INFOPRO-OUG
ゆるふわ検索エンジン
解説資料

2022年7月8日（金）14:05～14:35

INFOSTA　OUG　ライフサイエンス分科会

株式会社エムエムツインズ　田邊　稔

ゆるふわ検索エンジンの基本処理フロー

① 検索テキストを1つの文書と見なし、形態素解析を行い、全単語数に占める各単語の割合（TF値）を計算する

② TF値を使ってベクトル計算を行いTF-IDF値を算出し、次元（単語）と次元値（スコア）からその文書の特徴を示す上位n個の単語を抽出

③ 上位n個の単語を使って文書を検索し、単語n個のうち１個でもヒットすればその文書を選定対象とする

④ 選定対象となった文書についてベクトル計算を行い、TF-IDF値を算出する

⑤ 検索テキスト文書（1）と選定対象となった文書（n）の間で類似度（距離）を計算を行う

⑥ 検索テキスト文書と距離が近い文書順にソートして返す

※現在のところ、文書種別や特定の単語にウエイト付けはしていない

参考）ベクトル空間法による検索（1／4）

- TF-IDF ＝ TF（単語の出現頻度）＊ IDF（各単語のレア度）

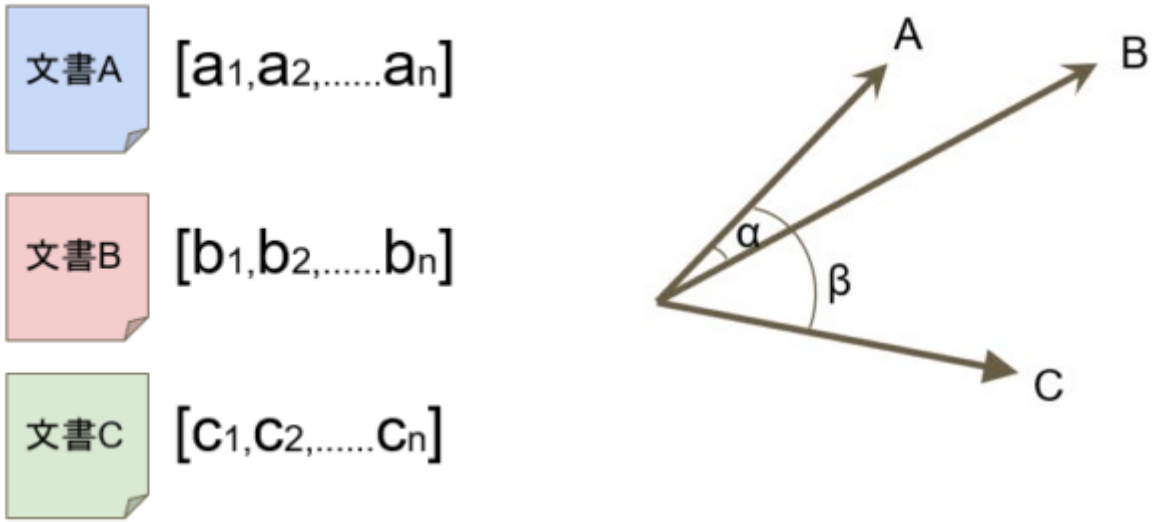
- つまり、「その単語がよく出現するほど」、「その単語がレアなほど」大きい値を示すものとなります。この計算を「各文書の各単語ごと」に行うことで、文書の特徴を判別し易くします。TF-IDF計算で求めた特徴（量）は「特徴ベクトル」とも呼ばれています。

参考）ベクトル空間法による検索（２／４）

- 文書をTF-IDFによってベクトル化したものとする。
- ベクトル化された文書との照合を行うためには、検索文字列もベクトル化する必要がある。検索文字列を一つの文書とみなすことで、前述の計算式を適用してベクトル化することができる。
- これによって、検索対象の文書群も検索文字列も同じ多次元空間（次元数は、全単語の数）上のベクトルとして表現される。
- あとはベクトル同士の比較を行う。
- ベクトル間の類似度は、ベクトル間の角度の小さい（ベクトルが近い）方が類似度が大きいものとする。

