コンプリメントタームを用いた情報検索

松本 晃　田中 穂積　徳永 健伸
東京工業大学大学院 情報理工学研究科

1 序論

ベクトル空間モデル [1] に代表される従来の情報検索システムでは二つの文書の類似度を計算する際に，文書に含まれるタームの共通部分に着目していた．通常の情報検索ではこのアプローチが有効であるが，あらかじめ内容をしぼりこんだ文書群に対してその中で更に同じ内容を表す文書をまとめる処理を行う場合には，不都合が生じる．同じような話題を扱った文書からなる文書群を対象にした場合，文書群の中の各文書はもともとある程度の話題を共有しているために，タームの共通集合はどの文書ペアにおいても似たようなものになってしまい，共通部分のタームの中にある2つの文書の違いをあらわすタームが「埋もれて」しまう可能性がある．

本研究では，比較する文書に対して，「一方には含まれるが，他方には含まれないターム」に着目する手法を提案する．このようなタームを「コンプリメントターム」と呼ぶ．重要度の高いコンプリメントタームが存在するなら，それは二つの文書が中心の話題で一致していないことを示す．共通部分でないところに重要度の高いタームが存在すれば，両文書は重要な話題の少なくとも一つを共有していないことになる．比較する文書が似ている文書であればあるほど，この差は，両文書の内容の決定的な違いを意味するようになると考えられる．

2 コンプリメントタームの利用

コンプリメントタームとは，二つの文書の比較に際し，一方には含まれるが他方には含まれないタームのこと指す．どちらに含まれてどちらに含まれないのかを明確に呼び表すために，コンプリメントタームを含まない記事を「基準記事」，含む記事を「対象記事」と呼び，基準記事 A，対象記事 B に対して計算されるコンプリメントタームを「基準記事 A の対象記事 B に対するコンプリメントターム」と呼ぶ．

2.1 重みづけ

コンプリメントタームにはそれぞれ重みを設定することができる．全コンプリメントタームの重みを合計したものが，基準記事から対象記事への類似度を計るための値となる．タームに対する重みづけは，[2] に詳しく述べられている．

本手法では，タームに対する重みづけに際して次のような点を重視する：

- コンプリメントタームに与える重みは，基準記事と対象記事の類似を表すためではなく，両者の相違を表すために付けられなければならない
- コンプリメントタームの部分，すなわち**基準記事に含まれない部分にその基準記事以外の記事には良く含まれているタームが存在**すれば，その事実は基準記事と対象記事が記事群中における重要な話題について相違があることを意味する

具体的には，式 (1) で計算する：

w_t ：ターム t の重要度
N ：文書群に含まれる文書数
f_t ：文書群におけるターム t の出現数
df_t ：文書群でターム t が現れる文書数

$$w_t = \frac{f_t}{\ln(N/df_t)} \qquad (1)$$

式 (1) では，df_t の値が N に近づくほど，w_t の値が大きくなる．これによって，文書群中の多くの文書に含まれるタームがコンプリメントタームの集合に存在した場合には，高い重要度が与えられることになる．

コンプリメントタームの集合に含まれるタームに付与された重みをすべて足し合わせることで，コンプリメントタームの集合の重みを計算する．「基準記事 A の対象記事 B に対するコンプリメントタームの集合の重み」を式 (2) に示す：

w_t ：ターム t の重要度
$CT(A,B)$ ：基準記事 A の対象記事 B に対するコンプリメントタームの集合

$WCT(A,B)$ ：$CT(A,B)$ の重み

$$WCT(A,B) = \sum_{t \in CT(A,B)} w_t \qquad (2)$$

2.2 記事のランキング

2.1 節で示されたコンプリメントタームの集合の重みの値は，大まかに言うと記事 A と記事 B の相違の度合いを表す「相違度」のようなものである．この値を使って，**文書のランキング**を行う．

ここで言う文書のランキングとは，文書群中のある記事 X について，記事群中のすべての記事を，記事 X に類似していると思われる順に上位からランキングすることである．このとき，記事 X はランキングをする上での検索要求であることから，記事 X のようなランキングの基準となる記事のことを **Query** と呼ぶ．

記事のランキングは，基準記事を Query にして，コンプリメントタームの集合の重みが小さくなる順に対象記事をランキングすることで行う．

2.3 正規化

2.2 節で説明したようにして，記事をランキングすることが出来る．ところが，この手法には問題が存在する．対象記事がタームを多く含む時は，基準記事が何であるかに関わらず，コンプリメントタームの集合が大きくなる．このような記事は、全体的に低い順位にランキングされやすい．

しかし，単純にコンプリメントタームの種類数で正規化を行うと，コンプリメントタームの種類数が多い対象記事のコンプリメントタームの重みが小さくなりすぎてしまい，うまくゆかない．そこで，本手法では「ランキングの反転による正規化」を行う．

