Blog 空間における拡大アンカーテキストと明示的リンク解析に基づくクラスタリング手法

神林 真実† 福田 直樹† 石川 博†

†静岡大学情報学部情報科学科 〒432-8011 静岡県浜松市城北 3-5-1

E-mail: †cs3037@s.inf.shizuoka.ac.jp, fukuta@cs.inf.shizuoka.ac.jp, ishikawa@inf.shizuoka.ac.jp

あらまし Blog はリアルタイムに一個人の意見を聞ける場として機能し，従来の Web ページよりも手軽に記述できる点から，多くの人に Blog サービスが利用されている．ユーザの増加に応じて Blog から得られる情報量も増加するため，Blog から情報を的確に得るための手法が重要になる．本論文では，Blog 記事におけるリンク方法の一種であるアンカーテキストや直接記載された URL に注目する．リンク先の内容は，記事の内容と密接に関わっていると考えられ，リンクの前後に現れる単語は，リンク先について特に言及している特徴語となりうると考えられる．そこで，本論文では，アンカーテキストと，リンクの前後に現れる名詞を含む“拡大アンカーテキスト”に重みを置いた文書解析の手法を提案する．提案手法を用いて作成する文書ベクトル空間モデルと，明示的リンク解析を行い作成するリンクベクトル空間モデルを併用することで，従来よりも精度の高い Blog クラスタリングを目指す．さらに，このクラスタリング結果を用いて，Blog から参照された Web ページを，有効な情報源としてユーザに推薦する機構を実現する．提案手法を 2 つの具体的な事例に適用し，その有効性を評価する．

キーワード データマイニング，ブログ，アンカーテキスト，リンク解析

A Clustering method based on Extended Anchor texts and Explicit Link Analysis in Blogspace

Mami KAMBAYASHI† Naoki FUKUTA† and Hiroshi ISHIKAWA†

†Department of Computer Science, Faculty of Infomatics, Shizuoka University

3-5-1 Johoku, Hamamatsu-shi, Shizuoka, 432-8011 Japan

E-mail: †cs3037@s.inf.shizuoka.ac.jp, fukuta@cs.inf.shizuoka.ac.jp, ishikawa@inf.shizuoka.ac.jp

Abstract Blogs allow us to get useful comments of or against blog user's opinions. Today, there a lot of people use Blog services because we can publish our opinions in Blog space much easier compared to ordinal authoring steps of Web pages. Useful information that we can get from blog space increases as blog users do, we need a method to instantly reach useful information in blogspace. In this paper, we pay attention to Anchor texts and URL contained in blog space since we can assume that the contents of linked Web pages are closely related to blog articles, and Anchor texts and nouns before and after the URLs are characteristic nouns that refer to the contents of Web pages. In this paper, we propose a clustering method of document analysis with Anchor texts and “Extended Anchor texts” that include nouns before and after the link URLs. We use the vector space model on both documents based on the extended Anchor text analysis and links based on explicit link analysis. We also realize a recommendation mechanism based on our clustering method. We show the effectiveness of our method by applying recent two popular topics in Blogspace.

Keyword Data Mining, Blog, Anchor Text, Link Analysis

1. はじめに

Web 上には莫大な量の情報が溢れ，その中から必要な情報を選択する必要が生じている．我々が現在最もよく用いる手段である検索エンジンでは，検索結果の一覧表示を表示するだけにとどまらず，類似した Web ページをクラスタリングすることによって，検索結果をよりユーザに解りやすくする手法が導入されつつある．

Web ページに対するクラスタリングの研究の代表的なものには，[1]などがあり，同じ URL を参照している（共参照）Web ページ間では，その内容は関連性が高く，リンク解析を取り入

れる事でクラスタリング結果の精度向上に繋がることが指摘されている．

近年登場した Blog は，Web ページとしては特殊な性質を持ち，新たな研究対象として注目を集めつつある．その背景に，近年見られる Blog 人口の増加が挙げられる．総務省の調べ[2]では，平成 18 年 3 月末において Blog 人口は 868 万人，前年度と比べて倍近い増加が確認されている．ユーザが Blog を通して，製品の感想を伝えるなど，口コミの場として利用している場合も多く見られるように，Blog はリアルタイムに一個人の意見を聞ける場であり，従来の Web ページよりも気軽に頻繁に更新されていることから，Blog から得られる情報は今後増加していくことが予測される．それに伴って，Blog から情報を的確に得るための技術が重要となる．

本研究では，Blog 記事内における明示的リンクを用いるリンクベクトル空間モデルと，アンカーテキストとその前後の単語までを含む“拡大アンカーテキスト”に重みを置いた文書ベクトル空間モデルの併用により，従来手法より精度の高い Blog クラスタリングを実現する．さらに，同クラスタに属する Blog 記事集合内で外部リンクが提示されていた場合，その中の有益なリンクをユーザに推薦可能とする．

2. 関連研究

Web ページ内のリンクは，製作者によって意図的に作られたものであるため，リンク先の内容とリンクを張っている元の Web ページの内容は関連している可能性が非常に高いと考えられる．Web ページのリンク構造を解析しクラスタリングに応用する事で，より精度の高いクラスタリングを実現できることが指摘されている．

