利用者の観点を反映した文書分類に関する研究

丸　田　　要

目 次

第1章　はじめに

1.1　本研究の背景

インターネットが普及して以来，多くの人が気軽にインターネットを利用して情報収集することができるようになった．また，文書群である記事もネット上に膨大な量が存在している．そのため，膨大な量の記事データの中から素早くかつ網羅的に目的の情報や記事を効率良く見つけ出す方法が必要である．

ユーザがWeb上にある目的の記事を探す場合，一般的にGoogle[1]やYahoo![2]などの検索エンジンを使用する．しかし，検索結果の中にユーザの求めるページが高い順位で表示されない問題[3]が指摘されている．そして，複数の分野で使用されるような検索クエリによっては検索結果である文書集合の中にユーザが必要としない分野の記事が含まれてしまう．その場合，効率良く目的の記事を探すことが困難となる．そのため，検索結果である文書集合を効率よく整理・検索する手法が必要である．その手法として，第2.4節で説明するようにページに順位を付けることにより，表示する順番を改良するランキング手法[4][5][6]と検索結果をクラスタリングする手法[7][8][9][10][11]がある．しかし，ランキング手法では表示の順番を改良するだけなので，元々の検索結果に不要な記事が含まれている場合はそのような記事が混在する問題を解決できない．検索結果をクラスタリングする手法は，理想的な分類ができた場合はユーザにとって最適なカテゴリの中のみを効率良く探すことができるため，本論文ではその手法に着目する．もちろんユーザの望む分類でなければ見当違いのカテゴリを探してしまうため，効率的に目的の文書を探すことができない．文書分類において留意しなければならない点がある．それは，ユーザの望む分類を行う際，分類結果は分類を行うユーザの目的・観点により異なってしまうことである．つまり，ある単一の文書データは観点が異なると分類されるクラスが異なる場合がある．その場合ユーザが考える分類とシステムによる分類に差異ができ，その差異部分に含まれる文書データはユーザの情報検索の阻害や見落とし

を発生させると考えられる．そして，ユーザの分類例を教師データと考えた場合，第 2.4.2 項で説明する検索結果のクラスタリング手法 [7][8][9][10][11] は教師無し分類とみなすことができる．そこで，ユーザによる文書分類例から教師データとなる観点を求め，それをテキスト分類に反映させることでユーザの望む教師有り分類を行う．

1.2 本研究の目的

本論文では，ユーザの観点とはユーザが文書分類の際に用いる無意識下の指標であると考える．そのため，ユーザがテキスト分類を行う際の観点を求めるには観点を明示的に表現することが望ましいが，ユーザの観点は人の感覚に依るところが大きいためそれは困難である．そこで，本論文では観点の特徴は分類結果である教師文書中の単語の分布に現れると考え，単語ごとに観点に対する寄与度を定めることにより近似的に観点を表現し，分類の指標に用いる．その観点を表す寄与度を行列分解を使用した文書分類手法に導入することで，ユーザの観点を文書分類に反映し，ユーザが目的のカテゴリに絞って目的の文書を効率よく検索することができることを目指す．

本論文では特に，（１）観点の特徴を抽出し分類に反映することでユーザが望む精度の良い分類を行うこと，（２）ユーザの負担が小さくするため少ない教師データで十分な精度の分類を行うこと，（３）計算こすとが少ないこと，の三つの点を考慮してユーザの観点に沿った分類を行う．

まず，観点の特徴はユーザが各クラスに分類するのに寄与した単語の分布に現れると考えた．そこで，ユーザが実際に分類した例から各クラスに対する単語の寄与度を観点抽出手法で算出し，観点行列にまとめる．実際に寄与度は各クラスにおける各単語の出現頻度を用いて算出する．その寄与度の算出法は様々な方法が考えられるため，どの手法がうまく観点の特徴を捉えるか実際に分類手法へ導入して比較検証する．まず，クラスタ内で単語の特徴量の平均値を寄与度とする手法が考えられる．しかし，この手法では一部のサブクラスのみに現れる特徴が埋没してしまうため，サブクラスの特徴を捉えるためには，クラス内で単語の特徴量の最大値を寄与度とする手法も考えられる．さらに，複数のクラスで多くの出現する単語は特定のクラスへ分類するのに寄与しているとは言えないので，当該クラスと他クラス

の比を寄与度とすることもできる．そこで，これらを組み合わせた４通りの観点抽出手法（EM-1，EM-2，EM-3，EM-4）を提案し，比較実験によりその効果を検証する．

次に，観点抽出手法で算出した観点の特徴を文書分類に反映するために，NMF[12][13]などの行列分解を利用した文書分類へ観点行列を導入する２つの分類手法（NMF-DV，NMF-IV）を提案している．NMF-DV は観点行列を行列分解式の基底行列部分に代入し特徴行列を算出する．NMF-IV は NMF における基底行列の初期値に観点行列を導入している．

さらに，単語と熟語を素性として同時に利用することで分類性能の向上を目指した MRNMF-N を提案している．MRNMF-N は NMF-IV に多重解像度 NMF[14] の手法を追加した手法である．このときの多重解像度 NMF では単語を素性とした文書ベクトルを高解像度データ，bigram を素性とした文書ベクトルを低解像度データとしている．

また，語義の曖昧性 [16] を無視した場合，複数の文書に出現する単語がそれぞれの文書で異なった意味で使用されていたとしても一つの素性として扱ってしまう．その際，文書分類では，多義語はクラスタリングにおける特徴として十分に機能しない恐れがある．これは，各クラスにおいて語義の重要度が異なるが，単語の意味による判別が出来ないからである．そのため，本論文では文書分類部分において多義語に対する曖昧性の解消を伴った多重解像度 NMF を提案している．

実験では，実際に文書データを分類することで，提案手法と既存の NMF，その他の NMF 手法 [15] や NMF 以外の手法 [17] との比較を行う．

1.3　論文構成

本論文の構成を以下に示す．

- 第２章 文書分類

  知識ベースによる分類手法，本論文で使用している機械学習による分類手法，文書分類における前処理技術，検索結果の表示手法と具体的な分類手法について述べる．

- 第３章 非負値行列因子分解

非負値行列因子分解 (NMF) を用いた従来の文書分類手法について述べる．

- 第 4 章 NMF を用いたクラスタリングの関連研究
  NMF によるクラスタリングついての関連研究について述べる．

- 第 5 章 提案手法 - 観点抽出 -
  観点と観点を表す寄与度を算出する手法について述べる．

- 第 6 章 提案手法 - 観点行列を用いた文書分類 -
  第 5 章で算出した観点行列を導入した文書分類手法について述べる．

- 第 7 章 実験
  実験や評価方法について説明し，実際に評価実験を行い、その考察について述べる．

- 第 8 章 おわりに
  本研究のまとめと今後の課題について述べる．

第2章　文書分類

多くの他の人工知能のタスクと同様に，テキスト分類には二つのアプローチがある．一つは知識工学 (knowledge engineering) 的なアプローチで，カテゴリに対する専門家の知識を宣言的，あるいは手続き的な分類ルールの形で直接システムにコード化しておくものである．もう一つは機械学習（Machine Learning) 的なアプローチで，あらかじめ分類済みの事例集合から学習を行う汎用的な帰納プロセスが分類器を構築するものである [17]．本章では，知識ベースによる文書分類と機械学習による文書分類について述べる．また，本論文の分類手法では知識ベースではなく機械学習による文書分類を行う．

2.1　知識ベースによる文書分類

文書分類に対する知識工学的なアプローチでは，手作業による分類ルールの作成が中心となる．対象領域の専門家が，あらかじめ与えられたカテゴリにラベル付けされるべき文書の十分条件の集合を定義する．この分類ルールの設定は，全て手作業で作成しなければならないため時間と労力がとても掛かる．

ここでは，テキスト分類に対する知識工学的アプローチの例を一つだけ挙げる．それは，カーネギーグループ社によってロイター社向けに作られた有名なCONSTRUEシステム [18] である．CONSTRUE システムで使われる典型的なルールは次のようなものである．例に挙げるルールは卓球カテゴリか否かを判断するものである．

```
IF((ラケット & 室内) or
   (テーブル & 球技) or
   (ネット & 室内 & 球技) or
   (ラケット & ラバー) or
   (ラケット & 球技 & ¬コート ))
then 卓球
else ¬卓球
```

CONSTRUE システムは，ロイター文書集合の小さな部分集合上でテストした場合には，適合率と再現率が同時に 90% を達成したと報告されている．しかし，ルール作成に長い時間がかかるため，高性能な機械的アプローチの方法が必要である．

2.2 機械学習による文書分類

機械学習によるアプローチでは，あらかじめ分類済みの訓練用文書集合を用いてカテゴリの性質を学習させることで，分類器を自動的に構築する．分類器は，訓練集合上で正しいカテゴリ割当を行う既知の関数を用いて処理されるため教師あり学習を行う．教師なしの処理はクラスタリング (clustering) と呼ばれる．

機械学習による文書分類には様々な手法があるが，本論文では，第 2.5.1 項では Naive Bayes について，第 2.5.2 項では Support Vector Machine について，第 2.5.3 項では多項ロジスティック回帰分析について，そして，第 3 章で NMF について述べる．

2.3 文書分類における前処理技術

自然言語処理における機械学習による文書分類はベクトル空間モデルで各文書をベクトルとして表現し分類処理を行う．文書ベクトルを作成するためには，各文書から素性ごとの特徴量を計算する必要がある．その時に用いる技術について以下で述べる．

2.3.1 特徴抽出

本論文で利用する特徴量は文書中の各名詞と動詞の TF-IDF 値を使用している．TF-IDF 値については 2.3.2.2 節で詳しく述べる．TF-IDF 値を求める前に文書を単語毎に分ける必要がある．そこで，単語毎に文書を分割するには形態素解析という手法を使用する．

2.3.1.1 形態素解析

形態素 (morpheme) とは，意味を担う最小の言語要素のことである．そして，コンピュータで形態素を自動的に制定する処理のことを形態素解析 (morphological

analysis) と呼ぶ [19]．形態素解析には，単語分割，読み振り，品詞付けといった基本的な機能があり，文章を意味のある単語に区切り，辞書を利用して品詞や内容を判別することができる．これはかな漢字変換や，機械翻訳などに用いられる．

