E-015

係り受け関係を用いた共起語による適合フィードバック
Relevance Feedback Using Co-occurrence Words Based on Dependency Relations

菅澤 匡倫 † **延澤 志保** ‡ **太原 育夫** §
Masanori Sugasawa Shiho Nobesawa Ikuo Tahara

1. はじめに

適合フィードバックを用いた検索支援システムでは, Web から検索した文書群に対して, 利用者がその文書群に対して上位の出力結果に適合性の評価を行ない, その評価がフィードバックされる [1]. 龍 [2] は, 評価を多値化しランク化して, 各ランクごとに単語の出現頻度に対して重みをつけ, その重みを用いてランクの高い文書群に対しては高い評価がつくような評価値を与えることにより, 評価値の高い順に表示を行なっている. 本稿では, 坂井 [3] の手法に基づき, ランク評価に係り受け関係に着目した共起語を適用する手法を提案する.

2. 適合フィードバックによる検索支援

2.1 ランクの分類

検索エンジンで固有名詞をキーワードとして検索した場合の文書の特徴を考慮して, 以下のような 4 段階のランク分類による評価を行う.

A 利用者の要求を十分に満たしている情報である.

B 利用者の要求の一部を満たしているが不十分な情報である.

C 検索語が要求している意味で用いられているが, 利用者の要求とは違う内容の情報である.

D 検索語が要求している情報とは異なった意味で使われている情報である.

2.2 文書の評価

文書に対する利用者の評価が高いほど大きな値が与えられるような評価値を設定する. そのため, 利用者によって行われたランク付けに基づき, 各ランク中の語への重み付けを行う. tfidf の考え方を拡張し, 単語の重みを評価する方法は数多く研究されている. そのうちで, 文書に付与されているあらかじめ分類したカテゴリの情報と単語の関係を考慮した重みを評価する方法が提案されている [2][3]. すなわち情報が付与された文書には, そのランクごとに特徴のある語が出現していると考えられることから, フィードバックに用いた文書中のランク r の文書数を N_r, その中で語 t の出現していた文書数を n_r^t として, 単語の重み w_t をランク情報と語の関係を反映した式 (1) で算出する.

$$w_t = \sum_r \alpha_r \log(\frac{n_r^t}{N_r} + 1) \quad (1)$$

システムはフィードバックされた文書の語とランクの情報を保持し, ランク情報が新たにフィードバックされたびに語の重み w_t を更新する. したがって文書のフィードバック件数が多くなるにつれて, より適切な重みを持つことになる. α_r はランク毎に異なる係数であり, ランク A，B の場合は正，C，D の場合は負として, ランク A，B，C，D を 1，0.5，-0.5，-1 と設定している. ただし, フィードバックする文書中に C と D が同時に存在し, かつ C と D のみしか出現しない場合は，C に出現する語に負の重みを与えると A，B の文章に出現する重みが下がってしまうため, 坂井 [3] では C，D を 0.5，-1 のように設定している.

評価の高いランクと関連の強い語であるほど重み w_t は大きな値を持つ. この重み w_t はフィードバックされた文章に出現する各語に対して付与されるので, 大きな重みを持つ語が多く出現するほどその文書の評価は高いと考えられる. そこで w_t を用いて文書 d の評価値 $score(d)$ を式 (2) を用いて算出する.

$$score(d) = \sum_{t \in d} w_t \quad (2)$$

3. 共起語に着目した文書評価

3.1 係り受け関係

類似文書の場合, その類似文書中の検索語に対して係り受け関係にある単語が出ている場合が多い. つまり, 利用者が要求している文書と同様に文書内で検索語に対して係り受け関係にある単語があるならば, その文書は類似しているといえる.

3.2 共起語

共起語を定義する際, ランク分類において分類した文書の内ランク A とランク C に着目する. 先行研究の実験では, ランク A とランク C の文書が混在している場合があった. そこで共起語を用いて, ランク A を上位にランク C を下位にすることにより, その混在化が解消されると期待できる. ランク分類を利用者が行ない, ランク A が全体で 1 つしかなかった場合, 検索語に対して係り受け関係にある単語を抜き出し, その単語の中で回数が多かったものを共起語とする.

また, ランク分類を利用者が行ない, ランク A が全体で複数あった場合, その文書ごとの係り受け関係にある単に着目し共通する単語を共起語とする. すなわち, 文書群中に係り受け関係にある単語が半数以上あったときそれを共起語とする. つまり, 同ランクの文書群の数を N とし, その中で係り受け関係にある単語 t が文書 N 群の中に含まれる回数を n_t としたとき, $n_t = N/2$ が成り立つならば, その単語 t を共起語とする. ランク C についてもランク A と同様に定める.

