情報検索における索引語の選択的利用

荻林裕憲　徳永健伸　田中穂積
東京工業大学 大学院情報理工学研究科
{ogi,take,tanaka}@cl.cs.titech.ac.jp

1 はじめに

情報検索において文書やクエリはタームの集合という形で表現し，それぞれのタームがその文書にとってどれだけ重要かという情報はタームの出現頻度などを元にした重要度で表現されるのが一般的である．ところが，実際の検索においては重要なタームがかえって検索意図と違う文書を検索してしまうことや，重要度の低いタームであっても文書の選別に役立つこともある．また，近年タームとしてのフレーズの利用が研究されているが，フレーズは必ずしも有効とは限らないという結果が出ている [1]．したがって，単語やフレーズの中から状況に合わせて適切なタームを選択し，検索を行なう技術が必要である．

本論文では，クエリに出現するタームの中から有効なものを選択し，そのタームを用いて再検索を行なう手法について述べる．

2 検索手順

本手法で提案する検索手順を図1に示す．

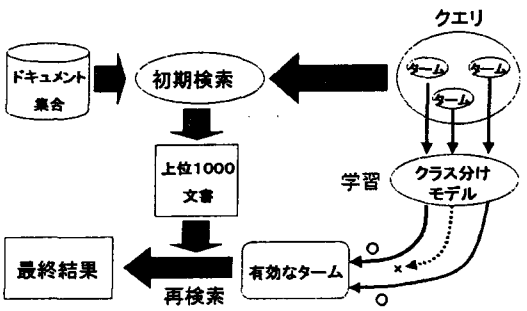

図 1: 検索手順

1. 初期検索
TREC7 において最も結果の良かった Okapi システム [3] の出力の上位 1000 文書を初期検索の結果として使用する．

2. 構文解析
ApplePieParser[4] を用いて，初期検索の結果得られた 1000 文書およびクエリの構文解析を行なう．

3. インデクシング
構文解析した 1000 文書およびクエリからタームを抽出し，重要度を付与する．本手法では品詞なし単一語のターム，品詞つき単一語のターム，フレーズの3種類のタームを使用している．

- 単一語のターム
ApplePieParser による構文解析結果から品詞として名詞，固有名詞，副詞，動詞，形容詞，外来語のいづれかが付与されており，かつ不要語でない単語を “品詞 単語” の形式で抽出する．また，抽出したタームから品詞の情報を削除したものを品詞なし単一語のタームとして使用する．
タームの重要度としては，TF×IDF を用いる．ただし，使用する IDF は品詞なし単一語のタームの場合は検索対象文書全体におけるものを，品詞つき単一語のタームの場合は 1000 文書内のものを利用する．

- フレーズ
ここで言うフレーズとは動詞-目的語，主語-動詞，形容詞-名詞，名詞-名詞の関係にある二単語とその関係の種類を組にしたものである．例えば “Document will discuss government assiscance to ...” という文書から動詞-目的後関係 (vo) として “vo:discuss:assistance” のようなものを構文解析結果より抽出する．
タームの重要度としてはターム頻度のみを用いる．これは，フレーズを用いた場合は TF ×IDF を用いるよりも TF のみを用いた方がやや精度が良いという予備実験の結果による．この理

由としては，もともとフレーズの場合単一語のタームと比較して具体的なタームであるために，IDF を用いなくても話題と関連した文書を検索する可能性が高いこと，そしてリランキング対象として取得した 1000 文書はすでにある話題を表現しており．この中でタームの弁別性を測ることでこの話題特有のタームの重要度が低く設定されてしまうことが考えられる．

さらに，フレーズの場合は 1 語の重要なタームが複数のフレーズに分散して出現するために 重要なタームを含むフレーズでも重要度が低く設定されてしまうという問題が考えられる．そこで，フレーズの TF を単一語のタームの TF を用いて修正を行なう．あるフレーズが "関係:単語 1:単語 2" の形をしているとき，このフレーズの TF を以下の式によって修正する．

$$(\text{新しい TF}) = \frac{(\text{元の } TF)}{2} + \frac{(\text{単語 1 の } TF)+(\text{単語 2 の } TF)}{4}$$