形態素とは，文書の要素のうち，意味を持つ最小の単位である．日本語において “私はあの公園で遊んだことがある” では，“私”，“は”，“あの”，“公園”，“で”，“遊んだ”，“こと”，“が”，“ある” がそれぞれ形態素に当たる．この時，“遊”，“ん”，“だ” などの一文字では，それのみでは意味を持たないため，形態素とは呼ばない．英語では原則として文章を単語ごとに区切って書く（分かち書き）ため，形態素ごとに分割することは容易である．

一方，日本語では単語ごとに区切らず続けて書くために，形態素ごとの分割が難しい．例えば，かな漢字変換の場合には，ひらがなのみで与えられた文章を区切る必要があるが，これは辞書を引きながら，いろいろな区切り方を試していくことになる．

この時，辞書にある名詞を形態素として区切ったり，前後の品詞を見て文法的におかしい区切り方は省くなどの処理をするが，複数の解釈が可能な文章もあり，区切り方を一意に決定することはなかなか難しい．特に長文になるほど区切り方の解釈が複雑になるため，ユーザの意図しない漢字変換をしてしまうことが増える．

日本語用の形態素解析システムは各社のワープロソフトやかな漢字変換ソフト (IME) などに内蔵されているほか，単体のソフトとしてはフリーソフトウェアの Chasen（茶筅）や MeCab（和布蕪）などが有名である．最近では，Java 製の比較的新しい形態素解析器として Kuromoji などもある．

2.3.1.2 MeCab

MeCab は，京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンである．

MeCab は言語，辞書，コーパスに依存しない汎用的な設計がとられている．パラメータの推定に Conditional Random Fields（CRF) を用いており，ChaSen が採用している隠れマルコフモデルに比べ性能が向上している．また，平均的に ChaSen，Juman，KAKASI より高速に動作する．

2.3.2 文書ベクトル

2.3.2.1 ベクトル空間モデル

ベクトル空間モデル (vector-space model)[20] は対象となるデータの個々の特徴量を要素とする多次元ベクトルで表現する．各ベクトルでの類似度 (similarity) を定義することにより，問い合わせと類似したものを探し出す方法である．

いま，比較対象の特徴として n 個の属性が備わっており，i 番目の属性を w_i とする．そして j 番目の文書データをベクトル $\vec{D_j}$ と表現する．これらのベクトルが線形独立であれば，n 次元のベクトルが定義できる．このように定義されたベクトルにおいて，j 番目のデータの特徴ベクトルは

$$\vec{D_j} = (d_{j1}, d_{j2}, d_{j3}, \cdots, djn) \tag{2.1}$$

のように表すことができる．ここで，d_{ji} は j 番目データの $\vec{D_j}$ における i 番目の要素であり，その属性値が $w_{d_{ji}}$ である．

各文書と同様に問い合わせも，またベクトルで表される．n 次元のベクトル空間に対するその問い合わせベクトル $\vec{Q}$ は

$$\vec{Q} = (q_1, q_2, q_3, \cdots, q_n) \tag{2.2}$$

のように表される．ここで，q_i は問い合わせベクトルの $\vec{Q}$ における i 番目の要素であり，その属性値が w_{q_i} である．

比較は比較対象の特徴ベクトル $\vec{D}$ と問い合わせベクトル $\vec{Q}$ の類似度を計算することにより行われる．

2.3.2.2 TF・IDF

文書を探すために，文書と単語の関連性の数値演算を行い，その値の高いものを候補とする．そこで用いられる評価値は文書中には重要な単語がどれくらい多く含まれているかを表している．文書中の単語がどの程度重要であるか重み付けに用いられているのが以下に述べる TF・IDF 法である．

この手法は次の 2 つの性質に注目している．

1. 文書に高い頻度で現れる単語は重要である

2. 少ない数の文書にしか現れない単語は重要である

まず，単語出現頻度 (Term Frequency: tf) を考える．単語 t が文書 d に高い頻度で現れるなら，t は d を良く特徴付ける．この考えによる尺度が単語出現頻度，tf(term frequency) である．ある文書 d における単語 t の出現頻度 $tf(d,t)$ は次式で定義される．$n_{t,d}$ は文書 d 内での単語 t の出現数，$\sum_k n_{k,d}$ は文書 d 中の前単語数である．

$$tf(d,t) = \frac{n_{t,d}}{\sum_k n_{k,d}} \tag{2.3}$$

次に，文書出現頻度 (Document Frequency: df) を考える．tf が大きいというのは重要な性質だが，それだけでは十分に文書を特徴付けることはできない．例えば，日本語文書で「は」という助詞はどんな文書でも高い頻度で現れるが，特定の文書を特徴付けないことは明白である．そこで，単語 t が検索対象となる文書集合のうちの少数の文書にしか現れないという性質が重要である．そこで，文書出現頻度 $df(t)$ は単語 t の出現する文書数とする．

$df(t)$ が小さいことが単語 t の文書を特徴付ける能力が高いことを表すので，実際にはこの逆数を log と文書集合中の文書総数 N による正規化した idf(inverse document frequency) を用いる．

$$idf(t) = log\frac{N}{df(t)} + 1 \tag{2.4}$$

単語 t について，その単語が文書内に出てくる回数とそれが全文書内に占める割合の積を計算することで，その単語の重要性と，その出現頻度によって文書の重要性を表すことできる．すなわち tf が大きく，df が小さいならば，単語 t は文書 d を真に特徴付けるといえる．この考え方を数値の尺度として表現したのが TF・IDF 値である．つまり，文書 d におけるキーワード t の重み TF・IDF 値は式 (2.5) で求まる．

$$\text{文書 } d \text{ にある単語 } t \text{ の } TF \cdot IDF \text{ 値} = tf(d,t) \cdot idf(t) \tag{2.5}$$

2.3.2.3 N-gram

N-gram とは文字あるいは形態素，または品詞情報が N 個つながった組合せのことである．例えば,「国境の長いトンネル」という文字列を，$N = 2$ で文字を単位として区切ると，図 2.1(a) の組合せになる．最初に [国-境] の 2 文字のペア，次に 1 文字右に移動して [境-の] の 2 文字のペアとなり，以下同様にして，最後に [ネ-ル] の組合せとなる．なお，N が 2 の場合はバイグラム (bi-gram，bigram) といい，N が 3 の場合はトライグラム (tri-gram，trigram) という．

形態素単位で分割し，N を 2 に取るならば，図 2.1(b) の組合せとなる．また，品詞情報でバイグラムを作れば,「国境」は名詞,「の」は助詞,「長い」は形容詞，そして「トンネル」は名詞と判断されるので，バイグラムは図 2.1(c) の組合せとなる．

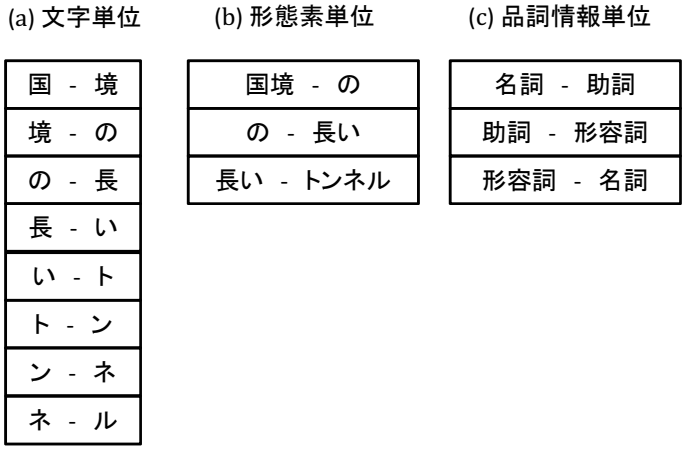

図 2.1: bigram

2.3.2.4 ベクトル間の類似度

先に述べたように，ベクトル空間モデルにおいて，ベクトルの比較を行うためにはベクトル間の類似度を定義しなければならない．類似度の尺度としては様々なものがあるが，ここではベクトル間の余弦を用いた手法を採用する．

- 類似度

  余弦の値を利用して類似度を求める．j 番目データの特徴ベクトル D_j と問い合わせベクトル Q の類似度 $sim(Dj, Q)$ は以下のようになる．

$$
\begin{aligned}
sim(D_j, Q) &= \frac{\vec{D_j} \cdot \vec{Q}}{\left|\vec{D_j}\right| \left|\vec{Q}\right|} \\
&= \frac{\sum_{i=1}^{n}(d_{ji} q_i)}{\sqrt{\sum_{i=1}^{n}(d_{ji})^2 \sum_{i=1}^{n}(q_i)^2}} \qquad (2.6)
\end{aligned}
$$

$sim(D_j, Q)$ の値の範囲は 0 以上 1 以下であり，1 に近づくほど類似度が高くなる．

2.4 検索結果表示手法

インターネット上の記事などを検索するには，Google や Yahoo!などの検索エンジンを利用するのが一般的であるが，検索結果の中にユーザの求めるページが高い順位で表示されない問題 [3] が指摘されている．

その問題を解決するための様々な手法が提案されている．その一つが第 2.4.1 項で説明するランキング手法による表示手法である．これは検索結果の各 Web ページに対して重要度によるランキングを計算し順位の高いページから表示する手法である．また他の手法として第 2.4.2 項で説明する検索結果をクラスタリングで表示する手法がある．

2.4.1 ランキング手法による表示

ランキング手法とは検索結果である各 Web ページの表示される順番をユーザが求める Web ページが上位に表示されるように順位付けする手法である．この手法の代表的な方法として PageRank[4][5] や HITS[6] がある．

PageRank は「Web ページは，他の重要なページからリンクされているページは重要である」というポリシーに基づき各ページに得点を付け順位付けを行う．

HITS では権威とハブとを定義している．権威とはそのページがリンクされているページ数のことであり，ハブとはそのページがリンクしているページ数のことである．そして，この HITS は「ページは，良い権威を指しているなら，良いハブに