高橋ら[1]は，共参照の関係である Web ページ間の特徴を基にしたリンクベクトル空間モデルと，文書ベクトル空間モデルを併用した Web クラスタリング手法を提案している．Yitong Wang ら[3]は，Web ページを対象にインリンクとアウトリンクを解析し，検索結果のクラスタリングの精度向上を実現した．阿部ら[4]は，文書全体ではなく，アンカーテキストのみに着目したテキスト解析とリンク構造の両面を考慮した Web 情報検索手法を提案している． また，鈴木ら[5]は Web のハイパーリンク構造とアンカーテキストをもとに複数サイトから階層的なディレクトリ構造を作成し，Web ページを分類する手法を提案している．しかしこれらの研究は一般の Web ページに対する手法であり，Blog 固有の特性は考慮していない．

Blog 固有の特徴を生かしたリンク解析の研究としては，石田[6]が Blog のリンク構造分析を行い，Blog と参照 Web ページにより構成される二部グラフにおいて，開発した最弱対アルゴリズムを用いて意味的集合がある部分二部グラフに分割する手法を提案しているが，我々が着目する Blog 特有のアンカーテキストの性質は考慮していない．また，大手ポータルサイトである Yahoo![7]は Yahoo!ニュースの URL を記事中に記載した Blog に対してリンクを張るというシステムを導入している．このシステムの導入により，今後は大手ニュースサイトからリンクを目当てとして URL を Blog 中に記載する事例は多くなると考えられる．Blog 記事中に，その情報源，あるいは参照情報がリンクとして示される機会の増えることが予想され，この Blog 固有のリンク情報やリンクの作法を考慮して，Blog 記事やその情報源となった Web ページを検索できるようになれば，有益ではないかと考えられる．

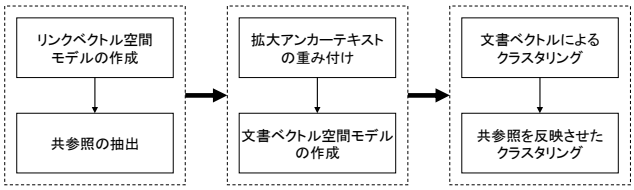

図 1　手法の流れ

本論文では，Blog の書き方の特徴を考慮した，拡大アンカーテキストにおける重み付けを提案する．高橋らのリンクベクトル空間モデルと文書ベクトル空間モデルを併用する方式を元に，提案手法ではそれを Blog に拡張する．クラスタリング結果を用いて，Blog から参照された Web ページを有効な情報源として推薦する機構を実現する．

3. 提案手法

Web のリンク方法には，トラックバックやコメント等があり，そのひとつに，<A>タグを用いてリンクを張る，ハイパーリンクと呼ばれる方法がある．ハイパーリンクを閲覧者がクリックするとリンク先にジャンプすることが可能となる．また，<A>タグで挟まれたテキストをアンカーテキストという．リンク先の内容とアンカーテキストの内容は類似している可能性が非常に高いと考えられる．近年の Blog サービスの 1 つに，URL とアンカーテキストを与えれば自動でハイパーリンクを作成する機能がある．また，エントリ作成画面でネット通販サイトの商品を検索し，商品画像つきリンクをエントリ中に記載可能とするサービスを提供している Blog もある．このようなサービスの充実から，Blog ユーザが手軽にハイパーリンクを用いることが可能となった．しかし，ハイパーリンクを用いてリンク先へジャンプをすると，アクセス解析によりジャンプ元が判明する場合がある．ハイパーリンクの存在をリンク先から秘匿する目的や，ユーザの表現方法の好みから，URL をエントリ中に「公式 HP→http://www…」のように直接記述している場合も多く見られる．この場合，アンカーテキストに相当するようなリンク先の内容を簡潔に表現した文が，直接的には指示されない．この問題を解決するために，本論文では“拡大アンカーテキスト”を提案する．

本提案手法の流れを，図 1 に示す．本手法は，大きく分けて，「リンクベクトル空間モデルの生成」，「文書ベクトル空間モデルの生成」，および「共参照を反映させた文書のクラスタリング」

の 3 つからなる．文書ベクトル空間モデル作成の際，提案する“拡大アンカーテキスト”による重み付けを考慮する．これらを考慮した上で，文書ベクトルを基にクラスタリングを行い，そのクラスタに対して共参照抽出結果を反映させる．

3.1. リンクベクトル空間モデル

リンクは“明示的リンク”と“暗示的リンク”に分けられ，前者は URL が明記されている事を，後者は URL が明記されていないが，エントリ内容から他の Web ページとリンクしている事が想定可能な Web ページ間の繋がりを示す．

このリンクベクトル空間モデルでは，画像を除いた，アンカーテキストのリンク先 URL，直接記述された URL を“明示的リンク”として用いる．

最初に各エントリから明示的リンクを全て抽出し，抽出 URL 総数を n 個とする．エントリ D_i における URL L_e の出現回数を用いて要素 W_{ie} とし，リンクベクトル空間モデルを作成する．

次に，リンクベクトル空間モデルを基にエントリ間のリンクの類似度を求め，階層的クラスタリングを行う．類似度は，ベクトルの内積から余弦を求めることで算出する．エントリ D_i, D_j 間の類似度 $sim(D_i,D_j)$を求める式を式①に示す．

$$sim(D_i,D_j)=\frac{W_{i1}W_{j1}+\ldots+W_{in}W_{jn}}{\sqrt{W_{i1}^2+\ldots+W_{in}^2}\times\sqrt{W_{j1}^2+\ldots+W_{jn}^2}} \quad (n\geq 1) \quad \cdots \quad ①$$