3.3 文書評価値

共起語が含まれた文書群について新たな重みを加算することにする. 加算する重みを決めるために, 坂井 [3] が

†東京理科大学大学院理工学部情報科学専攻
‡武蔵工業大学知識工学部情報科学科
§東京理科大学理工学部情報科学科

用いた評価値に基づきさらに新しい評価値を定める．まず, 文書群の評価値の平均値 h を式 (3) で算出する．

$$h = \frac{\sum Score(d)}{N} \quad (3)$$

その平均値を用いて新たに文書評価値 $Score(d)$ に対して新たな評価値を算出するために文書群の中で評価値の最高値を $Score(T)$, 最低値を $Score(B)$ として, ランク A の共起語に対する値 W_A, ランク C の共起語に対する値 W_C を以下のように求める．

$$W_A = \frac{Score(T) - h}{2} \quad (4)$$

$$W_C = \frac{h - Score(B)}{2} \quad (5)$$

共起語が含まれる文書 d に対して新たな評価値 $Score(d)$ を以下のようにする．

ランク A の共起語が含まれた場合

$$Score(d) = Score(d) + W_A \quad (6)$$

ランク C の共起語が含まれた場合

$$Score(d) = Score(d) - W_C \quad (7)$$

4. 実験結果

4.1 評価方法 1

A ランクから順に並ぶ状態を理想としてどのぐらい達成できているかを評価する値を評価値 1 とする．ここで理想状態とは, 例えば A が 2 個，B が 3 個，C が 3 個，D が 2 個の文書群の数が 10 個のとき, 出力結果が AABBBCCCDD といった順になっている状態を言う．そして各ランクごとに理想状態と同じか否かを測り，ランク数で割って平均を求めて評価値 1 を得る．

4.2 評価方法 2

理想状態と現在の各ランクの状態との距離を評価する値を評価値 2 とする．理想状態を 0 として 0 に近いほど評価が高い．理想状態と比べたときにランク A が理想状態の A ランクの範囲にある場合は 0 として，B ランクの範囲のときは 1，C のときは 2 といった感じに距離によって数値が上がり測定値が加算されていく．同様にそのほかのランクもそのランクから離れているランクの範囲にあるものほど数値が大きくなる．

4.3 先行研究と提案手法の実験結果

「カレー」を検索語とした場合, 評価値 1 においては坂井 [3] のほうがよい結果が得られていたが, 評価値 2 においては本手法のほうがよい結果を得られた．これは, カレーのおいしい店という情報を要求しているので, カレーのおいしい店の情報が載っている文書においてそのページの共起語がランク B の文書においても含まれてしまったためだと考えられる．これによりランク A だけでなくランク B まであがってしまい, その結果, 評価値 1 においては坂井 [3] よりも下がってしまったが, 理想状態に対する距離という点においては, ランク A と B がともに上がりランク C が下がったために結果は坂井 [3] よりもよい結果となった．

「The 有頂天ホテル」を検索語とした場合については, 坂井 [3] のほうが評価値 1 と 2 ともによい結果となっている．今回のランク付けにおいて DVD の販売のページを C としたが, そのページにも映画の感想が書かれている場合が多少あり, さらにランク A の共起語が現れたこと, ランク B においても先ほどのカレーの検索結果同様に共起語が現れたことにより, 評価値 1 と 2 ともに坂井 [3] よりも悪い結果になったと考えられる．

「ダイエット」と「ワイン」それぞれを検索語とした場合は, 評価値 1 と評価値 2 ともに坂井 [3] よりよい結果となった．これはランク A と C の共起語が検索結果の中のランク A と C において見られたためであり, 共起語に着目する本手法の有効性を示している．

表 1: 先行研究の実験結果

検索語	要求情報	評価値 1	評価値 2
カレー	おいしい店	0.345	64
The 有頂天ホテル	映画の感想	0.461	60
ダイエット	運動を用いる	0.336	70
ワイン	ワインの知識	0.292	52

表 2: 提案手法の実験結果

検索語	要求情報	評価値 1	評価値 2
カレー	おいしい店	0.326	56
The 有頂天ホテル	映画の感想	0.391	70
ダイエット	運動を用いる	0.456	58
ワイン	ワインの知識	0.356	51

4.4 おわりに

本稿では, 適合フィードバックに用いるランク評価に係り受け関係に着目した共起語を考慮する方法を提案した．実験結果より, 共起語の利用が有効であることを確認した．

参考文献

[1] 徳永健伸, 情報検索と言語処理, 東京大学出版会, 1999.

[2] 龍広海, 延澤志保, 太原育夫,“ ランク情報のフィードバックによる文書検索支援システム,” 信学会 2005 年総合大学講演論文集, 情報・システム 1,41, 2005.

[3] 坂井隆文,“ 適合フィードバックを用いた検索支援システム,” 平成 17 年度東京理科大学卒業論文, 2006.