なり，良いハブから指されているページなら良い権威となる」というポリシーに基づき各ページに権威得点とハブ得点を付け順位付けを行う．

2.4.2 検索結果のクラスタリング表示

検索エンジンの検索結果をユーザが利用しやすいように，グループに分類して表示することで，ユーザは効率良く目的の文書を探すことができるようになる．また，ランキング手法による表示手法では，使用する検索クエリによってはユーザが意図しない種類のページが混在した状態の検索結果になる場合がある．それに比べて，検索結果を分類して表示する方法はユーザは求める種類のカテゴリ内の検索結果のみを探すことができるという利点がある．

検索結果を分類して表示する方法 [7][8][9][10][11] は多く提案されている．その内のいくつかの手法を説明する．実際に公開されている検索エンジンには Carrot2[25] などがある．

既定のカテゴリによる手法

安形らは日本十進分類法 (NDC)[26] や Yahoo![2] のカテゴリなどの既定のカテゴリ名を使用して文書ベクトルの類似度による分類を行った [7]．

学習フェーズではすでに日本十進分類法で分類されている文書を用いて各カテゴリに対する各単語の重みを学習する．分類フェーズでは，Web ページの文書に対して形態素解析を行い，形態素をもとに特徴ベクトルを算出し，Web ページの特徴ベクトルと各カテゴリの特徴ベクトルの類似度による自動分類を行っている．

多視点クラスタリング

松村らは多視点クラスタリング [8] という手法を提案している．松村らは「ユーザの検索意図や検索対象に対する理解度に対して適切となるような検索結果の表示のされ方」を 1 つの「視点」として定義している．そして，クラスタの構造は変更せずにクラスタのラベルを切り替えることで，ユーザが検索結果クラスタを複数の視点で閲覧可能としている．

ラベリング手法として手法 1〜手法 4 の 4 つを提案している．手法 1 のラ

ベルはクラスタ内文書のいくつかに共通して出現し，Web 上で使用頻度が高い単語としている．手法 2 のラベルはクラスタ内文書のいくつかに共通して出現し，Web 上で使用頻度が低い単語としている．手法 3 のラベルは Web 上で使用頻度が低く，検索クエリに対する検索結果集合中では使用頻度が高くなる単語としている．手法 4 のラベルは Web 上で使用頻度が高く，検索クエリに対する検索結果集合中では使用頻度が低くなる単語としている．

2.5 分類手法

ここでは，提案手法の分類手法と比較を行う NMF 以外の 3 種類の分類手法とついて述べる．

2.5.1 Naive Bayes

ナイーブベイズ (NB : Naive Bayes)[17] は確率に基づいた分類器の一つである．ナイーブベイズによる分類器は，文書 d がカテゴリ c に属する確率 $P(c|d)$ をベイズの定理に適用して計算し求める手法である．

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{2.7}$$

周辺確率 $P(d)$ はカテゴリによらず一定なので，必ずしも計算する必要はない．しかし，$P(d|c)$ を計算するためには，文書 d の構造について何らかの仮定を置く必要がある．索引語数 n 個である文書 d の特徴ベクトルを $d = (w_1, w_2, \cdots, w_n)$ とするとき，最も一般的な仮定は，すべての索引語の特徴量は独立であるというものである．従って，式 (2.8) が成り立つ．

$$P(d|c) = \prod_i P(w_i|c) \tag{2.8}$$

この仮定から得られる分類器をナイーブベイズ分類器と呼ぶ．単純な分類器，つまりナイーブなベイズ分類器と呼ばれる理由は，座標を表す全ての索引語の特徴が独立であるという仮定が厳密に検証されておらず，そして明らかにその仮定は偽であることも多いからである．しかし，各要素が独立ではなく何らかの依存関係があ

ると仮定した確率的モデルを使用する試みは，これまでのところ性能面で目立った改善をもたらしてはいない．ナイーブベイズ分類器のこの意外な頑健性については一定の理論的検証が，Domingos and Pazzani(1997) で述べられている．

2.5.2 Support Vector Machine

サポートベクターマシン (SVM：Support Vector Machine) アルゴリズムは，テキスト分類問題に対して非常に高速かつ効果的である．

幾何的に言えば二値分類を目的とした SVM は，素性空間内でカテゴリの正例，負例を表す点を分離する超平面とみることができる．分類超平面は学習により，既知の正例と負例を最大マージンで分類するような一意に決まる超平面が選ばれる．マージンとは，超平面と，正例と負例の点集合の中でその超平面に最も近い点との距離である．図 2.2 は，2 次元における最大マージン超平面の例である．

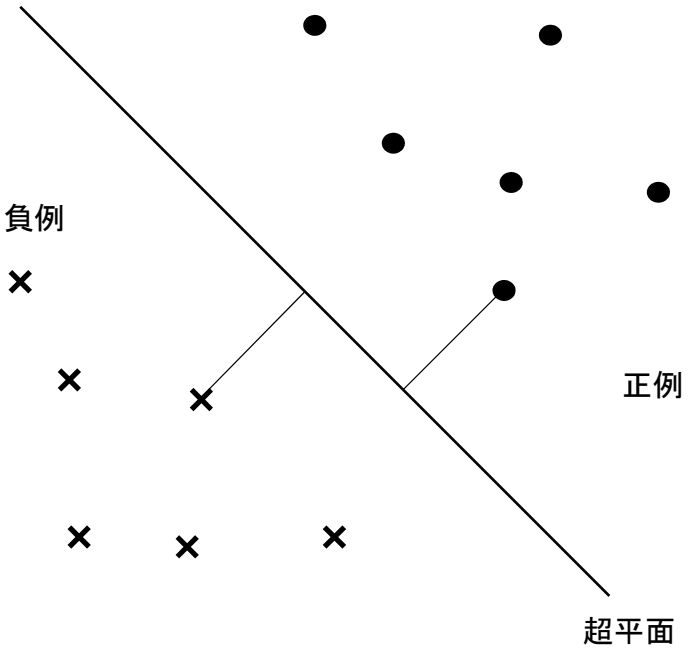

図 2.2: 二値線形 SVM

単純な SVM において，SVM 超平面はサポートベクター (support vector) と呼ばれる訓練事例の中の比較的小さな部分集合だけで完全に決定される．それ以外の訓練データは，最終的な分類器にまったく影響しない．つまり，特異な値が学習データに含まれていても，その特異なデータが分離超平面の計算に使用されない場合があり，過学習を避けることができる．この点に関して，SVM アルゴリズムは様々な

カテゴリ分類アルゴリズムの中で独特な性質であり，重要な優位性を持っている．であるといえる．SVM 分類器は，素性空間の次元に関係なく良い性能を示す．また，パラメータの調節も必要ない．

2.5.3 多項ロジスティック回帰分析 (MLR)

回帰分析とは，独立変数と従属変数との相関関係などを分析する手法である．例えば，変数 y と変数 x の 2 変数があり，この 2 変数が関数 f により $y = f(x)$ などのように定量的な関係を持つことを分析する事である．この時，関数 f が

$$f(x) = \alpha x + \beta \tag{2.9}$$

であったり，x が n 次元のベクトルで関数 f が

$$f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \tag{2.10}$$

のように線形モデルで表せる場合は線形回帰という．特に，式 (2.9) のように独立変数 x 一つと変数 y との関係を表すモデルを単回帰分析，式 (2.10) のように複数の独立変数で表すモデルを重回帰分析という．

この回帰分析を文書分類に用いる際には，独立変数 x を各文書の特徴ベクトルである文書ベクトルとし，従属変数 y を各クラスタへの確率変数と考える．単純に式 (2.10) の線形回帰分析で文書分類を考えた場合は，事前に教師データを使用して各係数 α を学習することで各クラスタへの分類器を作成することも可能である．

回帰分析の一つにロジスティック回帰分析 (logistic regression) がある．このロジスティック回帰分析は多変量を対象にした線形回帰分析の一種である．

文書分類を目的にロジスティック回帰分析を説明する．ある文書を表す文書ベクトルが $x = (x_1, \cdots, x_n)$ であり，その文書がクラスタ A である確率 p を式 (2.11) の関数 F でモデル化する．

$$p(x) = Pr\{\text{ クラスタ } A | x_1, x_2, \cdots, x_n\} = F(x_1, \cdots, x_n) \tag{2.11}$$

ここで，独立変数である文書ベクトル x の各要素を線形モデルで合成した合成変数 q と，その合成変数 q を用いたロジスティック関数を考えるとそれぞれ式 (2.12) と式 (2.13) になる．

$$q = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{2.12}$$

$$F(q) = \frac{\exp(q)}{1+\exp(q)} = \frac{1}{1+\exp(-q)} \tag{2.13}$$

式 (2.13) のロジスティック関数を用いてモデル化したものが式 (2.14) である．

$$p(x) = Pr\{\text{ クラスタ } A|x_1, x_2, \cdots, x_n\} = \frac{\exp(q)}{1+\exp(q)} = \frac{1}{1+\exp(-q)} \tag{2.14}$$

この時，図 2.3 のように，範囲が 0〜1 の確率 $p(x)$ が範囲が $-\infty$〜∞ である合成変数 q のロジスティック関数でモデル化したものがロジスティック回帰モデルである．