4. 使用するタームの決定

クエリに出現するタームのうち，どれを使用するかを決定する．決定はタームに付与した属性を決定木に与え，出力としてそのタームを使用すべきかを得ることにより行なう．決定木の学習には C4.5[2] を用いる．また，学習データの作成方法については 3 で述べる．属性としては，以下の 3 種類を用いる．

- タームの種類
  そのタームが品詞なし単一語のタームであるか，品詞つき単一語のタームである場合は品詞が何か，フレーズである場合はどんな関係のフレーズかを属性として用いる．
- クエリ中での出現位置
  TREC7 で使用されているクエリには <title>, <desc>, <narr>の 3 種類のタグが付与されている．そこで，タームが最初に出現するタグを属性として用いる．
- クエリ中での重要度
  クエリをインデクシングした際にタームに付与された重要度を属性として用いる．

これらの属性を元に，決定木を用いてタームを使用／不使用のいずれかのクラスに分類する．

5. リランキング

決定木によって有効と判断されたタームのみをクエリとして用い，1000 文書をリランキングする．文書のランキングはベクトル空間モデルに基づく．すなわち，ターム t_j の文書 D_i における重要度を d_{ij}，クエリにおける重要度を q_j としたとき，文書 D_i の重要度は

$$W_{D_i} = \sum_{all\ terms} d_{ij} \times q_j$$

となる．

3 有効なタームの抽出

タームの有効性を判別する決定木を作成するためには，実際に有効なタームを抽出し，学習用データとして与える必要がある．ここではテストコレクションを用いた学習用データの作成方法について述べる．

有効なターム集合の決定は，まず候補となるタームを有効度に従ってランキングし，次にランキングが上位のタームから 1 つづつ順にターム集合に加える．それぞれのターム集合をある基準に従って評価し，有効なターム集合に出現するタームをクラス "使用"，それ以外をクラス "不使用" とする．

よって有効なターム集合を決定するには，タームの有効度とターム集合の評価基準の 2 つを決める必要がある．以下でそれぞれについて説明する．

3.1 タームの有効度

テストコレクションによりクエリと文書集合，そしてどの文書が正解かと言う情報が与えられたとき，クエリに出現するそれぞれのタームに対して有効度を求めることができる．

ターム t が出現する文書数を D_t，そのうち適合する文書の数を T_t，クエリに適合する文書の数を T_{all} とした時，次の値を定義する．

$$\text{タームの精度 } P_{term}(t) = \frac{T_t}{D_t}$$
$$\text{タームの再現率 } R_{term}(t) = \frac{T_t}{T_{all}}$$
$$\text{タームの F-尺度 } F_{term}(t) = \frac{2 \times P_{term}(t) \times R_{term}(t)}{P_{term}(t) + R_{term}(t)}$$

タームの有効度としては精度 $p(t)$ または F-尺度 $F(t)$ を使用することが考えられる．有効度の値が高いタームほど，クエリに加えたときに適合する文書を検索する確率が高いと言える．

3.2 ターム集合の評価基準

次に，これらのタームを "使用"，"不使用" の 2 つのクラスに分類する．分類の方法は以下の 2 通りが考えられる．

タームの有効度	ターム集合の判定基準
F-尺度	F-尺度
F-尺度	精度
精度	F-尺度
精度	精度

表 1: タームのクラスの決定方法

1. まず, タームのランキングで一位のタームのみをクエリとして文書の重要度計算, ランキングを行ない, 平均精度 (non-interporated average precision) を計算する. 次にクエリに二位のタームを加え, 二つのタームにより文書のランキングを行なう. 以下同様にして有効度の高いタームから順にクエリに加え, それぞれのクエリに対する精度を求める.
   そして, 最も良い精度が得られた時のクエリに含まれるタームの集合をクラス "使用" とし, その他を "不使用" とする.
2. 1. と同様にタームのランキングが上位のものから順にクエリに加えていき, それぞれの時点での検索結果の F-尺度を計算する. クエリに出現するいずれかのタームが出現する文書の数を D, そのうち適合する文書の数を T, クエリに適合する文書の数を T_{all} とした時, 検索結果の F-尺度は次のように定義する.