クラスタリングは類似度を正の数の範囲とし，リンクによるクラスタリングでの同一クラスタに属するエントリ間の最終的な類似度は，リンククラスタリング結果に基づき高く評価を行う．

3.2. 文書ベクトル空間モデル

エントリおよびタイトルに対して形態素解析ツール Sen[8]を用いて形態素解析を行い，数や代名詞を除いた名詞のみ抽出する．この時の，総エントリ中の抽出名詞数を n 個とする．本研究では，文書ベクトル空間モデルの要素 W_{ik} に tf/idf 法による重み付けを用いる．tf_{ik} は文書 D_i における名詞 T_k がどれくらいの頻度で出現するかを，df_k は名詞 T_k が出現する文書数を，N は文書総数を表している．式②を用いて要素 W_{ik} を求める．

$$W_{ik}=tf_{ik}\times\left(\ \log\frac{N}{df_k}+1\ \right) \quad (1\leq k\leq n)\cdots② $$

算出した W_{ik} を用いて，文書ベクトル空間モデルを作成する．その際に，タイトルとはエントリの内容を最も簡潔に表現したものである可能性が高いと仮定して，タイトルに出現した名詞については重みをつける．

3.3. 拡大アンカーテキストによる重み付け

Blog におけるアンカーテキスト･URL 記載の前後の文脈に着目する．ここでは，画像 URL も対象に含む．リンクベクトル空間モデルの対象となる URL では画像を含まず，拡大アンカーテキストにおいて中心となる URL に画像を含む理由は，画像を載せたユーザにとって，画像の内容が持つ情報量が高いと仮定できるからである．画像を載せる大半のユーザは，エントリにおいて画像について何らかの言及を行っている．画像は参照ページと同様に情報源としての重要度を持つと本論文では仮定する．これらの前後の語は，画像を含むリンク先について特に言及している特徴語であると仮定できる．本研究ではこれらに重みを付けることで，文書の特徴をより明確にし，文書ベクトルのクラスタリングの精度向上を目指す．

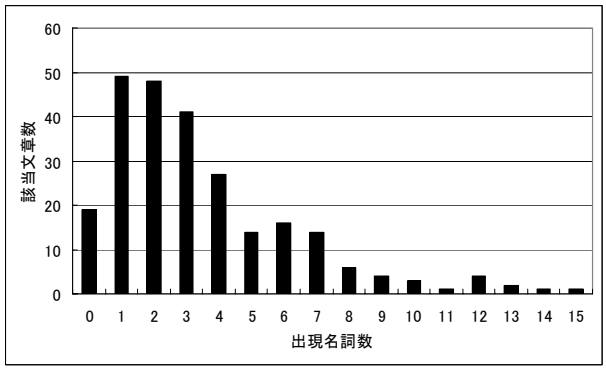

図 2　1 文中での出現名詞数

拡大アンカーテキストの抽出では，エントリ中において URL 記載場所を原点と考える．また，本手法では名詞をエントリ中の距離を測る単位として用い，原点からの距離を名詞の出現毎に 1 とする．ここで，アンカーテキストはリンク先 URL の真上，つまり原点に配置されているとするので原点からの距離は 0 となる．しかし，アンカーテキストを用いてリンクを張るユーザもいれば，URL を直接エントリに書き込むユーザもいるため，アンカーテキストのみに重点を置いては重み付けが適切とならない場合がある．本手法では，URL がエントリの冒頭部分に記載されている場合，URL 記載後リンク先内容言及を行い，エントリの非冒頭部分に記載されている場合，リンク先内容言及後 URL 記載であると判断し，前者では URL の直前の名詞群に，後者では URL の直後の名詞群に重みを置く．本論文では原点を中心に，重み付けを文単位に行う．

そこで，1 文が平均何個の名詞で構成されているのか調査を行った．250 個の文章を用いて，1 文が平均何個の名詞で構成されているのか調査した．使用した文章は複数の Blog から手動でランダムに選んだものである．その結果を図 2 に示す．出現名詞数は 0 から 15 まで分散していたが，1 文平均 3.5 個の名詞で構成されているという結果を得た．

3.3.1. 重み付け範囲と重みの定義

予備実験の結果から，重み付けを文単位で行うために 1 文を名詞 3 個とし，原点から M 文目まで重み付けを行うとして範囲 D を次のように定義する．

$$D=\{D_1,D_2\ldots,D_M\}=\{3,6,\ldots,3M\} \quad (M\geq 1)$$

また，範囲 D における重み G は G_1 の重み g を基準に次のよ

うに定義する．

$$G = \{G_1, G_2, ..., G_M\} = \{g, \frac{g}{2}, ..., \frac{g}{2^{M-1}}\} \quad (M \geq 1)$$

3.3.2. 重み付けの方向

本研究における URL の記載位置の基準として，“冒頭部分”と“非冒頭部分”がある．ここで URL の記載位置を，“先頭”と“非先頭部分”としない理由として，Blog の多くが一個人の日記として活用されているため，また第三者に向けて書かれたものであるため，挨拶や報告から始まる場合が多いことがあげられる．また，エントリ内容の情報源として，エントリの最後に情報源・関連 Web ページを紹介している場合も多く見られる．この場合，Web ページのタイトル・簡潔な紹介文をアンカーテキストに用いたハイパーリンク，または Web ページのタイトル・簡潔な紹介文の直後に URL を記載するという形がとられることが多く見られる．エントリの中盤でアンカーテキストまたは URL が見られる場合もあるが，この場合，日本語の形容詞・修飾語の関係から，名詞の前にその名詞と関連する語が来る場合が多い．以上から，本手法では URL の記載箇所がエントリの冒頭部分ならば URL 記載後に内容言及を記載，URL の記載箇所がエントリの非冒頭部分ならば URL 先の内容言及後に URL 記載していると仮定する．