参考）ベクトル空間法による検索（3／4）

例えば次のような3つのベクトルを考えてみる。



- 角度αが角度βより小さいので、文書Aに近いのは文書Bということがわかる。
- 実際には、ベクトル間の角度を計測する必要はない。
- cos類似度（コサイン類似度）を用いて、角度が小さいほど類似度が大きな値になるように計算を行う。

参考）ベクトル空間法による検索（4／4）

$$cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{||\boldsymbol{a}|| \ ||\boldsymbol{b}||}$$

※「a・b」はベクトルAとBの内積、||a||はベクトルaの大きさ

- 検索文とすべての文書の類似度を計算し、その類似度の大きな順に検索結果とする。
- すべての文書に対して、この計算式をそのまま適用すると、検索結果を得るまでの時間が大きくなる。実際には、様々な工夫を凝らすことで、検索時の計算量を減らす。例えば、文書のベクトルの大きさは、検索文に関係なく事前に計算できるので、その値を保持しておけば、検索のたびごとに計算する必要はない。

ベクトル計算を単純化してみる（１／６）

- **文書A**: ‘オロパタジン，オロパタジン，オロパタジン，投与，塩酸'
- **文書B**: ‘オロパタジン，皮膚炎，皮膚炎，アレルゲン'
- **文書C**: ‘オロパタジン，酸化鉄，アレルゲン'

まず表1では、全文書に含まれる全ての単語を重複なく列に取り、各文書（A／B／C）を行に取り、各セルにその単語の出現回数を記載した。

＼	オロパタジン	塩酸	アレルゲン	投与	酸化鉄	皮膚炎
文書Aでの出現回数	3	1	0	1	0	0
文書Bでの出現回数	1	0	1	0	0	2
文書Cでの出現回数	1	0	1	0	1	0

表1　各文書における各単語の出現回数（BoW：Bag of Words）

ベクトル計算を単純化してみる（2／6）

・表1のデータを基にtf値を計算する。各文書（例：「文書A」）における「ある単語」（例：「オロパタジン」）の出現回数（例：**3**回）を、全文書に含まれる全ての単語の数（例：表1の列数である**5**列）で割ればよい。3÷5＝**0.6**となる。
・同様にして全ての計算を行ったのが表2である。表2の「文書A」行の「オロパタジン」列を見ると、確かに**0.60**という数値が記載されている。

＼	オロパタジン	塩酸	アレルゲン	投与	酸化鉄	皮膚炎
文書Aでのtf値	0.60	0.20	0	0.20	0	0
文書Bでのtf値	0.25	0	0.25	0	0	0.50
文書Cでのtf値	0.33	0	0.33	0	0.33	0

表2　各文書における各単語の出現頻度（tf：term frequency）

ベクトル計算を単純化してみる（３／６）

次にidf値を計算する。計算の前提として、「ある単語」が含まれる文書の数を表3にまとめる。

＼	オロパタジン	塩酸	アレルゲン	投与	酸化鉄	皮膚炎
全文書（3個）中での文書の数	3	1	2	1	1	1

表3　各単語を含む文書の数

ベクトル計算を単純化してみる（４／６）

・表3を基に、全文書の数（例：表2の行数である**3**行）を、「ある単語」（例：「オロパタジン」）が含まれる文書の数（例：文書A／B／Cの**3**個）で割り、その計算値の自然対数（log）値を算出する。log(3÷3)＝**0**となる。「アレルゲン」は**2**つの文書に含まれるのでlog(3÷2)＝**0.41**となり、「皮膚炎」は**1**つの文書に含まれるのでlog(3÷1)＝**1.1**となる。
・同様にして全ての計算を行ったのが表4である。表4の「皮膚炎」列を見ると、確かに**1.10**という数値が記載されている。

