格フレームを用いた情報検索

宮川 和　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

1 はじめに

従来の情報検索システムでは, 検索対象となる文書と検索要求とを重み付きのキーワードベクトルで表現し, ベクトル間の距離によって関連性を判定するベクトル空間モデルを用いることが多い. 通常, ベクトルの各要素の重みの計算には語の出現回数を基にした統計量を用いるが, このような表層的な情報のみを利用した手法では, 検索要求に対して関連性が低い文書も検索してしまうという問題が生ずる. このような従来の手法に対し, TREC と呼ばれるプロジェクトでは, 大規模なデータにおいて自然言語処理を用いた場合, その効果がはっきりと現れる場合があると報告している [4].

我々は自然言語処理を用いた情報検索の一手法として, 従来の手法によって検索された文書集合に対し, 文書中の動詞, 格要素 (名詞, 助詞) からなる格フレームを用いて不適切な文書を排除する手法を提案する. 係受け関係の解析により抽出される格フレームを利用することで, 文書の構造的な情報を扱うことが可能となり, キーワードのみの集合に比べて精密に文書の内容を把握することができる. 抽出された格フレームと検索要求との関連性を判定し, 関連性の高い格フレームを持つ文書を検索することで, 内容的に関連性の低い文書が排除できる.

2 特徴的な格フレームの抽出

本論文では格フレームを, 動詞とその動詞に係る格要素 (名詞, 助詞) として定義する (図1参照).

全焼する	焼死する
で : 火災	が : 客数人
を : 数平方 m	で : 火災
	に : 12月14日

図 1: 格フレーム例

格フレームは, キーワード検索などに代表される従来の検索手法によって収集された文書集合中から, 係受け解析を行なうことによって抽出する. 本論文ではこれら抽出された格フレームに対して, 検索要求との関連性を判定する重要度と呼ばれる値を導入する. 重要度とは, その格フレームが検索要求に対してどの程度特徴的であるかを示す尺度である. 重要度を導入することにより, 検索要求に対して特徴的な格フレームを推定し, 文書と検索要求との関連性の判定に利用する. この節では, 格フレームの重要度の計算方法について述べる.

2.1 動詞, 格要素の重要度計算

格フレームに対して直接重要度を求めることは, データスパースネスの問題から難しい. そこで動詞, 格要素それぞれに対して重要度を計算し, それらを用いることで格フレーム重要度を計算する. 動詞, 格要素の重要度は TF, IDF を利用して計算する. 例えば動詞 V_iの重要度 I_{v_i}は式 (1) によって求める.

$$I_{v_i} = (f_i/f_{all}) \times (S_{all}/S_i) \quad (1)$$

ここで, f_iは動詞 V_iの出現頻度を, f_{all}は全動詞の出現頻度を示す. また, S_iはサンプル記事集合における動詞 V_iの出現記事数を, S_{all}は総サンプル記事数を示す. サンプル記事とはランダムに集めた記事を指す [8]. サンプル記事を用いた場合, 例えば動詞における, 「ある」, 「なる」などの一般的に使用されやすい語は S_iの値が大きくなるため重要度は低くなる. そのためより検索要求に対して特徴的な語を判定することが可能になる.

2.2 格フレームの重要度計算

格フレームの重要度 I_{cf}は式 (2) によって計算される.

$$I_{cf} = (\alpha \times I_v + (1-\alpha) \times \sum_i I_{ni}) \times a_{cf} \quad (2)$$

ここで, I_vは動詞重要度を, I_{ni}は, 格要素 n_iの重要度を示す. これらは 2.1 節で求めたものを正規化して使用する. $\alpha(0 \leq \alpha \leq 1)$ は格フレーム重要度に対する動詞重要度の貢献度を示しており, 動詞, 格要素の

どちらの重要度を格フレーム重要度に対して重要視するかを指定する．また，a_{cf}は格フレームの出現頻度を示す．式 (2) を用いることで，特徴的な動詞，格要素を持ち，出現回数が多い格フレームほど検索要求に対して特徴的であると判定される．

2.3 格フレームの重要度の修正

格フレームは格要素が多いほど話題を限定するため，話題に対してより特徴的になる．しかし一般的に，格要素数が多い格フレームほど出現回数が少なくなるため，式 (2) を用いると重要度は下がる．これは話題に対する特徴性と重要度の定義に矛盾する．そこで格フレーム間の類似関係や包含関係を利用してこの問題を回避する．すなわち，類似，包含関係にある格フレーム間で相互に重要度を修正し，多用される表現の格フレーム重要度を高くすることで，検索要求に対して特徴的な格フレームの判定を容易にする．

