用者数 df と複数回使用者数 df_2 について，回帰直線と離れている語を表 9 に示した．表 9 の df_2 の欄，df の欄は回帰直線からそれぞれ df_2 側，df 側に離れた上位 10 語を示している．表 9 で df_2 の欄にある語は 1 人が複数回利用するのが比較的多い語であると考えられるが，分野特定性の高い語が多いことに気づく．対して，df の欄にある語は 1 人が複数回利用するのが比較的少ない語であると考えられるが，分野特定性の低い語が多いのが特徴である．

上位 10 語			下位 10 語		
順位	df_2/df	語	順位	df_2/df	語
1	0.811	研究	1	0.152	主
2	0.653	細胞	2	0.161	学部
3	0.636	遺伝子	3	0.197	大学院
4	0.633	分子	4	0.203	連携
5	0.631	活動	5	0.206	中心
6	0.623	平成	6	0.211	テーマ
7	0.590	委員	7	0.243	構築
8	0.571	膜	8	0.257	成果
9	0.567	癌	9	0.270	方法
10	0.558	化学	10	0.271	関連

表 10: 高頻度語の df_2/df

上位 10 語			下位 10 語		
順位	tfidf	語	順位	tfidf	語
1	1073.9	細胞	1	311.3	構築
2	834.9	分子	2	315.4	主
3	778.7	開発	3	316.7	連携
4	772.3	環境	4	325.8	成果
5	766.5	平成	5	329.5	効果
6	747.2	委員	6	345.3	発生
7	734.0	化学	7	347.3	テーマ
8	729.7	構造	8	351.1	研究
9	722.7	遺伝子	9	351.2	実習
10	721.7	解析	10	352.7	影響

表 11: 高頻度語の tfidf

4.2 反復度と tfidf からの分析

反復度 df_2/df と df/N，df_2/df と tfidf の間には相関が低いことをそれぞれ 3.2 節，3.3 節で指摘したが，高頻度語に制限した場合に，df_2/df や tfidf が高い語や低い語には何か特徴があるのかどうかについて本節では調べる．

表 10 は高頻度語の df_2/df の上位 10 語ならびに下位 10 語を示したものである．表 10 から，上位 10 語には「細胞」「遺伝子」「分子」「膜」「癌」「化学」のような分野特定性の高い語が多いのに対し，下位 10 語には分野特定性の低い語ばかりであることがわかる．

表 11 は高頻度語の tfidf の上位 10 語ならびに下位 10 語を示したものである．表 11 から，上位 10 語には「細胞」「分子」「化学」「遺伝子」といった

分野特定性の高い語がある一方，下位 10 語は分野特定性の低い語ばかりであることがわかる．

以上のことから，df_2/df あるいは tfidf の低い値から分野特定性の低い語を抽出することができるが，逆に分野特定性の高い語を抽出するにはもう少し別の方法を考える必要があるといえる．

5 おわりに

本稿では教員の研究紹介部分の内容語について，総頻度，使用者数，複数回使用者数の 3 つの観点から分布状況ならびにそれらの値の関連について調査した．さらに，高頻度語について特徴的な点を明らかにした．本稿では教員データにおいて，(1) これら 3 つの尺度ではジップの法則が成り立つこと，(2) 総頻度と使用者数，ならびに使用者数と複数回使用者数の間には強い相関があること，(3) 反復度の持つ特徴が成り立つこと，(4) 反復度と tfidf の相関は低いこと，(5) 高頻度語に限ればこれら 3 つの尺度の関連が分野特定性の識別に有効なこと，(6) 高頻度語に限れば反復度や tfidf の低い語から分野特定性の低い語の抽出が可能なこと，を示した．しかし，実験データ作成時に，形態素解析器「茶筌」を利用しているが，必ずしも専門用語がうまく取り出されていない可能性が指摘される．その点は専門語辞書を搭載するなどしての解決が望まれる．

先に df_2/df と tfidf は別の指標であることを指摘したが，どう違うのかについては全く議論できなかった．これらの本質的な違いがどこにあるのかを明らかにするのが今後の課題である．また，本稿では全部局のデータを対象として分析を進めたが，分野によって状況が変わることが考えられる．部局ごとの分析から分野ごとの違いを明らかにすることが今後の課題として挙げられる．そして，教員データのうち他の項目の場合はどうなのか，他大学の教員データではどうなのかについての実験，あるいは教員データベース以外のデータを用いて同様の実験をすることも考えられよう．

なお，本実験は国立情報学研究所で開発された汎用連想計算エンジン GETA（[6]）を利用した．

謝辞. 本稿執筆にあたり，多数の助言をいただいた九州大学 石野明氏に感謝します．

参考文献

[1] K.W.Church. Empirical estimates of adaption: The chance of two noriegas is closer to $p/2$ than p^2, *Coling*, pp.173-179, (2000).

[2] Y.Takeda, K.Umemura and E.Yamamoto. ”Determinig indexing strings with statistical analysis”, *IEICE Transactions on Information and Systems*, vol.E86-D, No.9, pp.1781-1787, (2003).

[3] 相澤彰子．語と文書の共起に基づく特徴度の数量的表現について，情報処理学会論文誌，Vol.41, No.12, pp.3332-3343, (2000).

[4] 九州大学研究者情報. http://hyoka.ofc.kyushu-u.ac.jp/search/

[5] 杉本典子，金丸玲子，池田大輔，竹田正幸，井上 仁，廣川佐千男. 九州大学自己点検・評価関連情報システム，情報処理学会 第 41 回デジタル・ドキュメント研究会資料，(2003).

[6] 汎用連想計算エンジン GETA. http://geta.ex.nii.ac.jp/

[7] 武田善行，梅村恭司．キーワード抽出を実現する文書頻度分析，計量国語学，Vol.23，No.2，pp.1-26，(2001).

[8] 田中省作，関 隆宏，石野 明，金丸玲子，杉本典子，竹田正幸，廣川佐千男．大学経営における大学評価システムの活用，情報処理学会 第 67 回全国大会予稿集，(2005).

[9] 徳永健伸．情報検索と言語処理，東京大学出版会．(1999).

[10] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原 正幸．日本語形態素解析システム『茶筌』 version 2.2.1 使用説明書. http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1-j.pdf