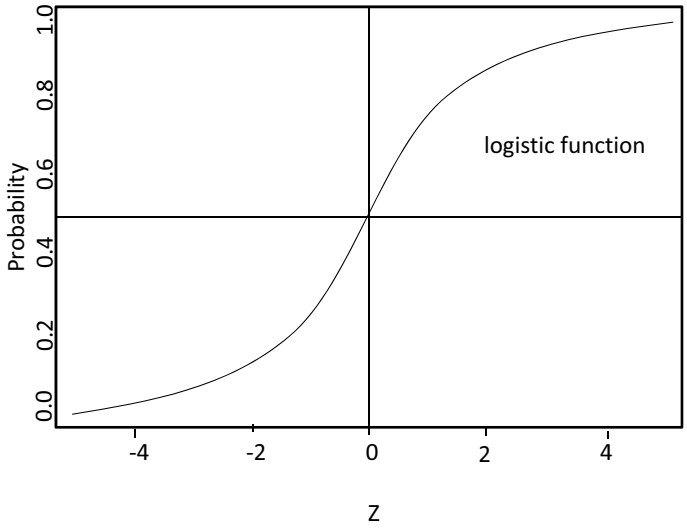

図 2.3: ロジスティック関数

式 (2.14) を整理すると，重回帰モデルである式 (2.15) になる．

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n \tag{2.15}$$

このクラスタ A である確率 q を求めるモデルの場合は 2 値分類である．3 つ以上のクラスタに分類する場合は多項に拡張した多項ロジスティック回帰分析を用いる．例えば，3 つのクラスタに分類する場合を考える．クラスタの種類を A，B と C とし，それぞれへの確率を p_A，p_B と p_C とする時，多項ロジスティック回帰分析は式 (2.16) と式 (2.17) になる．

$$\log\left(\frac{p_A}{p_C}\right) = \alpha_A + \beta_{1A}x_1 + \beta_{2A}x_2 + \cdots + \beta_{nA}x_n \tag{2.16}$$

$$\log\left(\frac{p_B}{p_C}\right) = \alpha_B + \beta_{1B}x_1 + \beta_{2B}x_2 + \cdots + \beta_{nB}x_n \tag{2.17}$$

第3章　非負値行列因子分解

本研究では，文書分類に非負値行列因子分解を用いるため，この章では非負値行列因子分解について述べる．本論文の以降では非負値行列因子分解を英語表記の略称である NMF と表記する．

3.1　非負値行列因子分解（NMF）

3.1.1　NMF の概要

NMF[12] とは，Non-negative Matrix Factorization の略であり，第 2.3.2 節で説明したベクトル空間モデルで表現されたデータを次元縮約することで，クラスタリングを行う．文書クラスタリングにおいて NMF は式 (3.1) のように m 個の文書データと n 個の索引語から作られる $n \times m$ の索引語文書行列 X を $n \times k$ の基底行列 U と $k \times m$ の特徴行列 V^T の積の形に分解することにより文書データを次元圧縮することができる．ここで，k はクラスタ数である．

$$X = UV^T \tag{3.1}$$

NMF では行列 X，U と V の各要素は非負値である．U は索引語を特徴量とする各クラスタの基底ベクトルを表し，V^T は各文書と各クラスタの関連度を表している．

つまり，n 次元の文書データである索引語文書行列 X が k 次元の文書データである特徴行列 V^T へと次元圧縮される．NMF を文書クラスタリングへ適用する際には次元圧縮後行列である特徴行列 V^T を利用する．特徴行列 V^T の h 行目の要素の値が，各文書と h 番目のクラスタとの関連度の大きさを表している．そのため，i 番目の文書データのクラスタは式 (3.2) で得られる．そのため，式 (3.2) のように

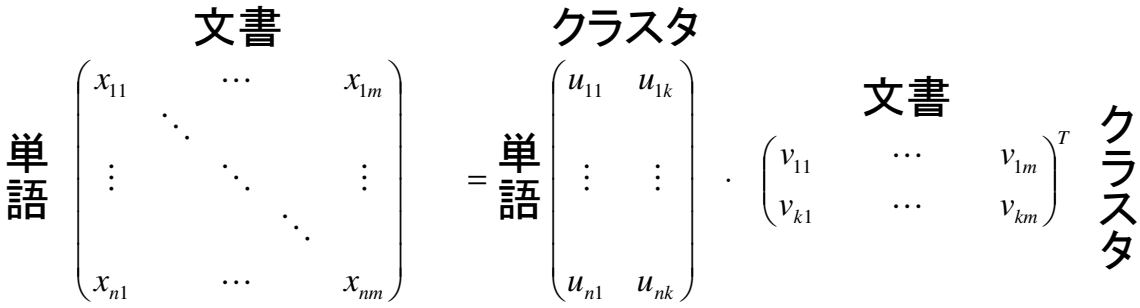

図 3.1: NMF による行列分解

関連度が一番大きいクラスタに i 番目の文書データを分類することができる．ここで，v_{ih} は行列 V の i 行 h 列の要素を表す．

$$\arg\max_{h} \ v_{ih} \tag{3.2}$$

基底行列 U と特徴行列 V への分解は NMF の目的関数である式 (3.3) の J を最小にするような基底行列 U と特徴行列 V を推定することで求まる．

$$J = ||X - UV^T||_F \tag{3.3}$$

そして，ラグランジュの未定乗数法を用いて式 (3.3) の J を最小にする基底行列 U と特徴行列 V の乗算型更新式を求める．r を反復更新の更新回数として式 (3.4), (3.5) のように表される．

$$v_{ij}^{(r+1)} \leftarrow v_{ij}^{(r)} \frac{(X^T U)_{ij}}{(VU^T U)_{ij}} \tag{3.4}$$

$$u_{ij}^{(r+1)} \leftarrow u_{ij}^{(r)} \frac{(XV)_{ij}}{(UV^T V)_{ij}} \tag{3.5}$$

ここで，$u_{ij}^{(r)}$ と $v_{ij}^{(r)}$ はそれぞれ更新回数 r 回目である U と V の i 行 j 列の要素を表し，$(X)_{ij}$ は行列 X の i 行 j 列の要素を表す．

また，各繰り返し後には発散を防ぐためと各基底を単位ベクトルにするために基底行列 U を以下の式 (3.6) に従い正規化を行う．

$$u_{ij} \leftarrow \frac{u_{ij}}{\sqrt{\sum_i u_{ij}^2}} \tag{3.6}$$

通常，基底行列 U と特徴行列 V の初期値 $U^{(0)}$ と $V^{(0)}$ はランダムな値を与えることで作成される．

3.1.2 NMF のアルゴリズム

NMF による文書クラスタリング

1. 索引語文書行列 X を入力
2. 初期値 $v_{ij}^{(0)}$，$u_{ij}^{(0)}$ を乱数で与える
3. 式 (3.4)，(3.5) により $v_{ij}^{(r)}$，$u_{ij}^{(r)}$ を更新
4. 一定回数まで 3. を繰り返す
5. 式 (3.2) で各文書のクラスタを判定

3.1.3 NMF の問題点

初期値に関する問題点

NMF では，$U^{(0)}$ と $V^{(0)}$ の値によって，最終的に得られる $U^{(R)}$ と $V^{(R)}$ は大きく異なる．ここで R は最大更新回数とする．つまり，$V^{(R)}$ はクラスタリング結果を表しているため，クラスタリング結果は初期値 $U^{(0)}$ と $V^{(0)}$ に依存していると言える．

ここで，新納等が提案しているピンポン型クラスタリング [21][22] の結果から，新納等が提案しているピンポン型クラスタリング [21][22] は NMF 以外のクラスタリング手法と NMF の繰り返しにより NMF における初期値を洗練させているとみなすことができる．その結果，高いクラスタリング精度が可能となっている．それにより，初期値 $U^{(0)}$ により精度の高い初期値を与えることで，より精度の高いクラスタリング結果を得ることができると期待できる．また，NMF の初期値の設定方法として堀田らが提案している特異値分解を利用した手法 [23] などもある．

局所解に関する問題点

NMF では，式 (3.4)，(3.5) による繰り返しでは式 (3.3) の J に対する局所最適解に収束するが，文書分類の観点からみると最適解であるとは限らない．

つまり，NMF による文書クラスタリング [13] はある程度の成果が確認できるが，その繰り返しによる収束した結果である局所解が必ずしもユーザが望む分け方としての良いクラスタリング結果であるとは限らない．なぜなら，式 (3.4)，(3.5) の目的は式 (3.3) における J を最小にするであるのに対し，初期値が異なるクラスタリング結果を比較すると J の小さい方が必ずしも良いクラスタリング結果であると判断できないからである．

そのため，精度の良い分類のためには既存 NMF の目的関数を改良することで，繰り返しによるクラスタリング結果がなるべくユーザの望む分け方としての最適なクラスタリング結果に近づくような目的関数にする必要がある．

3.2 NMF-I：初期値に対する制約

NMF では第 3.1 節で挙げたように収束結果が初期値 $U^{(0)}$ と $V^{(0)}$ に依存するという問題が存在する．一般的には NMF での初期値は乱数で与えるが，単純な乱数ではクラスタリング結果が悪い局所解に収束するような初期値となる可能性がある．そこで，U の初期値に教師制約を追加することで良好な初期値による更新を行い分類性能の向上を図る．

まず既知である教師文書ベクトルの各クラスタにおける平均ベクトルを求め，それを平均教師ベクトルとする．その平均教師ベクトルを各クラスタ毎に並べ基底とした初期基底行列 U_s を考えた場合，NMF での理想的な文書分類がなされた場合の基底行列の収束値は U_s に近いものであろうとの期待に基づき，この初期基底行列 U_s を教師あり NMF における基底行列 U の初期値とする手法を以前提案した [24]．初期基底行列 U_s は式 (3.7) で与える．

$$U_s = X_{train}(V_{train}^T)^+ \tag{3.7}$$

ここで，教師データ数を t とした時，$X_{train}(n$ 行 t 列$)$ は教師データのみの教師文書行列であり，$V_{train}^T(k$ 行 t 列$)$ は各文書の正解クラスタに対応する要素を 1 としてそれ以外の要素を 0 とした教師特徴行列である．また，“+” は擬似逆行列である．

初期基底行列 U_s を U の初期値 $U^{(0)}$ として既存 NMF と同様の更新式を実行する手法を，以降は NMF-I，(NMF with Initial basis value by training data) と呼ぶ．

3.3 NMF-S：局所解に対する制約

第 3.1 節で挙げた局所解に関する問題に対処するため，既存 NMF の目的関数に既知である教師データに対する分類が成功する目的を追加し多目的の最適化を行う手法を以前提案した [24].