類似，包含関係にある格フレームは，動詞，格要素がどれだけ類似しているか，また類似した格要素をどれだけ共通に持つか (あるいは包含するか) によって定義できる．そこで本論文では，動詞，格要素がどれだけ類似しているかを評価する尺度として以下のように類似度を定義する．類似度を計算することにより格フレーム間の類似性を評価し，類似性に応じて相互の重要度を修正し，特徴的な格フレームを明らかにする．

2.3.1 格要素の類似度

格要素の類似度は，シソーラスを利用することで計算する．シソーラスとは，分類コードを用い，語を意味によって分類，配列した辞書のことである．ある格における名詞n_{ai}とn_{bi}を，シソーラスを用いて分類コードに変換し抽象化する．両者が共通の分類コードに変換される場合，両者は同じ概念に属するため類似していると判断する．また，共通する分類コードの割合が高いほど類似した名詞であると判断する．そこで式 (3) を用いて格要素の類似度S_{ni}を計算する．

$$S_{ni} = \frac{(\text{共通する分類コード数})^2}{(n_{ai}\text{の分類コード数} \times n_{bi}\text{の分類コード数})} \quad (3)$$

類似しない名詞間，あるいは一方の格要素がない場合の類似度は 0 となる．例えば，ある格において名詞がそれぞれ図 2 のように分類コードに変換された場合，2 つの名詞の類似度は $3^2/(4 \times 3) = 3/4$ となる．

地震後 (2646,2698,2702,2711)
直後 (2646,2698,2711)

図 2: 名詞の分類コード変換例

2.3.2 動詞の類似度

動詞の類似度を計算する場合もシソーラスを利用する．しかし，動詞v_a, v_bに共通の分類コードがあったとしても，動詞単独からでは両者が本当に共通の意味で使用されているかどうかは判断できない．例えば「発生する」，「起きる」という 2 つの動詞があった場合，後者の「起きる」は「目覚める」という意味で使用されているかもしれない．このような動詞の曖昧性の解消には，動詞が取る格要素を利用する [6]．動詞に共通の分類コードがある場合，両者の格フレームの各格要素がどれだけ似ているかによって，動詞がどれだけ似ているかを判定できる．そこで本論文では動詞の類似度として，共通する格要素の類似度の平均値をとる．例えば図 3 において，動詞の類似度は $(3/4 + 1/2 + 1)/3 = 3/4$ となる．また動詞間に共通の分類コードがない場合は，異なった意味であると判断し，類似度は 0 とする．

発生する 起こる		地震後に 直後に	数箇所で 住宅数棟で	火災が 火災が
3/4	←	3/4	1/2	1

図 3: 動詞の類似度の計算例

2.3.3 類似度を用いた格フレーム重要度の修正

以上のように計算した動詞，格要素の類似度を用いて，格フレーム間相互の重要度を修正する．格フレーム A に対する格フレーム B の修正値$M_{(a|b)}$は以下の手順で求める．まず格フレーム A,B 間の動詞，格要素類似度を計算する．動詞類似度が 0 の場合，異なった格フレームであるので$M_{(a|b)} = 0$とする．動詞類似度が 0 でない場合，$M_{(a|b)}$を式 (4) によって計算する．

$$M_{(a|b)} = (\alpha \times I_{v_b} \times S_v + (1-\alpha) \times \sum_i (I_{n_{bi}} \times S_{ni})) \times a_{cf_b} \quad (4)$$

ここでI_{v_b}は格フレーム B における動詞v_bの重要度を，S_vは格フレーム A,B 間の動詞類似度を示す．同様に，$I_{n_{bi}}$は名詞n_{bi}の重要度を，S_{ni}は名詞類似度を示す．また，a_{cf_b}は格フレーム B の出現頻度を，αは式

(2) における動詞重要度の貢献度を示す.

式 (4) は, 格フレーム B の重要度を求める式に, 動詞, 格要素の類似度を掛け合わせたものである. このため, 格フレーム A,B 相互の類似性が高いほど, 格フレーム A に対する修正値は高くなる. また, 共通しない格要素に対しては類似度が 0 であるため, 修正値は, 動詞, 格要素の類似性と共通する格要素のみを用いて計算されることになる.