ここで，冒頭部分と非冒頭部分を決定するために，予備実験を行う．エントリ 9117 件を解析したところ，名詞出現回数は 379913 回となり，1 エントリの平均名詞出現回数は 41.67 回となった．また，1 文平均 3.5 個の名詞で構成されている（3.3 参照）ので，[平均名詞出現回数／3.5]より 1 エントリが平均何文で構成されているかを調べた．本研究においては 1 エントリ平均 11.9 文で構成されているという結果を得た．

前述したように，エントリの最初の 1 文を挨拶または報告の文である可能性を考慮して，冒頭に URL 記載をする場合，2 文目前後に URL を記載すると推測する．

よって本研究では，エントリ D_i における名詞総出現回数を Ci とし，(C_i／5）番目の名詞出現までを冒頭部分，それ以降を非冒頭部分と定義する．そして，冒頭部分に URL が記載された場合, 原点から後ろに向かって M 文目までに出現した名詞に対して特徴語としての重みを置き，非冒頭部分に URL が記載された場合, 原点から前に向かって M 文目までに出現した名詞に対して重みを置く．

3.4. クラスタリングと情報源の推薦

作成した文書ベクトル空間モデルにおいて，拡大アンカーテキストによる重み付けを行った後，階層的クラスタリングを行う．その際，式①を用いて，類似度を算出する．また，１クラスタに属するエントリが２以上になった場合は，最長距離法に基づき，類似度計算対象の２つのクラスタに含まれるエントリの組み合わせで最も類似度が低いものを，２クラスタ間の類似度とし，類似度が閾値以上ならばクラスタリングを行う．その後に共参照解析によるエントリ間類似度の重み付けを行う．本研究においては，共参照の関係にあるエントリを，ひとつのクラスタにするとする．

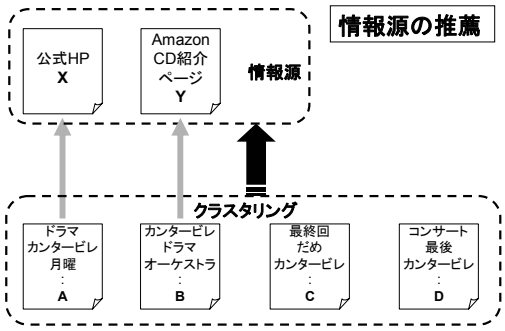

図 3 情報源の推薦

またクラスタリング結果において，同クラスタに属すエントリ群が, それぞれ別の URL を参照していた場合, それらの URL を１つにクラスタリングすることで同一情報を扱う情報源と仮定し，複数の情報源の推薦を行う．

明示的リンクを用いるユーザは増加しているが，全体的には 2 割弱のエントリにしか明示的リンクが見られないというのが現状であり，圧倒的に明示的リンクのないエントリの方が多い（本実験においては約 320000 件のエントリ中 4800 件強にしか明示的リンクは見られなかった）．

本論文における情報源の推薦では，あるクラスタに属するエントリが，情報源となる Web ページを参照していた場合，同クラスタに属する他のエントリも, 仮想的にその情報源 Web ページを参照しているとし，そのクラスタにおいて参照された情報源 Web ページをユーザに推薦する．

4. 評価実験概要

本実験では, クローラで収集した Livedoor ブログをベースに，特定の Web ニュースに共参照している Blog のうちランダムで 7 件を追加したデータセットを用いる．エントリの追加は，異なるユーザによる 3 件以上のエントリからの共参照が，商用 Blog 以外に抽出されなかったためである．このデータセットを用いて，提案手法に対する比較実験を行い，手法の有効性の評価を行う．

実験は大きく 2 つに分けられる．Blog における明示的リンク解析実験と，類似度によるクラスタリング実験である．クラスタリング実験はさらに，拡大アンカーテキストの妥当性検証実験, 複数の名詞出現エントリリストを基にした手法の比較実験, 情報推薦の妥当性検証実験に分かれる．

5. 明示的リンク解析実験

まず，リンクベクトル空間モデルを用いた共参照抽出結果を提示する．用いたデータセットはエントリ 37279 件である．解析の結果，URL を持つエントリは全体の約 13%にあたる 4775 件，URL 記載回数は 23693 回，URL 種類数は 10790 件であっ

た．また，Web ページを参照したエントリが 1 件のみの，共参照が抽出できなかった URL が全体の約 8 割の 8035 件であった．参照エントリが 2～10 件の URL 数は 2714 件，参照エントリが 11～30 件の URL 数は 28 件であった．参照エントリ数が 30 以上の URL 数については, 最高は 147 エントリからの共参照で, 31 エントリ以上から参照された URL 数は 13 であった．

解析結果から，大半の URL が共参照されていないことがわかる．また，共参照されている URL において，8 以上のエントリから参照されている URL はほぼ全てが, ネット通販サイト，オークションサイト，Blog に設置可能なアプリケーションサービス等などであった．このことから，抽出された URL の大半から共参照の関係は発見できないが，上位の共参照関係の抽出から，現在 Blog において多数見られる広告・商用専門の Blog をクラスタリングすることが可能になると考えられる．また，共参照として最も情報量の高い可能性のある URL は，現時点のデータセットサイズにおいては，その URL が出現するエントリ数が 2 以上 6 未満程度の間であると推測できる．この値については，データセットが大きくなればそれだけ共参照が抽出できる可能性は高くなり，上限値は増加すると推測できる．