＼	オロパタジン	塩酸	アレルゲン	投与	酸化鉄	皮膚炎
idf値	0	1.10	0.41	1.10	1.10	1.10

表4　各単語の文書間でのレア度（idf：inverse document frequency）

ベクトル計算を単純化してみる（５／６）

・最後に、tf値（表2）とidf値（表4）を掛け算すればよい。文書Aの「オロパタジン」は0.60×0＝**0**となり、文書Bの「皮膚炎」は0.50×1.10＝**0.55**となる。
・同様にして全ての計算を行ったのが表5である。表5の「文書B」行の「皮膚炎」列を見ると、確かに**0.55**という数値が記載されている。

＼	オロパタジン	塩酸	アレルゲン	投与	酸化鉄	皮膚炎
文書Aでのtf-idf値	0	0.22	0	0.22	0	0
文書Bでのtf-idf値	0	0	0.10	0	0	0.55
文書Cでのtf-idf値	0	0	0.14	0	0.37	0

表5　各文書における各単語の重要度（tf-idf）

ベクトル計算を単純化してみる（6／6）

以上でtf-idfの計算は完了となる。文書A／B／Cに対する3つの特徴ベクトルが作成されている（表5の3行）。この3つを表の行と列に取り、各行と各列で示す2つのベクトル同士のコサイン類似度の値を表の各セルに記載したのが表6である。

＼	文書A	文書B	文書C
文書A	1	0	0
文書B	0	1	0.06
文書C	0	0.06	1

表6　各文書の特徴ベクトル同士で計算したコサイン類似度の一覧表

例えば表6の「文書B」行を見ると、
- 「文書B」と「文書A」の類似度は**0**（無関係）
- 「文書B」と「文書B」の類似度は同じなので**1**（同じ）　★原則、検索結果から除外する
- 「文書B」と「文書C」の類似度は**0.06**（わずかに類似）

となっている。

単純化したベクトル計算結果の考察

- **3文書ともに出現している単語(「オロパタジン」)は，文書の特徴を表しているとはいえない**と評価されたことを表している．
- 逆に「皮膚炎」のTF-IDF値は0.55となっているため，**文書Bにのみ出現している「皮膚炎」は文書Aの他の単語「投与」や文書Cの他の単語「酸化鉄」と比較して文書の特徴をよく表している**
- 文書Aと同様に，文書B、C共に出現している「オロパタジン」についてはTF-IDF値が0.0になっているため，文書の特徴を表しているとはいえないと評価されている．
- 一方，「皮膚炎」はどちらも文書Bにしか出現しないため，TF-IDF値は0.55と大きくなっている．ここで注目してほしいのは，「皮膚炎」のTF-IDF値(0.55)と比べて，「酸化鉄」のTF-IDF値(0.37)が低くなっている点である．これは**一方にしか出現しない単語が複数ある場合，出現頻度が高い方が文書の特徴をよく表している**と評価されたことを表している．
- このように，TF-IDF値を計算することで文書に出現している単語のうち，どの単語が文書の特徴をよく表しているかが評価できる．

添付文書Aの検索テキストの単語種類と出現回数(上位40)
※参考として”mecab”による形態素解析の結果を示す

単語	出現回数	出現率
錠	12	0.0547945
患者	11	0.0502283
投与	11	0.0502283
性	10	0.0456621
こと	9	0.0410959
アレロック	8	0.0365297
剤	8	0.0365297
名	7	0.0319635
有効	7	0.0319635
コード	6	0.0273973
等	6	0.0273973
販売	6	0.0273973
皮膚	6	0.0273973
オロパタジン	5	0.0228311
塩	5	0.0228311
塩酸	5	0.0228311
機能	5	0.0228311
承認	5	0.0228311
場合	5	0.0228311
年	5	0.0228311
番号	5	0.0228311
mm	4	0.0182648
アレルギー性	4	0.0182648
そう	4	0.0182648
回	4	0.0182648
開始	4	0.0182648
期間	4	0.0182648
小児	4	0.0182648
注意	4	0.0182648
貯法	4	0.0182648
痒	4	0.0182648
季節	3	0.0136986
月	3	0.0136986
参照	3	0.0136986
治療	3	0.0136986
疾患		
症		
障害		
成人		
成分		

・汎用的な単語の出現率が高く、それらが上位にくると、文書として特徴が出にくい（違った特徴を示す可能性が高い）ため、検索キーワードとして除外する必要があるかもしれない