NMF では特徴行列 V がクラスタリング結果を表している．そのため，教師データが全て正しいクラスタへと収束しているクラスタリング結果を表す教師特徴行列 V_s を考えた場合，クラスタリング結果である V を V_s に近づくような方向が，教師データに対する分類が成功する方向であると考えられる．そこで，クラスタリング結果を表す V を教師特徴行列 V_s に近づけるような教師制約を追加する．

この手法を NMF-S，(NMF with Supervised constraint) と呼ぶ．NMF-S では式 (3.8) を目的関数とする．

$$J_s = ||X - UV^T||^2 + \mu||L * (V_s - V)||^2 \tag{3.8}$$

ここで，$(V_s)_{ij}$ は既知データであり正解クラスタならば 1，既知データであり正解でないクラスタか未知データならば 0 とする．そして行列 $L(m$ 行 k 列) は既知データならば 1，未知データならば 0 とする．また “$*$” は要素毎の乗算とし，μ は教師制約項に対する重みである．

そして，ラグランジュの未定乗数法を用いて式 (3.8) の J_s を最小にする基底行列 U と特徴行列 V の新たな乗算型更新式を求める．U の更新式は既存の式 (3.5) と同じで，V の更新式は式 (3.9) となる．

$$v_{ij}^{(r+1)} \leftarrow v_{ij}^{(r)} \frac{\{X^T U + \mu(L * V_s)\}_{ij}}{\{VU^T U + \mu(L * V)\}_{ij}} \tag{3.9}$$

さらに，NMF-S では更新毎に特徴行列 V を以下の式 (3.10) に従い正規化を行う．これは，特徴行列 V の各要素である文書と各クラスタの関連度に上限を付け効率的に教師データを正しいクラスタへ収束させるためである．

$$v_{ij} \leftarrow \frac{v_{ij}}{\sqrt{\sum_j v_{ij}^2}} \tag{3.10}$$

NMF-S の教師制約により既知である教師データが正しいクラスタへと収束する方向へも収束方向を多目的に最適化されると，教師データが正しいクラスタへ収束

するのに影響されて間接的に未知データも正しいクラスタへ収束するのではないかと期待された．この期待は以下の理由に基づくものである．NMF において基底行列 U と特徴行列 V の更新式はお互いに影響している．そのため，V の教師部分などの一部分が理想的な値に近づけば各クラスタの基底ベクトルである基底行列 U も理想的な値に近づくと考えられる．さらに各クラスタの基底ベクトルが理想的な値に近づくと未知データに関してもよりよいクラスタリング結果に収束すると考えられる．以上が先ほどの期待に対する理由である．

第4章　NMFを用いたクラスタリングの関連研究

この章ではNMF を用いた分類手法に関連した研究について述べる．

4.1　ピンポン型文書クラスタリング

ピンポン型クラスタリング [21] ではまずNMF により，あるクラスタリング結果を導く．次にLBR によりそのクラスタリング結果を改善する．改善されたクラスタリング結果を利用して，NMF の初期値となる行列 V_0 と U_0 を作成し，NMF を実行する．この処理を繰り返す．

4.1.1　LBR

LBR[21][22] はグラフスペクトル理論を用いたクラスタリング手法である Mcut の結果を修正する目的で提案された．グラフスペクトル理論を用いたクラスタリング手法では，その結果に「ねじれ現象」と呼ばれる不具合が生じることがあり，それを解消するために提案された手法がLBR である．

4.2　Hybrid Method

Hybrid Method[27] はNMF の初期値である V_0 に乱数ではなく，pLSI によるクラスタリング結果をNMF の初期値として与える．

これは，NMF のクラスタリング結果が初期値である V_0 と U_0 に大きく依存するためである．つまり，初期値に乱数を与える標準の使い方では，あまり精度の良くないクラスタリング結果にしかならない．そこで，pLSI によるクラスタリング結果を用いることで，NMF の初期値を整理する．

4.3 Semi-Supervised NMF

Semi-Supervised NMF(SSNMF)[15] は H.Lee 等が提案した半教師あり NMF であり，3.1.3 節で挙げた目的関数に関する問題を解決する手法の一つである．そのため，既存 NMF の目的関数に対して，我々とは異なる教師情報を含んだ制約項を追加することで多目的の最適化を行なっている．SSNMF における目的関数を式 (4.1) に示す．

$$J_{ss} = ||X - UV^T||^2 + \lambda||L * (Y - WV^T)||^2 \tag{4.1}$$

式 (4.1) において Y は教師データの正解クラスタがラベル付けされた $k \times\ m$ のラベル行列であり，教師データの正解クラスタを 1 としそれ以外を 0 としている．W は $k \times\ k$ の第二項目における基底行列であり，クラスタ間の関係を表している．そして，L は式 (4.2) のように教師データのみを制御するための $k \times\ m$ の重み行列である．また，λ は第二項目に対する重みである．

$$L_{ij} = \begin{cases} 0.001 & \text{if } \mathrm{Y_{ij}} = 1 \\ 1 & \text{if } \mathrm{Y_{ij}} = 0 \\ 0 & \text{if } \mathrm{Y_{ij}} \text{ is unknown.} \end{cases} \tag{4.2}$$

式 (4.1) の第二項目が SSNMF における制約である．この制約により，基底行列 W と特徴行列 V^T の積がラベル行列 Y に近づく方向へ特徴行列 V の収束方向は制御される．つまり，特徴行列 V における文書とクラスタの関連度に対して似ているクラスタ同士の関連度を上げる効果があると考えられる．

さらに，SSNMF では最終更新後の $V^{(R)}$ に対して K-means を適用している．その結果をクラスタリング結果としている．

SSNMF における更新式は式 (4.3)〜(4.5) となる．

$$v_{ij}^{(r+1)} \leftarrow v_{ij}^{(r)} \frac{(X^TU)_{ij} + \lambda(L * Y^T)W}{(VU^TU)_{ij} + \lambda(L * VW^T)W} \tag{4.3}$$

$$u_{ij}^{(r+1)} \leftarrow u_{ij}^{(r)} \frac{(XV)_{ij}}{(UV^TV)_{ij}} \tag{4.4}$$

$$w_{ij}^{(r+1)} \leftarrow w_{ij}^{(r)} \frac{((L * Y)V)_{ij}}{(L * (YV^T)V)_{ij}} \tag{4.5}$$

このSSNMFと第3.3節で説明したNMF-Sの相違点は二つ存在する．一つ目はSSNMFの制約ではVは教師データにおいて正解クラスタと正解クラスタに類似するクラスタの両方の関連度を上げるが，NMF-Sでは正解クラスタのみの関連度を上げる．二つ目はSSNMFにおける基底行列Wは既存のNMFにおける基底行列Uや特徴行列Vと同じように乗算型更新式で求める必要があり，初期値$W^{(0)}$はランダムな値を与える．そのため第3.1.3節で挙げた初期値に関する問題点に関して更に初期値に対する依存度を高めてしまう恐れがある．

4.4 Kernel NMF

Kernel NMF(KeNMF)[28]はD.Zhang等が提案したNMFにカーネルトリックの概念を導入した手法の一つである．このKeNMFはカーネル関数を使用することで，対象データの特徴空間で張られる特徴ベクトルを非線形変換して，その写像空間でNMFによる識別を行う手法である．

$$J_{ker} = ||K(X) - YV^T||_F \tag{4.6}$$

ここで，関数Kはカーネル関数であり，行列Yはカーネル関数により写像された特徴空間における各文書ベクトルの基底ベクトルである．

4.5 多重解像度NMF(MRNMF)

多重解像度NMF[14] (MRNMF : Multi Resolution Non-ngegative Matrix Factorization)は，東工大の伊東らが提案した手法である．伊東らは高フレームレート低空間解像度と低フレームレート高空間解像度の異なる2つの動画像から，正確な細胞形態と詳細な細胞活動系列を抽出し，両解像度から得られる特徴を同時に利用するために提案したNMF手法の一種である．伊東らは多重解像度NMF-model1とmodel2と名付けられた2通りの手法を提案した．

4.5.1 多重解像度NMF-model1

多重解像度NMF-model1は第4.1項や第4.2項で説明したピンポン型文書クラスタリングやHybird Methodと同じ様に2種のクラスタリング手法を段階的に順次

行っている．伊東らの多重解像度 NMF-model1 では同対象物に対しての異なる 2 つの解像度データに対してそれぞれ NMF を 1 回実施している．まず，式 (4.7) で高解像度データに対する NMF を行う．そこで推定された基底行列 U を式 (4.8) の基底行列 U に代入し次の低解像度データに対する NMF において既知情報として扱い，特徴行列 V_2 に対してのみ更新式を適用し反復計算による推定を行う．

$$J_1 = ||X - UV_1||^2 \tag{4.7}$$

$$J_2 = ||Y - MUV_2||^2 \tag{4.8}$$

ここで，行列 X が高解像度データであり，行列 Y が低解像度データである．

また，文書分類に対するピンポン型の手法は新納らの研究 [21] から組み合わせる手法によっては異なる手法のクラスタリング手法を繰り返し行うことで良い分類結果になることが示されている．

4.5.2 多重解像度 NMF-model2

多重解像度 NMF-model2 は異なる 2 つの解像度データに対する NMF を同時に実施する．そのために目的関数である式 (4.9) を提案している．

$$J = (1 - \lambda)||X - UV_1||^2 + \lambda||Y - MUV_2||^2 \tag{4.9}$$

各解像度における基底行列は共通している．また，低解像度の行列因子分解における行列 M は，ビギニングを表現する行列である．多重解像度 NMF-model2 は，異なる解像度のデータを用いて同時最適化を行うため各解像度で逐次的に行列因子分解を行い各要素の推定を行う多重解像度 NMF-model1 と比べて，高い推定精度が見込まれる．

第5章　提案手法：観点抽出

この章では人が文書を分類する際に用いる観点について説明し，その観点の特徴を算出する手法を提案する．

5.1　人の観点

本論文では，ユーザの観点とはユーザが文書分類の際に用いる無意識下の指標であると考える．また，通常，人の観点は人によって異なっている．そのため，表5.1の様に分類結果も人によって異なっている．

表5.1は実際に6人の被験者にマルチラベル文書を任意に分類してもらった分類結果である．使用した文書集合はWeb朝日で公開されている記事の中で政治とスポーツの両カテゴリが付与されたマルチラベル文書の19記事である．被験者には各記事を政治かスポーツの一方のクラスに分類してもらった．結果である表5.1を確認すると，6人全員が同じクラスに分類した記事は19記事中6記事であった．また，経済とスポーツのマルチラベル文書の29記事を同様に片方のクラスに分類してもらった場合は29記事中10記事を同じクラスに分類した．このように，観点が異なる人が文書分類するとそれぞれの結果は同じにはならない．