$$\text{検索結果の精度 } P_{result}(Q) = \frac{T}{D}$$
$$\text{検索結果の再現率 } R_{result}(Q) = \frac{T}{T_{all}}$$
$$\text{検索結果の F-尺度 } F_{result}(Q) = \frac{2 \times P_{result}(Q) \times R_{result}(Q)}{P_{result}(Q) + R_{result}(Q)}$$

   そして, 検索結果の F-尺度を最大とするときのクエリに含まれるタームをクラス "使用" とし, その他を "不使用" とする.

よって, タームのクラスの決定方法は表 1 の 4 通りが考えられる. これらの違いに関しては 4.2 節で実験を通して議論する. このようにして得られたタームとクラスの組を C4.5 に与え, タームの使用／不使用を判断する決定木を学習する.

4 実験

3 節で述べた有効なタームの抽出方法, および 2 節で述べたタームの選択を用いたリランキング方法を実験により評価する.

4.1 実験対象

実験対象としては TREC7 のテストコレクションを使用する. TREC7 のテストコレクションはクエリ番号 351～400 の 50 個のクエリと 528,155 の文書から構成される.

また 2 節で述べた通り, 本手法では対象の文書全てに対して検索を行なうのではなく, まず全ての文書に対して従来手法により初期検索を行ない, その結果の上位 1000 文書をリランキング対象としている. 初期検索の手法としては, TREC7 で最も結果の良かった Okapi の手法を用いた.

4.2 有効なタームの抽出方法に関する実験

表 1 で示したとおり, 有効なタームの抽出方法は 4 通り考えられる. ここではそれぞれの手法にどのような違いがあるかを実験により確認する.

まず, 表 2 はそれぞれに対する平均精度の比較である. 表において例えば "Fm-prec" は有効度としてタームの F-尺度を使用し, ターム集合の評価基準として平均精度を用いたことを表す. これを見ると, 有効度としてどちらを使用するかは平均精度の面では大きな影響を与えず, ターム集合の評価基準による影響が大きいことが分かる.

Fm-Fm	Fm-prec	prec-Fm	prec-prec
26.42	31.80	27.26	32.00

表 2: 手法による平均精度の違い

次に, 表 3 は選択されたターム数の比較である. "x/y/z" は選択された品詞なし単一語のターム, 品詞つき単一語のターム, フレーズの数がそれぞれ x,y,z 個であることを表す. また, 下段は選択された品詞なし単一語のタームの数を 1 としたときの他の種類のターム数との比率である. ターム集合の評価に F-尺度を用いた場合の方が選択されるタームの数が少なくなること, 有効度としてタームの F-尺度を用いた場合は選択される選択されるフレーズの数が少なくなることが分かる.

Fm-Fm	Fm-prec	prec-Fm	prec-prec
0.4/0.7/0.4	4.1/4.2/0.8	0.8/1.5/2.0	2.9/5.0/2.5
1.0/1.9/1.0	1.0/1.0/0.2	1.0/1.9/2.5	1.0/1.3/0.6

表 3: 選択されたターム数の比較

さらに, これらを再現率-精度グラフを描いて比較した

ところ次のことが分かった.

- タームの有効度に関しては, F-尺度を用いた場合は再現率が高い領域で, タームの精度を用いた場合は精度の高い領域でそれぞれ有効である.
- ターム集合の評価基準に関しては, 平均精度を用いた場合は再現率の高い領域で, F-尺度を用いた場合は精度の高い領域でそれぞれ有効である.