・逆に、薬名・病名などの専門用語だけになってもレア度が薄まってしまうため、汎用単語との組み合わせが必要だと思われる

添付文書A　検索結果1件目の単語データ(文書内の単語)

単語	TF値
皮膚	0.146908799731450
投与	0.138959356115520
改善	0.113236884041180
オロパタジン	0.110384558751670
痒	0.102218423720910
皮脂	0.088508234485528
発症	0.086839194931629
塩酸	0.072028998257008
欠乏	0.071181478095087
例	0.070461627972919
治療	0.067554052598947
症状	0.067542592903624
検討	0.061303137771454

1件目のデータ(単語の種類数:29)

薬物	0.056625879212158
療法	0.055438252844027
炎	0.054439551287936
予防	0.053999928569476
スコア	0.050162055445495
塩	0.048119623733256
週	0.043177873367830
病因	0.041875119383244
全般	0.034223734334758
＃＃	0.033173789400023
Ｏｌｏｐａｔａｄｉｎｅ	0.032827626776812
そう	0.030829936669388
～)／	0.028995631747207
効果	0.027613449467600
)／	0.026825669292128
群	0.026095295223200

添付文書A　検索結果2件目の単語データ(文書内の単語)

単語	TF値
重症	0.112165691473690
塩酸	0.104078598398890
アレルギー性	0.103405030066460
心身	0.102022236884260
鼻炎	0.101861564997970
副作用	0.090485900611388
障害	0.078372907991971
毒性	0.075752650754368
フェキソフェナジン	0.074980950608016
プソイドエフェドリン	0.073294328190358
配合	0.063482833369641
Ｆｅｘｏｆｅｎａｄｉｎｅ	0.063245886450739
行動	0.060430157647438
Ｐｓｅｕｄｏｅｐｈｅｄｒｉｎｅ	0.060066312227987
投与	0.059493302312453
誘発	0.058672143621269

2件目のデータ(単語の種類数:33)

交感神経	0.057851187335234
異常	0.056582674877825
治療	0.054229170664292
閉	0.051445708418109
鼻	0.050365235044774
錠	0.045365823819914
発現	0.045245487917357
塩	0.034765324486183
剤	0.034012808381132
耳鼻咽喉科	0.033563319919627
アレルギー	0.031161521378989
ディレグラ	0.030692977381085
化学	0.029282996607214
抗	0.027464382397979
呼吸	0.026138168678442
刺激	0.025393618894754
＃＃＃＃	0.024286536992455

添付文書A　検索結果149件目と150件目の単語データ(文書内の単語)

149件目のデータ(単語の種類数:13)

単語	TF値
疼痛	0.23961954586283
治療	0.21016816364117
Ｔｒａｍａｄｏｌ	0.19825389834301
毒性	0.19572240431943
副作用	0.19482396378551
Ａｃｅｔａｍｉｎｏｐｈｅｎ	0.17771620410419
トラマドール	0.16516364320957
配合	0.15376952971758
アセトアミノフェン	0.14208001923851
慢性	0.13584928911021
塩酸	0.11204510839979
錠	0.09767653918757
投与	0.09607066595641

・1件目、2件目に比べて単語の種類数が少ないため、各単語のTF値は高い
・150件目は記号、数字などノイズが多い

150件目のデータ(単語の種類数:20)

単語	TF値
バラシクロビル	0.212457981477380
ｍｇ	0.192441636734650
帯状疱疹	0.180907656577080
試験	0.138694335250320
盲	0.114380083169390
臨床	0.105086770803210
投与	0.097150111641309
検	0.094578497551682
塩酸	0.084978031651526
比較	0.077965669301479
錠	0.074080521293942
相	0.067495346441535
至	0.063737980584906
用量	0.061639299952474
重	0.053994014027481
3000	0.053812841047677
＃＃	0.052183488943856
適用	0.049453221624494
Ｖａｌａｃｉｃｌｏｖｉｒ	0.049042906613113
検討	0.048215951056200