ユーザが文書分類を行う際の観点を求めるには観点を明示的に表現することが望ましいが，ユーザの観点は人の感覚に依るところが大きいためそれは困難である．そこで，ユーザが分類した教師文書から観点の特徴を抽出し，それを分類に反映させる．次節の5.2節でその手法について述べる．

5.2　観点抽出手法

本論文では観点の特徴は教師文書中の単語の分布に現れると考えた．そこで，各クラスに対する単語の寄与度を観点抽出手法で算出し，全寄与度の集合を近似的な観点として表現する．

表 5.1: 分類結果の違い

Docs-No	Tester01	Tester02	Tester03	Tester04	Tester05	Tester06
doc-01	政治	政治	政治	スポーツ	政治	政治
doc-02	スポーツ	スポーツ	政治	政治	スポーツ	政治
doc-03	スポーツ	スポーツ	スポーツ	政治	スポーツ	スポーツ
doc-04	政治	スポーツ	スポーツ	スポーツ	スポーツ	スポーツ
doc-05	政治	政治	スポーツ	スポーツ	スポーツ	スポーツ
doc-06	スポーツ	スポーツ	スポーツ	政治	スポーツ	スポーツ
doc-07	政治	政治	政治	政治	政治	スポーツ
doc-08	スポーツ	政治	スポーツ	政治	スポーツ	スポーツ
doc-09	政治	政治	政治	政治	政治	政治
doc-10	政治	政治	スポーツ	政治	スポーツ	政治
doc-11	スポーツ	スポーツ	スポーツ	スポーツ	スポーツ	スポーツ
doc-12	スポーツ	スポーツ	政治	政治	スポーツ	政治
doc-13	政治	政治	政治	政治	政治	政治
doc-14	政治	政治	政治	政治	政治	政治
doc-15	政治	政治	スポーツ	政治	スポーツ	政治
doc-16	政治	政治	政治	政治	政治	政治
doc-17	スポーツ	スポーツ	スポーツ	政治	スポーツ	政治
doc-18	政治	政治	政治	政治	政治	政治
doc-19	政治	スポーツ	政治	政治	スポーツ	スポーツ

提案手法の流れを以下に示す．まず，観点抽出手法により寄与度を算出する．全寄与度の値を観点行列 $U_m \in \mathbb{R}^{w \times k}$ として扱う．ここで，w は単語数，k はクラス数である．次に，行列分解を利用した分類手法に観点行列 U_m を導入し，ユーザの観点を反映した分類を行う．観点抽出手法として 4 つ（EM-1，EM-2，EM-3，EM-4）の方法を提案し検証している．

各観点抽出手法の説明では任意のクラス A に対する単語 t の寄与度の算出式について述べる．また，$D = \{d_1, d_2, ..., d_n\}$ を全文書ベクトルの集合，n は文書数とする．その時，$D_A(\subseteq D)$ はクラス A 所属の文書ベクトル集合であり，$\overline{D}_A$ は D_A の補集合である．特徴量である $f(D, d, t)$ は文書 d における単語 t に対する TF・IDF 値を使用する．ここで，任意のクラスに対する全単語の寄与度を要素としたベクトルを観点ベクトルとし，全クラスの観点ベクトルを整列させて作成する行列を観点行列とする．

5.2.1 EM-1 クラス内での平均特徴量

EM-1 での寄与度を計算する方針は以下の通りである．まず，クラス内のあらゆる文書で多く出現する単語はクラスに対する寄与度が大きい．次に，一つの文書内

で多く出現していてもクラス内で出現する文書は少ない単語はクラスに対する寄与度は小さい．

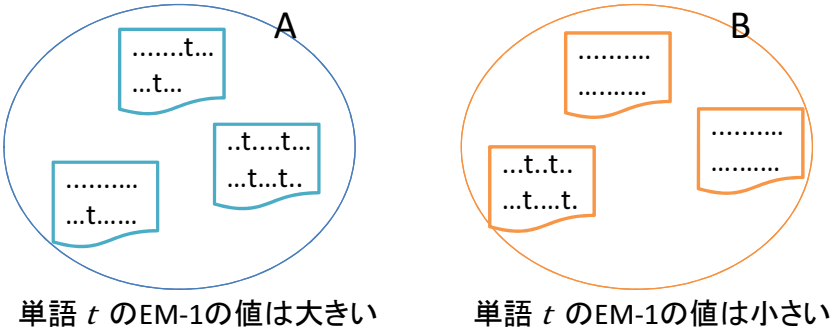

図 5.1: EM-1 当該クラス内での平均特徴量

その方針に基づいて EM-1 は，D_A に属する全文書における単語 t に対する特徴量の平均値をクラス A に対する単語 t の寄与度とする．寄与度の算出式は式 (5.1) である．EM-1 では，当該クラス内のあらゆる文書で多く出現する単語が観点を表す単語として重要であると仮定した手法である．

$$\underset{d \in D_A}{\text{mean}}\ f(D, d, t) \tag{5.1}$$

クラス A を当該クラスとした時，図 5.1 を用いて説明すると，クラス A 内の文書に出現する単語 t の特徴量をクラス A 内において平均化した値を寄与度とする．そのため，EM-1 では，単語 t がクラス A に属するあらゆる文書に沢山出現する時，寄与度は高くなる．また，EM-1 では計算中の当該クラス以外のクラスについては考慮しない．

5.2.2 EM-2 平均特徴量のクラス間比率

EM-1 では当該クラスのみでクラスに対する単語の特徴を計算している．しかし，他のクラスに対する寄与度も高い単語は当該クラスへの寄与度は小さい方が望ましいとの方針で EM-2 を考える．つまり，他のクラスでも多く出現する単語の寄与度を小さくする．

その方針を満たすため EM-2 は，EM-1 で求めた寄与度に対してクラス間の比率を使用する．EM-2 における寄与度は式 (5.2) で算出する．EM-2 は，当該クラスで

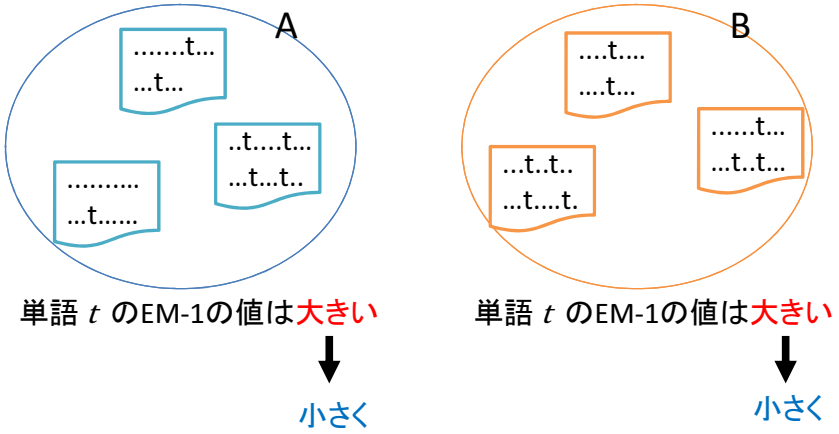

図 5.2: EM-1 値では両クラスに対する寄与度が大きいケース

のみ平均的に使用されている単語が観点を表す単語として重要であると仮定した手法である．

$$\underset{d\in D_A}{\text{mean}}\ f(D,d,t)/\underset{d\in \overline{D}_A}{\text{mean}}\ f(D,d,t) \tag{5.2}$$

クラス A を当該クラスとした時，図 5.2 ではクラス A に属する文書に出現する単語 t の特徴量をクラス A 内で平均化した値とクラス B とクラス C に属する文書に出現する単語 t の特徴量をクラス B とクラス C 内で平均化した値のクラス間で算出した比率を寄与度としている．そのため，EM-2 では，単語 t がクラス A 以外のクラスに属する文書に多く出現する時，寄与度は低くなる．

5.2.3 EM-3 クラス内の最大特徴量

EM-1 ではクラス内の少数の文書にのみ出現する単語の寄与度は小さい．このような EM-1 ではクラス内で平均化するため，サブクラスの特徴を捉えることができない．しかし，各メインクラスには図 5.3 のようにサブクラスが存在する．EM-1 で計算するとメインクラスのスポーツにおいて重要な単語は寄与度が大きくなると予想できるが，サブクラスであるサッカーや野球における重要な単語はメインクラス全体ではあまり多く出現するとは言えない．そのため，そのような単語の寄与度はクラス内で平均化され小さくなる．

サブクラスの特徴は間接的にメインクラスの特徴を表しており，サブクラスの特徴も観点を捉えるために重要な要素であると言える．そして，少なくとも一文書内で多く出現する単語はサブクラスの特徴を持つと考えた．