この抽出した URL を元にリンクベクトル空間モデルを作成し，URL を持つエントリ 4775 件をクラスタリングする．類似度が正の数である限りクラスタリングを行う，すなわち 1 つでも共参照ページを持てばクラスタリングを行うという条件下で行った．その結果，1429 回のクラスタリングが行われ，3346 のクラスタが得られた．その内 390 のクラスタにおいて共参照が抽出され, 残り 2956 のクラスタにおいて共参照は抽出されなかった. この時, 最高 147 のエントリが 1 クラスタに分類され, それらは広告系サイトを共参照とし, 商業目的の Blog エントリであることが確認された.

6. 類似度によるクラスタリング結果解析

文書ベクトル空間モデル作成には，リンクベクトル空間モデルを作成したデータセットと同じエントリ 37279 件を用いる．その中から特定の名詞が出現するエントリに注目し，クラスタリングの精度を検討する．今回は拡大アンカーテキストについて，原点から 2 文目までを範囲，クラスタリングは類似度 0.4 以上のクラスタ間において行った．また，重み G は式③によって算出された 4.0 を用いる．N は全エントリ数，E_m は m 番目のエントリ，m は 1≦m≦N を示す．

$$G = \frac{\sum_{m}^{N} \{E_m(\text{最大}tf/idf) - E_m(\text{平均}tf/idf)\}}{N} \quad \cdots\text{式③}$$

また，“今日”や“前”などの特徴語として成り立たない名詞への重み付けを回避するために，平均 tf/idf 値以下の tf/idf 値を持つ名詞に対して，拡大アンカーテキストの重み付けを行わないとする．

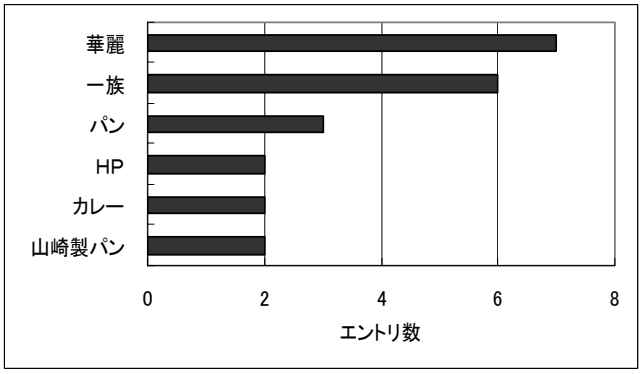

図 4 “華麗”に関する拡大アンカーテキスト該当名詞

6.1. 拡大アンカーテキストの妥当性

実験で用いる検索語は“華麗”と“ソフトバンク”である．実験データにおいて, 両名詞についての情報源 Web ページの参照が抽出されている．

6.1.1. 拡大アンカーテキスト抽出実験 1

拡大アンカーテキストによる重み付けが同じトピックで検証できるように，実験では“華麗”と，“一族”が出現するエントリ群において, トピックと関連があると思われる明示的リンク，または明示的リンクを持ち，このトピックがエントリのメイン話題であるエントリのみに注目する. このとき，“華麗”, “一族”などのドラマに関連する名詞が拡大アンカーテキストとして抽出することが望ましい.

データセットエントリ 37279 件中，明らかな商用サイトを除き，上記に該当するエントリは 9 件であった．この 9 件の特徴として，アンカーテキストを用いているエントリが多い事があげられる．9 件のエントリにおいて，拡大アンカーテキストに含まれ, ニュースと関連が深いと判断される名詞は延べ 27 個であった．拡大アンカーテキストに含まれ，ニュースと関連が浅いと判断される名詞は延べ 50 個であった．また，2 エントリ以上で重み付けが見られたものを図 4 に示す.

6.1.2. 拡大アンカーテキスト抽出実験 2

昨年からニュースとして何度も取り上げられている，“ソフトバンク”に注目する．“ソフトバンク”と“携帯”について扱っているエントリのみを抽出し, さらにその中から，“ソフトバンク”に関連する情報源を参照しているエントリ 10 件を用いる．このとき，“ソフトバンク”と“携帯”と関連が深い名詞が，拡大アンカーテキストとして抽出する事が望ましい．このエントリ 10 件中半分ほどが直接記述 URL，残りがアンカーテキストによる URL であった．この実験において，拡大アンカーテキストに含まれ，ニュースと関連が深いと判断される名詞は延べ 50 個であった．拡大アンカーテキストに含まれる，ニュースと関連の浅いと判断される名詞は延べ 36 個であった．また，2 エントリ以上で重み付けが見られたものを図 5 に示す.