最終的な格フレーム A の重要度 I_{cf_a} は, すべての格フレーム間の組合せに対して修正値を求めた後, 式 (5) によって求める.

$$I_{cf_a} = I_{cf_a} + \sum_i M_{(a|i)} \tag{5}$$

3 格フレームを用いた文書評価

重要度を計算することで, 重要な格フレームを推定することが可能になる. 重要度が高い格フレームは検索要求に対して特徴的であり, 内容的に関連性が高いと判断できる. 従ってこのような格フレームを文書中に持つ文書は, 検索要求に対して関連性が高いと推定できる. つまり, 重要度の高い格フレームがどの程度文書中に含まれているかによって文書と検索要求との関連性を評価できる.

文書の評価を行なう前に, 関連性が低い格フレームを削除することで評価の誤差を少なくする. 具体的には, 全重要度の合計β%以下の格フレームを下位から削除する. 重要度の合計値を利用することで, はっきりと特徴的な格フレームを推定できた場合は少ない格フレームで, そうでない場合は多くの格フレームを用いて関連性を判定する.

文書の評価は, 上位の格フレーム集合を各文書に適用し, 適用した格フレーム重要度の合計を文書の重要度とすることで行なう. しかし, 文書によって抽出される格フレーム数が異なり, 適用される格フレーム数も異なる. そこで文書 a の重要度 I_a を式 (6) を用いて計算する.

$$I_a = \frac{\text{適用された格フレーム重要度の合計}}{\text{文書 } a \text{ から抽出される格フレームの総数}} \tag{6}$$

式 (6) を用いることで, 各文書の格フレーム数に影響されることなく評価が可能になる. 文書の重要度を求め, 重要度によって文書をランキングすることで, 検索要求に対して関連性の高い文書を推定する.

4 実験

実験には, 毎日新聞 94 年度版 CD-ROM データの中からランダムに収集し人手で解析が行なわれた, 5080 記事を用いる [7]. この記事集合に対しては, 用意された Query に対して, 関連性がある (A Rank), 関連性が曖昧 (B Rank), 関連性なし (C Rank), というランクが人手で付与されている. 今回の実験では 2 節で述べた Query による初期検索にキーワード検索を利用している. そのため, キーワード検索によって 1 記事以上の検索が可能な, 30 Query を用いて実験を行なう. システムの評価は, 各 Query に対して A Rank, B Rank の文書がどの程度上位にランキングされるかを再現率-適合率曲線上の 21point で求め, それらの平均値を評価することで行なう. また従来法として, ベクトル空間モデルを用いた SMART システム [3] と同等のランキング法を利用し, 本手法との比較を行なう.

本手法ではいくつかのパラメータと資源を利用する. 初期検索の対象文書は, 毎日新聞 91 年度から 95 年度までの CD-ROM データ約 43 万記事とした. またいくつかの予備実験から, 2.2, 2.3.3 節で述べたαを 0.3, 3 節で述べたβを 90 とした. 2.3.1 節で述べた名詞用のシソーラスは NTT シソーラス [1, 2] を, 2.3.2 節で述べた動詞用のシソーラスは分類語彙表 [5] の上位 6 桁を利用した.

本手法により各 Query に対して検索された文書数, 抽出された格フレーム数, 重要度を求めることにより下位を削除した場合の格フレーム数の平均値を表 1 に示す. また, 30 Query を用いた実験結果を図 4 に示す.

表 1: 各 Query 毎の文書, 格フレーム抽出結果

	平均値
検索された文書数	1,206
抽出された格フレーム数	30,676
下位を削除した格フレーム数	11,005

図 4 は従来法と本手法がほぼ同等であることを示している. しかし本手法の場合, 初期検索した文書を対象とし, その文書集合中で特徴的な格フレームを抽出し評価を行なうため, 初期検索でうまく話題を限定できなかった場合によい結果が得られない. そこで話題が限定されている場合とそうでない場合の結果を比較した. 話題が限定されているかどうかの判定には, 表 1 で示した平均検索文書数を用いた. 平均文書数以下の文書を収集する Query は, ある程度話題を限定していると判断し, それ以上の文書を収集する Query は話題が限定されていないと判断した. その場合の結果を図 5, 図 6 に示す.