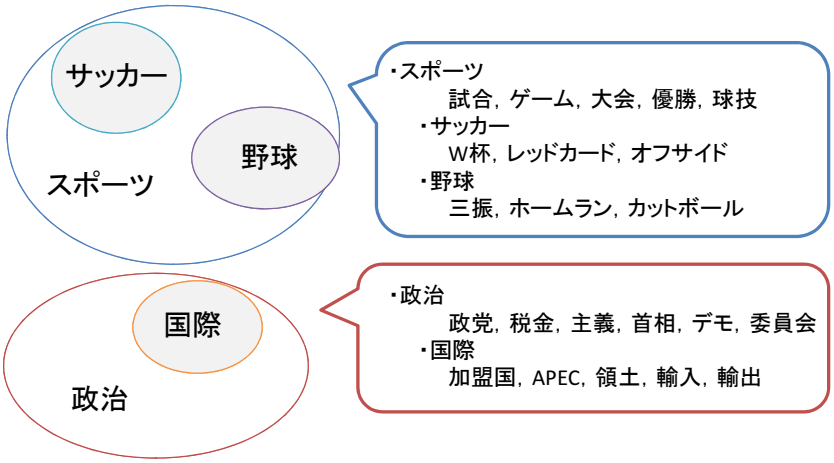

図 5.3: メインクラスとサブクラスにおける重要単語

実際には EM-3 は，D_A の全文書における単語 t に対する特徴量の最大値をクラス A に対する単語 t の寄与度とする．EM-3 では，一つの文書内で多く出現する単語がサブクラスの特徴を持ち観点を表す単語として重要であると仮定した手法である．EM-3 における寄与度は式 (5.3) で算出する．例えば，単語 t に対して，文書 $d \in D_A$ における特徴量が高く，それ以外の D_A に属する文書における特徴量が全て低い場合を考える．その場合，EM-1 の寄与度は平均化されるため低くなり，EM-3 の寄与度は最大値を取るため高くなる．

$$\max_{d \in D_A} f(D, d, t) \tag{5.3}$$

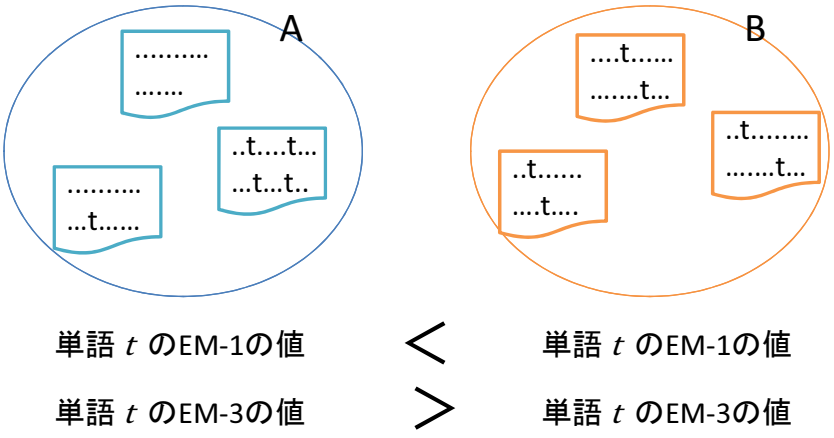

図 5.4: EM-1 と EM-3 の比較

図 5.4 を用いて説明すると，EM-1 値を考えると，クラス B ではあらゆる文書で単語 t が出現しているためクラス A よりクラス B の方が寄与度は大きい．しかし，

EM-3 値を考えると，クラス A では高々一つではあるが文書内に単語 t が多く出現しているためクラス B よりクラス A の方が寄与度は大きい．また，EM-3 では計算中の当該クラス以外のクラスについては考慮しない．

5.2.4 EM-4 最大特徴量のクラス間比率

EM-4 は，EM-2 の様に EM-3 で求めた寄与度に対してクラス間の比率を使用する．EM-2 における寄与度は式 (5.4) で算出する．EM-4 は，当該クラスにのみ出現し，一つの文書内で多く使用されている単語が観点を表す単語として重要であると仮定した手法である．

$$\max_{d \in D_A} f(D, d, t) / \max_{d \in \overline{D}_A} f(D, d, t) \tag{5.4}$$

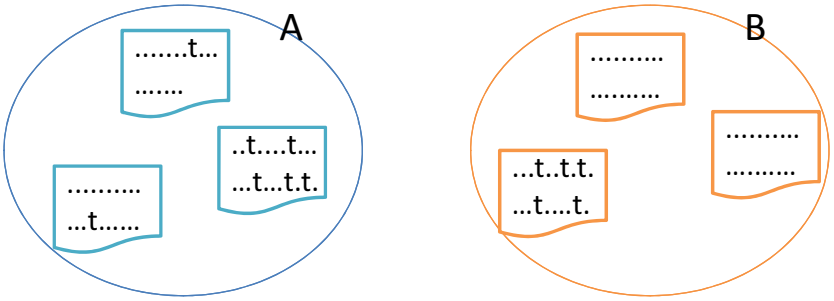

図 5.5: EM-4 当該クラスと非当該クラスでの最大特徴量の比

クラス A を当該クラスとした時，図 5.5 ではクラス A に属する文書の中で単語 t が最も多く出現する文書における単語 t の特徴量とクラス B とクラス C に属する文書の中で単語 t が最も多く出現する文書における単語 t の特徴量のクラス間で算出した比率を寄与度としている．そのため，EM-4 では，単語 t がクラス A 以外のクラスに属する文書に多く出現する時，寄与度は低くなる．

5.2.5　各観点抽出手法による寄与度ランキング

各観点抽出手法で観点の特徴をうまく捉えることができているか，実際に寄与度を算出し，寄与度が大きい単語について調査した．その結果が表5.2である．表5.2を見ると，4種類全ての観点抽出手法で算出した寄与度が大きい単語は各クラスにおいて特徴的な単語であることが確認できる．

表 5.2: ps:政治とスポーツの記事集合に対する寄与度ランキング

	EM-1		EM-2		EM-3		EM-4	
Rank	政治	スポーツ	政治	スポーツ	政治	スポーツ	政治	スポーツ
1	氏	メートル	氏	メートル	女性	メートル	女性	メートル
2	首相	位	首相	高	認定	高	認定	高
3	女性	—	女性	位	船	ダンス	船	ダンス
4	認定	W杯	認定	帝京大	竹島	—	竹島	田原
5	経済	高	経済	ダンス	寄付	ペア	寄付	ペア
6	都知事	五輪	寄付	W杯	行使	田原	行使	帝京大
7	寄付	帝京大	細川	早大	経済	帝京大	経済	本田
8	大統領	目	大統領	ＰＫ	甘利	本田	甘利	ＰＫ
9	細川	早大	アベベ	決勝	アベベ	部	アベベ	早大
10	アベベ	ダンス	都知事	目	容認	ＰＫ	容認	下村

このことから，4種類全ての観点抽出手法で算出する寄与度は十分観点の特徴を捉えることができていると考えられる．

第6章　提案手法：観点行列を用いた文書分類

第5章で説明した観点抽出手法で観点情報を抽出した後，その観点情報を文書分類に反映させる．この章では，観点抽出手法で算出した観点行列を文書分類に導入する方法について述べる．

本論文では，その際の文書分類に行列分解を利用する．そのため，観点を表す算出した各クラスと各単語に対する寄与度を行列空間表現として整理する．整理して出来た行列を観点行列U_mとして文書分類に反映させる．観点行列U_mは第5章の各抽出手法における各クラスごとの観点ベクトルを全てのクラス分整列することで作成する．

6.1　行列分解による分類

本論文では近年注目されているクラスタリングの一つであり第3章で説明したNMFという手法を利用している．その理由は大きくわけて三つある[13]．第一に，他のクラスタリング手法と比べ比較的計算が容易であり，第二に，類似の手法である特異値分解においては基底ベクトルを直交にしなければならない制約があるが，NMFではその制約がないため柔軟に基底ベクトルを定めることができるという利点があり，第三に，スパースな文書ベクトルに対する文書クラスタリングが良いクラスタリング結果になると報告されているからである．

既存のNMFへ観点行列U_mを導入する手法として文書と各クラスとの関連度を表す特徴行列V^Tを直接算出する手法と初期値に導入する手法の二つを提案する．特徴行列V^Tを直接算出する手法を直接的行列分解（NMF-DV：NMF with Direct matrix decomposition using Viewpoint matrix）と呼ぶことにする．初期値に導入する手法をNMF-IV（NMF with Initial basis value by Viewpoint matrix）と呼ぶことにする．

6.1.1 直接的行列分解 (NMF-DV)

直接的に既存の NMF へ観点行列 U_m を導入する NMF-DV では，NMF で使用していた式 (3.1) を利用する．式 (3.1) では索引語文書行列 X を基底行列 U と特徴行列 V^T の積の形に行列分解している．実際に NMF で行列分解を行う際には，基底行列 U と特徴行列 V^T は未知であり更新式を反復計算することで各行列に対する近似値の推定を行う．しかし，観点行列 U_m を利用することで更新式の反復計算せずに行列分解が可能となる．

具体的に NMF-DV では，式 (3.1) における基底行列 U に観点行列 U_m を代入し特徴行列 V^T を解く式 (6.1) の形に整理することで特徴行列 V^T を算出する．

$$V^T = U_m^+ X \tag{6.1}$$

ここで，式 (6.1) において A^+ は行列 A の擬似逆行列である．

NMF-DV は NMF で行う反復計算が不要であるため，良好な教師文書があれば短時間で文書分類を行うことができる．しかし，NMF-DV は教師文書から作成される観点行列 U_m に強く依存しており，教師文書と実際に分類する文書の差が大きい場合には上手く機能しないと考えられる．

6.1.2 初期値へ観点行列を導入する手法 (NMF-IV)

既存 NMF の初期値へ観点行列 U_m を導入する NMF-IV では，NMF-DV の様に直接的に特徴行列 V^T を算出するのではなく，NMF における基底行列の初期値に導入することで，反復計算による最適解の推定を行う．具体的には，NMF における基底行列 U の初期値へと観点行列 U_m を代入し既存の NMF と同様に更新式の反復計算を行い基底行列 U と特徴行列 V^T を推定する．それにより，通常の NMF より早くて良質な基底行列の解へと収束ができ，NMF-DV よりも最適な分類結果に収束することが期待される．

NMF-DV と NMF-IV を比較すると計算量は反復計算を行う必要がある NMF-IV の方が多く，教師文書に対する依存性は NMF-IV の方が小さいと考えられる．

第 4.3 節で説明した SSNMF は教師文書と類似する文書を教師文書のクラスに分類する手法であるのに対し，提案手法の NMF-DV や NMF − IV は教師文書から観点の特徴を取得し観点に沿った分類を行う．

6.2 N-gramモデルを低解像度とした多重解像度NMFによる分類

第6.1.1節で説明したNMF-DVや第6.1.2節で説明したNMF-IVは単語を単位とし単語の出現頻度を利用している．つまり，単語間の結びつき具合に関する情報は使用していない．しかし，単語の結びつき具合を表す共起情報は文書の特徴として重要な要素である．そのため，単語の共起情報を利用することで文書分類の性能が向上すると期待した．

そこで，形態素単位のN-gramを素性とし，各N-gramの出現頻度を特徴量とする文書行列Yを考えた．その時，N個の単語が一つのN-gramになっており，複数の要素の重ね合わせになっている．つまり，単語単位の文書行列XよりN-gram単位の文書行列Yは低解像度データであるとみなせる．