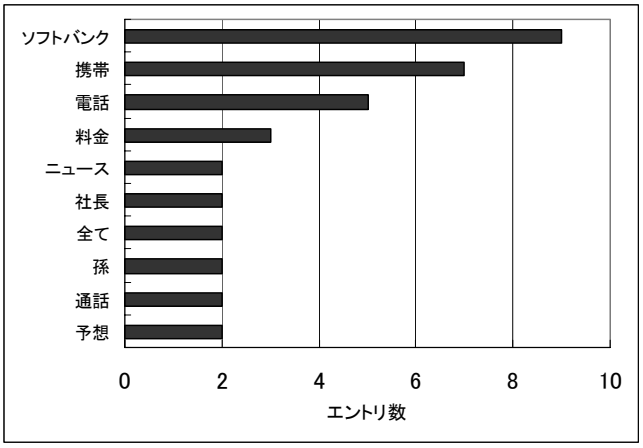

図 5 “ソフトバンク”に関する拡大アンカーテキスト該当名詞

6.1.3. 考察

実験 1, 2 から，拡大アンカーテキストのルールに従って抽出した名詞には，取り上げるトピックに関連の深い名詞が多い事がわかった．特に，抽出が成功しやすいのは，URL がエントリの冒頭部分又は文末部分に記載されているときであることがわかった．

複数トピックを同時に扱う傾向の見られたエントリに関しては，関連の浅い語が特に多くなってしまった．これは実験 1 に強く現れている．また，エントリ中盤に明示的リンクを記載する場合，本論文においては前述したように，明示的リンクの前部分に向かって重み付けを行っているが，これが特徴語抽出の失敗に繋がることもあった．抽出に成功している事例も多くあるため，この成功の事例を元に，中盤に明示的リンクが記載されていた場合の拡大アンカーテキストの重み付けルールの詳細を改変する必要があると考えられる．たとえば，まず，エントリ中の改行等を基準としてエントリ全体のトピックの切れ目を見つけ出し，そこでエントリを分割した後に，分割後のエントリに拡大アンカーテキストの現在のルールを適用する方式があげられる．

6.2. クラスタリング実験

データセットであるエントリ 37279 件から特定の名詞（以下キーワードとする）が出現するエントリ郡をクラスタリングする実験を行う．本論文で提案する手法の有効性を確認するために, 文書ベクトルのみを用いた類似性によるクラスタリング(手法 1)，文書ベクトルとリンクベクトルを用いた類似性によるクラスタリング（手法 2）と，提案手法である拡大アンカーテキストを用いた文書ベクトルとリンクベクトルを用いた類似性によるクラスタリング（手法 3）を比較する．

6.2.1. クラスタリング実験 1

クラスタリング実験 1 で用いるキーワードは“納豆”である．また，大手ニュースサイトの“納豆とコレステロールで捏造発覚”というトピックに共参照し，そのニュースについて議論しているエントリが 5 件存在する．その 5 件を Co-A～E とする．

表 1 クラスタリング実験 1 による生成結果

手法	クラスタ	特徴語
1	Ⅰ	納豆 ダイエット 放送 捏造 辞典
	Ⅱ	捏造 コレステロール 納豆 アミ 関西テレビ
2	Ⅲ	納豆 捏造 ダイエット 放送 コレステロール
3	Ⅳ	納豆 放送 朝晩 辞典 ダイエット
	Ⅴ	納豆 捏造 コレステロール ダイエット 疑惑

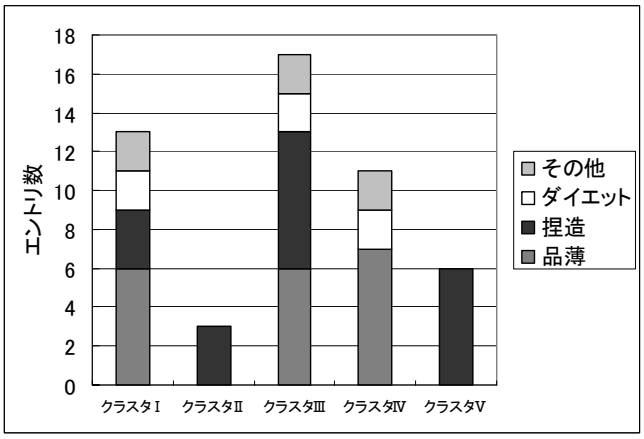

図 6 実験 1 によるクラスタ内トピック

“納豆”が出現したエントリは 172 件であった．手法 1 において生成されたクラスタは 143 個，手法 2 において生成されたクラスタは 137 個，手法 3 において生成されたクラスタは 138 個であった．

手法 1 によるクラスタリングによって生成されたクラスタにおいて，“ダイエット”，“捏造”などを特徴語として持つ，クラスタサイズが 13 のクラスタに注目する. これをクラスタⅠとする．そして，クラスタⅠに含まれる全エントリの名詞別 tf/idf 値総和を算出し，上位 5 個をクラスタの特徴語とした場合の，このときの特徴語を表 1 に示す（以下クラスタⅠ～Ⅴの特徴語を表 1 に示す）．また，前述の Co-A～E のうち，Co-A～C はクラスタⅡに属し，Co-D のみクラスタⅠに，Co-E はクラスタサイズ 1 の別クラスタに属している．ここで，クラスタⅠに属するエントリを解析すると，特徴語となる“捏造”に関連するエントリは，13 エントリ中，僅か 3 エントリのみであった．“納豆ダイエット”の方法について記述しているものが 2，“納豆が品薄である”という内容のものが 6 エントリ，その他が 2 であった．クラスタⅠ～Ⅴ内のトピックについては図 6 に示す．

手法 2 によるクラスタリングでは，手法 1 で生成されたクラスタⅠとクラスタⅡが統合され，クラスタⅢが生成された．この時のクラスタサイズは 17 であった．その結果，特徴語は表 1 のように“捏造”がまた上位となり，クラスタ内のエントリも，“捏造”が話題のエントリ，“品薄”が話題のエントリと，話題の混在が見られた．