図 1 に，ランキングの反転の様子を示した．図の上段の表は，記事 1 ～ 6 について，各記事を Query としたランキングの順位の値を示したものである．上段の表を横に読むと，基準記事を Query とした反転前のランキングの順位を読み取ることができる．上段の表を縦に読むと，対象記事が各基準記事において何位にランキングされているかが読み取れる．これが，反転に使用するランキング順位の値になる．上段の表を縦に読んだときの対象記事を Query として，順位の値によって，基準記事をランキングする．

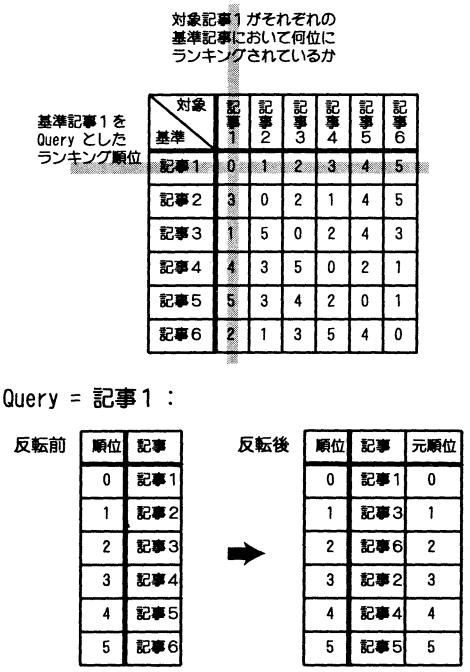

図 1: ランキングの反転

下段には，反転前と反転後のランキングを示した．なお，下段の表の反転後のランキングにおいて，反転前の元順位が全て異なっているのは，あくまで説明を分かりやすくするためである．上段の他の対象記事の列を見ると分かるように，実際には同順位の記事が多く存在し，この場合には，さきに説明したように，コンプリメントタームの集合の重みが小さい順にランキングする．

3 実験

3.1 実験の概要

実験には，毎日新聞 CD-ROM 91 年～ 95 年の記事データを利用した．この中には，約 43 万の記事が収録されており，それぞれの記事には，半自動的に抽出されたキーワードが付けられている．本実験でも，このキーワードを記事に含まれるタームとして利用する．この記事データの中から，いくつかのキーワードで検索を行って得られた記事群に対し，同じ話題を持つと思われる記事に手作業で印を付けた上で実験を行い，評価を行う．

実験に使用する記事群の情報をまとめて，表 1 に示した．なお，記事群 A のみ，2 種類の印がつけら

表 1: 使用する記事群の情報

	記事群名	検索された記事数	同じ内容の記事数	ターム共有率
A	「日本テレビ爆弾事件」・「都庁爆弾事件」	37	13・24	0.145
B	「日本航空・中華航空の墜落事故」	88	5	0.130
C	「北海道の地震」	157	50	0.099
D	「94 年・東京の殺人事件」	96	11	0.095
E	「アメリカ核実験」	69	11	0.069

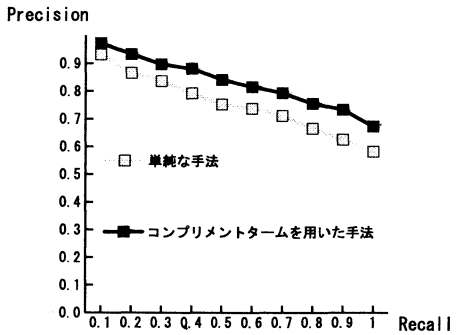

図 2: 再現率-精度曲線の例（記事群 A）

れているが，これは，記事群 A のみは記事群中の 2 種類の話題を利用したことを意味する．

ターム共有率とは，記事群の中の各記事に含まれるタームがどの程度共通しているかということを示すための指標で，式 (3) で計算される：

$$T(A,B) = (\text{記事 } AB \text{ の共通集合})$$
$$RT(A,B) = \frac{|T(A,B)|}{|A|}$$

$$(\text{ターム共有率}) = \frac{\sum_{X \neq Y} RT(X,Y)}{(N-1)^2} \qquad (3)$$

ターム共有率の大きい記事群ほど，互いの記事がより多くのタームを共有していることになる．

評価は，本手法と従来手法（ベクトル空間モデルに基づいた類似度計算）の 2 つの手法で記事群ごとに再現率-精度曲線を描くことで行う．

3.2 結果

再現率-精度曲線の結果全ては紙面の関係から掲載できないが，その一例を図 2 に示す．

表 2: 各記事群における，「コンプリメントタームを用いた手法」の精度と「単純な手法」の精度の差

記事群	A	B	C	D	E
共有率	0.145	0.130	0.099	0.095	0.069
精度差	0.0795	0.1838	0.0422	-0.0127	-0.0214