4.3 リランキングに関する実験

実際に決定木を学習し, それによって選択されたタームを用いて初期検索結果の文書をリランキングする実験を行なった. 学習データの作成方法としては, タームの有効度として精度を, ターム集合の評価基準として平均精度を用いた. また, 学習データ作成の際には "使用" クラスのデータを複製し数を増やすことにより, 多少不要なタームを選択してしまうことはあるが必要なタームを落しにくいルールを学習させた. これは, 普通に学習を行なって得られたルールでは選択されたターム数の平均が "0.66/1.44/1.66" であることからも分かる通り選択条件が厳しく, 例えば タイトルが "orphan drug" であるクエリで選択されたタームが "orphan" のみで "drug" が落されてしまったりと, 非常に重要であるタームでも "不使用" クラスに分類されてしまうことが多かったためである. 複製数としては 5 倍, 7 倍, 10 倍の 3 通りについて実験した. 実験は 50 クエリのうち 49 クエリで学習, 残りの 1 クエリでテスト, という方法で全てのクエリに対して行なった.

表 4 がその結果である. "改善数" の行は, リランキングによって初期検索よりも精度が向上するクエリの数である. 学習データと同じ, すなわち理想的にタームが選択できたとき, 50 クエリのうち 31 クエリは本手法により精度が改善する可能性がある. 結果を見ると, "使用" クラスの学習データを 7 倍に増やして学習を行ない, 得られたルールでタームを選択した場合に最も良い結果が得られ, このときこの 31 クエリのうち 25 クエリについて, 本手法によるリランキングを行なうことで精度が改善されていることが分かる. 残りの 19 クエリについては Okapi の手法で用いている検索要求拡張が非常に有効であるため, 元のクエリのみを用いている本手法でリランキングを行なうと却って結果が悪化してしまう. 実際に表 4 の "平均精度 (改悪)" の行を見ると, 本手法で改悪されてしまうクエリでは初期検索の精度が非常に高いことが分かる.

	初期検索	学習データ	5 倍	7 倍	10 倍
改善数	—	31	20	25	21
平均精度 (改善)	24.27	32.27	25.67	26.74	26.36
平均精度 (改悪)	40.20	31.56	23.42	26.87	24.99
平均精度 (全て)	30.33	32.00	24.82	26.79	25.84

表 4: リランキングの評価

5 まとめ

本研究では, 検索要求に含まれるキーワードやフレーズの中から有効と思われるタームを選択し, そのようなタームを用いて文書をランキングする手法を提案した.

本手法では, 検索要求と文書集合, そしてどの文書が正解かと言う情報が与えられたときに, タームの有効度という尺度を設定し, これに基づいて各検索要求に対して有効なターム集合を選択する.

次に, 選択されたターム集合をもとに学習を行ない任意の検索要求に対して有効なタームを選択するための規則を得る. 学習, 規則使用の際にはタームの種類, クエリ中での出現位置, クエリ中での重要度を選択の手がかりとした.

TREC7 で使用された 50 個のクエリを用いて評価実験を行なった結果, 本手法で精度の向上が可能である 31 クエリのうち 25 クエリに対してはほぼ適切にタームを選択し, それを使用したリランキングによって初期検索以上の精度を得ることができた.

参考文献

[1] M. Mitra, C. Buckley, and C. Cardie. An analysis of statistical and systactic phrases. In *Proceedings of RIAO'97 Conference on Computer-Assisted Information Searching on the Internet*, pp. 200–214. C.I.C and C.A.S.I.S,USA, 1997.

[2] J.R. QuinLan. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann, 1993.

[3] E. Robertson, S, S. Walker, and M. Beaulieu. Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive. In *Proceedings of Seventh Text REtrieval Conference (TREC-7)*, pp. 152–163, 1999. NIST Special Publication.

[4] S. Sekine and R. Grishman. A corpus-based probabilistic grammar with only two non-terminals. In *Proceedings of the International Workshop on Parsing Technologies*, 1995.