手法 3 によるクラスタリングにおいては，手法 1 におけるク

表 2　クラスタリング実験 2 による生成結果

手法	クラスタ	特徴語
1	A	プラン ホワイト ソフトバンク 通話 月額
	B	ソフトバンク 商法 非常識 告発 幻想
	C	守秘 義務 一個人 社員 告発
2	D	プラン ソフトバンク ホワイト 通話 月額
	E	ソフトバンク 告発 商法 企業 銘柄
3	F	ソフトバンク プラン ホワイト 熊本 親父
	G	月額 円 プラン 通話 ソフトバンク
	H	ソフトバンク 通り道 新型 サービス 日経
	I	ソフトバンク 告発 商法 店員 詐欺

ラスタ I に属していた“納豆が品薄である”，“納豆ダイエット”の話題のエントリが，“捏造”の話題のエントリとは別にクラスタⅣに属す結果となった．クラスタⅣは，クラスタ I には属していなかった，“品薄”について書かれたエントリも属し，クラスタサイズは 11 であった．また，手法 3 を用いた場合，“捏造”を話題にした Co-A～D は，クラスタⅤに分類された．この時のクラスタサイズは 6 であり，Co-A～D 以外にも，クラスタ I に属していた“捏造”を話題としたエントリ 1 件が属し，さらに，専門用語をあまり使わずに，ニュースに対する感想を述べている Co-E も属す結果となった．よって，手法 3 においては，手法 1，2 よりも話題の混在が少なくなったといえる．

6.2.2.　クラスタリング実験 2

クラスタリング実験 2 で用いるキーワードは“ソフトバンク”である．また，大手ニュースサイトの“ソフトバンクの商法を詐欺告発”というトピックに共参照し，そのニュースについて議論しているエントリが 2 件，ニュース配信 Blog と思われるエントリが 2 件存在している．その議論されているエントリ 2 件を Co-F，G とする．また，“ソフトバンク”，“携帯”に関連する拡大アンカーテキスト該当エントリが 10 件含まれている．

“ソフトバンク”が出現したエントリは 125 件であった．手法 1 において生成されたクラスタは 99 個，手法 2 においては 90 個，手法 3 においては 90 個であった．

手法 1 によるクラスタリングによって生成されたクラスタにおいて，“ホワイト”，“プラン”などを特徴語として持つ，クラスタサイズが 5 のクラスタに注目する．これをクラスタ A とする．クラスタ A の特徴語を表 2 に示す（以下クラスタ A～I の特徴語を表 2 に示す）．クラスタ A に属しているエントリは全て，ホワイトプランについて記述されており，内容的に非常に類似していると言える．また，前述の Co-F，Co-G はそれぞれが別クラスタであるクラスタ B，C に属していた．クラスタ A～I 内のトピックについては図 7 に示す．

手法 2 によるクラスタリングにおいては，手法 1 で生成されたクラスタ A に対し，共参照によるクラスタの結合が行われ，クラスタ D が生成された．このときのクラスタサイズは 7 であった．今回の結合は，同ユーザ作成エントリに見られる共参照によって行われた．特徴語の大幅な変化は見られなかったが，異なるトピックの混在が起こってしまった．また，Co-F と Co-G は共参照によりクラスタ結合が行われ，クラスタ E に属する結果となった．この時クラスタ E のエントリは，前述したニュース配信 Blog を含めた 4 件であった．

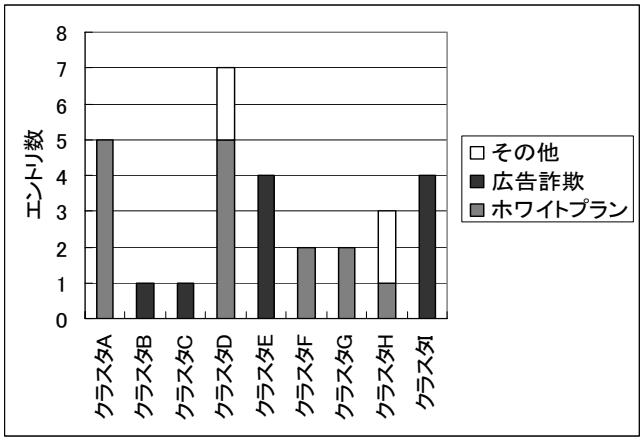

図 7　実験 2 によるクラスタ内トピック

手法 3 によるクラスタリングにおいて，手法 1 によって生成されたクラスタ A に属する 5 エントリは，3 クラスタに分かれてしまう結果となった．5 エントリ中，2 エントリがクラスタ F，2 エントリがクラスタ G，1 エントリがクラスタ H に属した．クラスタサイズはそれぞれ 2，2，3 となっている．このとき表 2 からわかるように，クラスタ G～H は，ホワイトプランについて記述しているが，手法 1，2 に比べて特徴語もトピックに関係のないものが上がってきてしまう，類似しているのに別クラスタという結果となった．また，Co-F，Co-G は手法 2 と同じエントリが属すクラスタ I であったが，拡大アンカーテキストによってトピックと関連語の深い名詞への重み付けが行われたために，クラスタ I の特徴語は，手法 2 よりもトピックに関連の深い名詞となった．

6.2.3.　考察

実験 1 は，提案手法である手法 3 でのクラスタリング結果が最も良いものとなった．提案手法が，共参照関係のエントリ感の類似度が低くても，クラスタリングが成功することが関係すると考えられる．