添付文書Aの検索結果の考察について

- 正直なところ、正確には説明が付かない
- 検索テキストおよび文書データ中の単語数によっても評価が変わる
- 登録する文書データが増えれば評価も変わる
- 薬名・病名などの専門用語＝「特徴的な単語」と評価されるとは限らない
- 検索テキストや登録済みの文書データ中の記号や数字などノイズによる影響もある
- 秘伝のタレ的な要素もある

TF-IDFの特性（１／２）

- 文書中の「特徴的な単語」を捉えること
- 「特徴的な単語」とは、その文書を代表する単語。
- 検索の場合には、すべての単語を均等に扱うのではなく、特徴的な単語を用いることで、より適切な文書を順位付けすることができる。
- その意味では、BOW（Bag of Words）よりはTF-IDFの方が、より特徴的な単語を表すことができていると言える。
- もちろん、TF-IDFの計算結果が特徴語を正確に表しているとは限らないが、単純な計算式で、それなりの効果が出ていると言える。

TF-IDFの特性（２／２）

- 文書に含まれる単語数が多いほどTF値が小さく、単語数が少ないほどTF値が大きくなる（分母に「すべての単語の出現回数の和」をとっているため）
- そのため、複数の文書から重要度の高い単語を抽出して絶対評価でTF-IDF値を比較する場合、文書毎の単語数（つまり、ゆるふわ検索エンジンに登録する文書毎のテキストの量＋単語の種類数）の差による影響が多く出てしまう。

発展途上のゆるふわ検索エンジン

a. 形態素解析の課題（薬名や病名など専門用語が正確に解析できていない。記号や助詞などのノイズが正しく除去されていない。）
b. 特徴量（ベクトル）計算の課題（形態素解析が正しくないと特徴量計算も不正確となる）
c. 文書間の類似度（距離）計算の課題（ベクトル計算が正しくないと距離計算も不正確となる）
d. 検索速度と検索精度のトレードオフ（速度を出すために精度をやや落とさざるを得ない）
e. レスポンス向上のため、現在複数台のサーバに分散して並行検索（マルチサーチ）を行い、結果をマージしてスコア順にソートしているため、サーバ1台のみの場合とスコアが異なる可能性がある

検索精度に関する当面の課題

1. 形態素解析における課題
   - 日本語用語抽出にあたっては形態素解析器および形態素辞書が必要であるが、多くのシステムで用いられている形態素辞書は、日常用いられる日本語文の解析を想定したものであり、対象分野の特殊性を考慮したものとはなっていない。
   - 文書単位の辞書テーブルに記号などノイズが含まれている

2. 登録済みドキュメントのタイトルや抄録をそのまま入れてもランクインしないケースがあることに対する疑問に対して説明がしづらい。
   - 類似度がまったく同じもの（類似度を1とする）については、基本的に検索結果から除外される仕様としていることも一要因。

将来的な拡張計画案（１／２）

案１）医学専門用語辞書の活用

- 医学論文からのサンプルテキストに対して、一般用語からなる一般辞書ファイルと専門用語からなる専門辞書ファイルを用いて形態素解析を行い、辞書ファイルの違いによる形態素解析の精度を比較する。

案２）文書種別や特定の単語にウエイト付けを行う

- 原著論文のウェイトを高める、主題と判断できる語のウェイトを高める等

案３）機械学習エンジンと従来型の全文検索エンジンの併用

- ２つの検索エンジンの検索結果をどのように重みづけ＆ランキングするか

将来的な拡張計画案（２／２）

案４）キーフレーズ抽出の活用

- あまり本質的でない文言（記号、挨拶文、どのテキストにも出てきそうな文言など）を無視できる可能性が高いため

案５）患者がよく使う表現の活用

- 患者表現辞書を使って患者テキストの表記ゆれを吸収した意味構造検索を実装するなど一般的な利用者向けのエンジンとする
- 「胸が苦しい」と「胸がムカムカする」など症状の表記の揺れによりヒットしない問題
  - （１）MEDNLPの患者表現辞書
- 「頭が痛い」と「頭は正常だが胃が痛い」など主語が違うのにヒットしてしまう問題
  - （２）係り受け解析：病名に紐づく患者表現を取得