それぞれの記事群のグラフで，各再現率における，「コンプリメントタームを用いた手法」の精度と，「従来手法」の精度との差の平均値を，表 2 に示す．

この表から，ターム共有率の高い記事群，すなわち内容の良く似ている記事群ほど，「コンプリメントタームを用いた手法」が優れていることが分かる．

本手法がターム共有率の高い記事群に対しては有効だが，そうでない記事群に対しては「単純な手法」に劣ってしまう理由は，次の点にあると考えられる：

> 本来「コンプリメントターム」は，『**共通部分に重要なタームが多く含まれていても，非共通部分に重要なタームが含まれているならば**，両者は重要な話題で食い違っていると判断できる』という考え方に基づいている

従って，2 つの文書がそもそも無関係であり，共通部分にほとんど重要なタームが含まれないような場合には，非共通部分に重要なタームがあるかどうかは，2 つの文書間の類似性の判定には重要ではなくなってしまう．このようなケースでは，逆に従来手法などによって，2 つの文書の共通部分を考慮したほうが，うまく行くと思われる．

以上のことから，本手法はターム共有率の高い文書においては有効であるが，ターム共有率の低い文書に対してはあまり有効ではなく，従来手法などによってある程度類似する記事をしぼりこんだ後に本手法を適用するといった利用法が有効だと考えられる．

4 記事のタイトルを考慮した手法

4.1 手法の概要

次に，本手法の今後の改善の可能性を探る意味で，本手法に記事のタイトル情報を考慮した変更を加えた手法について実験を行い，その結果を分析する．実験対象・実験の手法など実験に関連する要素は，全て前節の実験と同じである．

変更した点は，コンプリメントタームの重みの集合を計算する部分である．

$w_{A,t}$ ：記事 A における ターム t の

表 3: 各記事群における，「記事のタイトル + コンプリメントタームを用いた手法」の精度と「コンプリメントタームを用いた手法」の精度の差

記事群	A	B	C	D	E
共通率	0.145	0.130	0.099	0.095	0.069
精度差	0.0292	0.0242	0.1032	-0.0385	0.0503

重要度

N : 文書群に含まれる文書数

$f_{X,t}$: 文書 X におけるターム t の出現数

df_t : 文書群でターム t が現れる文書数

$REL(A,X)$: 基準記事 A と記事 X の関連性

$$w_{A,t} = \frac{\sum_X (f_{X,t} \cdot \frac{1}{REL(A,X)})}{\ln(N/df_t)} \quad (4)$$

$$WCT(A,B) = \sum_{t \in CT(A,B)} w_{A,t} \quad (5)$$

式 (4) における $REL(A,X)$ は，基準記事 A と記事 X の関連性で，式 (6) で計算される：

A : 基準記事

$T(A)$: 記事 A のタイトルに含まれるターム

$$REL(A,X) = \exp(\sum_{t \in T(A)} f_{X,t}) \quad (6)$$

4.2 実験の概要

実験の方法は，3 節の実験と同じである．本手法と 4.1 節で説明した「記事のタイトル + コンプリメントタームを用いた手法」によって記事のランキングを行った上で再現率-精度曲線を描き，比較を行う．

4.3 結果

それぞれの記事群の再現率-精度曲線で，各再現率における，「記事のタイトル + コンプリメントタームを用いた手法」の精度と，「コンプリメントタームを用いた手法」の精度との差を合計した値を，表 3 に示す．

表から分かるように，「記事のタイトル + コンプリメントタームを用いた手法」は，「コンプリメントタームを用いた手法」の結果を改善する傾向にあるが，逆に悪くなる場合もある．また，結果の改善／改悪率は，ターム共通率をはじめとする，記事群そのものの特性とはあまり関係ないようである．

この結果から，若干の問題点は見られたものの，記事タイトルの情報を利用したことによる全体的な結果の改善は，「コンプリメントタームを用いた手法」に対する改良の可能性を十分に示していると言える．記事間の関連性を表す他の情報を併用することで本手法の性能を更に高めることに対しての見通しは，きわめて明るいと思われる．

5 まとめ

本研究では，情報検索において，ある記事群における 2 つの記事間の類似性（相違性）を判定する手法として，2 つの記事に共通しないターム（「コンプリメントターム」）を利用することを提案した．

実験では，特定のキーワードによってあらかじめ記事を絞り込んだ 5 つの記事群を用いて，記事のランキングの性能を調べた．この結果，互いに共通するターム数の多い記事群に対しては，本手法は良い結果を示すことが分かった．

また，記事のタイトルの情報を考慮した重みづけを用いた参考実験により，本手法に記事間の関係を示唆する様々な情報を加えることで，より高性能なシステムになりうる可能性を示した．

今後の改良点として，正規化の手法の改良，ランキングを途中で区切る閾値の設定，記事のタイトルをはじめとする様々な情報の考慮，重みづけそのものの改良，そして，コンプリメントタームの集合の中から重み計算に使用するタームの選別などが考えられる．

参考文献

[1] G.Salton and M.J.McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill Publishaing Company, 1983.

[2] B. Umino. Some principles of weighting methods based on word frequencies for automatic indexing. *Library and Information Science*, Vol. 26, pp. 67–88, 1988.