実験 2 においては，提案手法が手法 1，2 と比較し，悪い点と，良い点が見られた．実験 2 では，同ユーザの記述したエントリに見られる，参照ページではない“自己推薦ページ”（ブログランキング，又は，他に自分が運営している Web ページなど）を，共参照として抽出してしまったために起こる，トピックの混在が起こってしまった．この為，同一ユーザのエントリを複数扱う場合の明示的リンクは，誤った共参照抽出が行われないようにする必要がある．

実験 2 での欠点を除けば，共参照によるクラスタリングを用いると，参照 URL を記載しているため詳細を多くは記述せず，

意見を多く記述しているエントリ間のクラスタリングが可能になる，という長所を持ち合わせている．商用 Blog やニュース配信を目的とした Blog の除去があれば，共参照は非常に有効な手法であり，クラスタリング後のラベル付を行う際に，拡大アンカーテキストは効果があると考えられる．

6.3. 情報源の推薦の妥当性について

情報源の推薦の妥当性について，クラスタリング実験 1，2 のデータを用いて検証を行った．実験結果から，情報推薦数が 1〜3 個が最も多いことがわかった．また，21 以上推薦しているクラスタも複数存在していた．21 個以上を推薦しているクラスタの情報推薦数は，最高 70 以上にもなり，エントリのほとんどがネット通販を目的とした Blog であることが確認された．今回，大手のアフィリエイト URL は削除して実験を行ったが，このように，小規模でも商用目的の Blog はリンク解析を行うことで判別が可能であることがわかる．今後の課題としては，一定以上の情報推薦数を持つクラスタを除去することで，商用 Blog を除くことがあげられる．

また，情報推薦数が 1〜10 個のクラスタに関して，全く商用 Blog が混在しないとは言えない結果が見られたが，ほとんどの推薦先が，公式ページ，ニュースリンク，自己推薦 Web ページであった．

7. 実験による全体的考察

複数の実験から，提案手法による成功例，失敗例を得て，次のような課題を得た．

拡大アンカーテキストは，明示的リンクが冒頭又は文末に記載されているときに重み付けが適切である可能性が高い．逆に不適切な可能性が高いのが，明示的リンクがエントリ中盤に記載されている場合である．この対応策として，改行を目安にエントリ内のトピックの切れ目でエントリを分割，重み付けを行うことがあげられる．他に，特徴語が適切に抽出できてもクラスタリングの精度向上に繋がらない場合もあることがわかった．

共参照のクラスタリングへの利用は，参照 URL を記載し詳細を記述せず，意見を多く記述しているエントリでもクラスタリングが可能であることが長所としてあげられる．しかし，内容に関係なく，お奨めの Web ページをエントリに記載するユーザもいるため，誤った共参照によるクラスタリングが起こり，トピックの混在の発生の原因にもなることがわかった．

8. おわりに

本論文では，拡大アンカーテキストによる重み付けを提案し，文書ベクトル空間モデルとリンクベクトル空間モデルの併用による従来よりも精度の高い Blog クラスタリングと，Blog から参照された有効な情報源の推薦手法を提案した．実験により，拡大アンカーテキストに基づいてリンク先と関連のある語の抽出が可能であること，さらに，Blog における共参照関係がクラスタリングへ有効であることがわかった．

今後の課題としては，エントリ中盤に明示的リンクが記載された場合の拡大アンカーテキスト重み付けルールの改善である．今回の実験のサンプル数は少なく，クラスタサイズが小さくなっている．様々な Blog サービスから成る大規模データである本来の Blog スペースにどの程度まで適用可能であるかを調べることも課題の 1 つである．

明示的リンクの中にはアフィリエイトなどの広告・商業サイトへのリンクが多く含まれ，現状ではそれらの除去は完全には行えていない．複数の情報源の推薦において，これらの除去は必須であり，除去方法については研究の余地が多くある．また，Blog の場合，ユーザの趣旨によって同意義の名詞を，漢字，平仮名，片仮名などで表現するため，本来ならば一つの名詞が複数の異なった名詞として扱われてしまう．ベクトル作成の際に用いる名詞の選出の洗練は，今後の課題である．

参考文献

[1] 高橋 功, 三浦 孝夫, “ハイパーリンクの共起性を用いたクラスタリング手法” , DEWS2005 1C-i12, 2005

[2]総務省, “ブログ・SNS の現状分析及び将来予測” http://www.soumu.go.jp/s-news/2006/060413_2.html, 2006.4

[3] Yitong Wang, Masaru Kitsuregawa, “Use Link-based clustering to improve web search results”, Proc.WISE’01 , 2001

[4]阿部 匡史, 豊田 正史, 喜連川 優, “アンカーテキストとリンク構造解析を用いた Web 情報検索の改善”, DEWS2003 2-P-04

[5]鈴木 祐介, 松原 茂樹, 吉川 正俊, “アンカーテキストとハイパーリンクに基づく 文書の階層的分類”, JSAI2005 3C2-02

[6]石田 和成, “潜在的ウェブログコミュニティの抽出のための二部グラフ分割アルゴリズム”, 第 9 回 セマンティックウェブとオントロジー研究会, 人工知能学会, SIG-SWO-A404-01, 2005

[7]Yahoo! Japan http://www.yahoo.co.jp/

[8]日本語形態素解析システム Sen http://ultimania.org/sen/