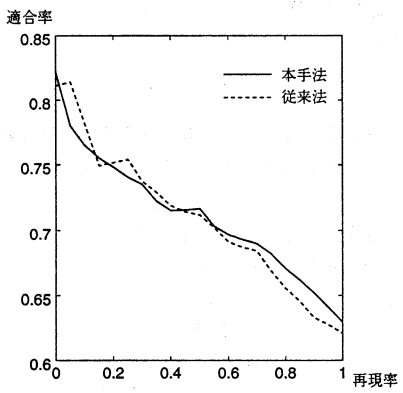

図 4: 30 Query の平均再現率-適合率曲線

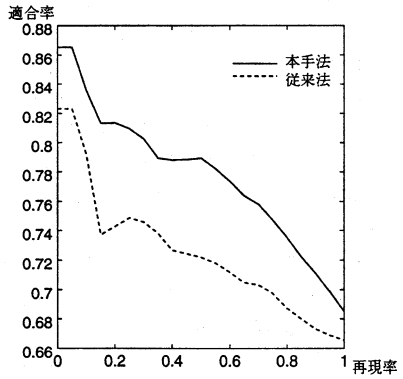

図 5: 検索記事数が平均値以下の再現率-適合率曲線

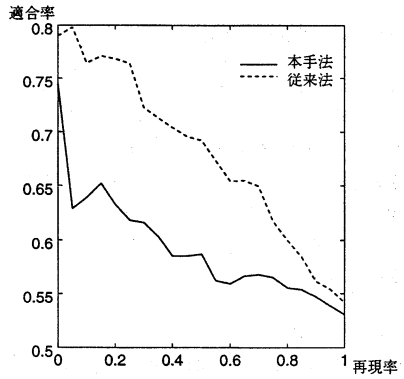

図 6: 検索記事数が平均値以上の再現率-適合率曲線

図 5 に明らかなように, 収集文書が平均以下, つまりある程度初期検索で話題が限定できる場合, 従来法に比べ最大 8%ほどの改善が見られた. しかし話題が限定できない場合, 本手法は図 6 に示すように従来法と比べ結果が悪くなった. これは本手法が, 初期検索によってある程度話題を限定した文書集合に対して, 従来法に比べ検索結果を改善できることを示している.

5 おわりに

本論文では, 言語処理を用いた情報フィルタリングの手法の一つとして, 格フレームを用いて記事と検索要求との関連性を判定し, 関連性の高い文書を検索する手法を提案した. 実験では, 初期検索によってある程度話題が限定された場合, 本手法は従来の手法に比べ結果を改善できることを確認した.

実験では初期検索としてキーワード検索を用いたが, その他の検索手法の結果に対して本手法を適用し, どのように結果が改善されるかを確認する必要がある. また本論文では動詞, 名詞, 助詞からなる格フレームのみに注目したが, 形容詞, 形容動詞などにも注目することで, より文書の内容に踏み込んだ解析が可能になると思われる. そのような手法を考案することが今後の課題としてあげられる.

参考文献

[1] 池原悟, 宮崎正弘, 横尾昭男. 日英機械翻訳のための意味解析用の知識とその分解能. 情報処理学会論文誌, Vol. 34, No. 8, pp. 1692–1704, 1993.

[2] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩己, 小倉健太郎, 大山芳史, 林良彦. 日本語語彙体系 — 全 5 巻 —. 岩波書店, 1997.

[3] Gerard Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Prentice-Hall, 1971.

[4] T. Strzalkowski, F. Lin, J. Perez-Carvallo, and J. Wang. Building effective queries in natural language information retrieval. In *Proc. of the fifth Applied NLP*, pp. 299–306, 1997.

[5] 国立国語研究所（編）. 分類語彙表. 秀英出版, 増補版, 1996.

[6] 内元清貴, 宇津呂武仁, 長尾眞. 動詞の語彙的知識獲得における類義語の用例を用いた多義性の類別. 情報処理学会 自然言語処理研究会, Vol. 101, No. 15, pp. 105–112, 1994.

[7] 木谷ほか. 日本語情報検索システム評価用テストコレクション. BMIR-J2, 情処研究会報告 DBS-114, 1998.

[8] 吉田和広, 徳永健伸, 田中穂積. 新聞記事の要約のためのテンプレートの自動抽出. 言語処理学会第 2 回年次大会, pp. 105–108, 1996.