解像度が異なる2種のデータがあるため，節4.5で説明した多重解像度NMFを利用する．多重解像度NMFは画像処理において異なる解像度を利用することで各解像度で抽出できる2つの要素を同時に扱うことができ細胞検出などで有効性が示されている[14]．特に多重解像度NMF-model2を文書分類に適用させた手法を提案する．

6.2.1 多重解像度NMFをN-gramを使用した文書分類へ適用

文書分類における高解像度の文書ベクトルは単語を素性とした文書ベクトルとし，低解像度の文書ベクトルはbigramを素性とした文書ベクトルとする．

使用するbigramは形態素単位で構成する．また，付属語は全て同等な形態素として扱う．つまり，素性には“付属語+自立語”や“自立語+付属語”も含むこととした．

伊東等の論文[14]における多重解像度NMFの目的関数である式(4.9)は画像処理を目的としている．そのため，本論文では文書分類を目的とした多重解像度NMFの目的関数として式(6.2)を提案する．

$$J_m = \alpha||X - U_1V^T||_F + \beta||Y - MU_2V^T||_F + \gamma||U_1 - M^TMU_2||_F \qquad (6.2)$$

ここで，$Y \in \mathbb{R}^{b \times n}$ は bigram を素性にした低解像度文書行列であり，b は bigram の要素数である．$M \in \mathbb{R}^{b \times w}$ は各 bigram 要素とそれを構成する単語群の関係を表す行列であり，構成する単語に対応する要素は 0.5，それ以外の要素は 0 である．α，β と γ は各項に対する重みである．式 (6.2) の第三項目は各解像度における基底行列の差を小さくする．目的関数 J_m から求めた更新式は式 (6.3)，(6.4)，(6.5) である．この手法を MRNMF-N(MRNMF with N-gram model) と呼ぶ．この手法により高解像度文書行列からは単語の特徴を，低解像度文書行列からは熟語の特徴を同時に取得しより良い文書分類が行えることを期待する．

6.2.2 更新式

ラグランジュの未定乗数法を用いて式 (6.2) の J_m を最小にする基底行列 U_1，U_2，特徴行列 V の乗算型更新式を求める．

$$V_{ij} \leftarrow V_{ij} \frac{(\alpha X^T U_1 + \beta Y^T M U_2)_{ij}}{(\alpha V U_1^T U_1 + \beta V U_2^T M^T M U_2)_{ij}} \tag{6.3}$$

$$(U_1)_{ij} \leftarrow (U_1)_{ij} \frac{(\alpha X V + \gamma M^T M U_2)_{ij}}{(\alpha U V^T V + \gamma U_1)_{ij}} \tag{6.4}$$

$$(U_2)_{ij} \leftarrow (U_2)_{ij} \frac{(\beta M^T Y V + \gamma M^T M U_1)_{ij}}{(\beta M^T M U V^T V + \gamma M^T M M^T M U_2)_{ij}} \tag{6.5}$$

式 (6.4)，(6.5) の初期値に観点行列を使用することで多重解像度 NMF へ観点を導入する．

6.3 多義語に対する曖昧性の解消を伴った多重解像度NMFによる分類

これまで提案した手法，NMF-DV，NMF-IV と MRNMF-N は文や単語の意味を考慮せずに字面の情報のみを使用していた．しかし，人が文書を分類する際には文や単語の意味を考えて分類している．そして，単語には字面が同じで意味が異なるものがある．そのような単語を多義語というが，字面の情報のみの場合には，この多義語は意味が違っても同じ要素として扱ってしまう．そのため，多義語の曖昧性を解消すること (WSD) で分類性能の向上が期待される．

第 6.3.1 節で WSD については詳しく後述するが，自然言語処理では多義語の曖昧性を解消することは重要な問題になる．それは，文書分類でも重要な問題である．語義の曖昧性を解消しない場合，複数の文書に出現する単語がそれぞれの文書で異なった意味で使用されていたとしても一つの素性として扱ってしまう．実際に多義語が文書で使用されている割合を調査した結果が図 6.1 である．図 6.1 の多義語数は，文書中の単語が第 6.3.1 項で説明する辞書の WordNet において多義である数をカウントしている．

表 6.1: 多義語が含まれている割合

データ	文書数	索引単語数 t	多義語数 p	多義語割合 p/t
ps	99	3149	1926	61.2%
se	149	3317	2162	65.2%
et	500	7048	4372	62.0%
it	500	7374	4632	62.8%

調査結果では文書集合 4 種の全てにおいて多義語の割合が 60%以上であった．つまり，多義語を考慮しない場合はこの 60%の単語の意味の違いを無視していることになる．

このように多義語が多い文書集合に対する文書分類では，多義語はクラスタリングにおける特徴として十分に機能しない恐れがある．これは，各クラスにおいて語義の重要度が異なるが，単語の意味による判別が出来ないからである．更に，文書分類における WSD は一単語のみを対象にするのではなく，全ての多義語を対象にする all-words WSD であるため処理は難しくなる．そのため，本論文では一単語を対象にした NMF による WSD を考え，それを全ての単語に拡張し多義語に対する曖昧性の解消を伴った文書分類手法を提案する．

6.3.1 多義語の曖昧性解消 (WSD)

自然言語処理における重要な問題の一つに，多義語の意味的曖昧性の解消 (WSD: Word Sense Disambiguation)[16] がある．多義語とは複数の意味を持つ語である．例として多義語として「勉強」が出現する次の 2 つの文がある．

- 例文 1：試験に向けて勉強する

- 例文 2：店員が勉強してくれたおかげで安く買えた

例文 1 における「勉強」は学問や芸術などを学ぶことという意味で使用される．例文 2 における「勉強」は商人が商品を値引きして安く売ることという意味で使用される．このように，複数の意味を持つ多義語の曖昧性を解消する技術は，機械翻訳や自動要約，情報検索システムなどにおいて性能を向上するために必要な技術である．文書が電子データとして簡単に保存できるようになり，大量のコーパスが利用可能になったため統計的学習に基づいた語の曖昧性を解消する研究が多く行われてきた．

WSD の手法には語義が付与されていないコーパスを用いる手法と人手により語義が設定されているコーパスを用いる手法がある．

Lesk アルゴリズム

WSD の手法は大きく教師有り学習と教師無し学習の二つの方法がある．教師有り学習では，語義タグ付きコーパスなど人手で準備された教師データを利用して，第 2.5 節で述べたような機械学習による分類手法で語義の推定を行う．

教師なし WSD の一つに Lesk アルゴリズム [35] がある．この手法は対象語に対する辞書に記載されている語義の説明文や例文の中で，最も重複する単語が多い語義を選択する．その他の教師無し WSD として周辺語の語義との関連性を利用した手法 [29] や確率モデルを利用した手法 [30] がある．辞書には下で説明する WordNet などを利用する．

WordNet

WordNet [31] とは，英語の大語意データベースである．名詞，動詞，形容詞，副詞が “Synset” という異なる概念を表す単位によってグループ化されている．また，各 synset は上位下位関係などの多様な関係で結ばれている．

そして本論文で使用している日本語 WordNet[1]は大規模な日本語の意味辞書であり，独立行政法人情報通信研究機構 (NICT) が作成したものである．これはプリンストン大学で開発された Princeton WordNet [2]やヨーロッパの

[1]http://nlpwww.nict.go.jp/wn-ja/
[2]http://wordnet.princeton.edu

EuroWordnet 協会が推進する Global WordNet Grid [3] に着想を得て開発されている．

6.3.2 一単語対象の NMF による WSD

語義候補数 s_h 個の多義語 h である時の一単語を対象にした NMF による WSD を考える．その時，分類対象の文書 d においてその語 h が出現する周辺前後 5 単語を素性とした各文書の文書ベクトルを作成する．その文書ベクトル群である文書行列 $Y_h \in \mathbb{R}^{a_h \times d}$ を入力行列とする．また，辞書に記載されている単語 h の定義文と例文に出現する単語と各文書における単語 h の周辺前後 5 単語の重複単語を素性として基底行列 $R_h \in \mathbb{R}^{a_h \times s_h}$ を作成する．そして，各文書における語義候補の選択を行う特徴行列 $W_h \in \mathbb{R}^{s_h \times d}$ を作成する．ここで，d は文書数，a_h は単語 h における周辺単語数である．

式 (6.6) のように文書行列 Y_h を基底行列 R_h と特徴行列 W_h の積の形に分解する．

$$Y_h = R_h W_h^T \tag{6.6}$$

各文書における語義候補の選択を行うための特徴行列 W を推定するため，各多義語 h の曖昧性解消を行う NMF の目的関数 J_h として式 (6.7) を定義する．また，この NMF による WSD は教師なし WSD である．

$$J_h = ||Y_h - R_h W_h^T||_F \tag{6.7}$$

WSD を行うには通常の NMF と同様に特徴行列 W_h の推定を行う．そして式 (6.8) のように語義候補の中で関連度が大きい語義と判別する．

$$\text{文書 } i \text{ の多義語 } h \text{ の語義} = \arg\max_{s_h} w_{is_h} \tag{6.8}$$

次に，この NMF による WSD の精度を計る実験を行った．実験では Agirre らの手法 [36] と比較する．Agirre らは WordNet が持つ「Synset」という概念の単位を利用して生成したグラフに対してページランクによる語義を表すノードの重要度を計算し重要度が高い語義を選択している．実験で使用するデータは SemEval-2007[34]

[3] http://globalwordnet.org

で公開されている語義付きコーパス「English Lexical Sample」である．このデータには100単語の多義語が準備され，その多義語が出現する数十から数百の文書が付属している．この文書中の多義語はWordNetに基づいてアノテーションされている．実験における評価値は適合率と再現率のみでの調和平均$F1$とした．実験結果を図6.2に示した．

表 6.2: 各手法での一単語WSD

	NMFによるWSD	Agirreらの手法
F1	0.401	0.417

結果は提案手法のNMFによるWSDの方がF1は少し低い．しかし，最近の教師なしWSD手法と評価値は同等であるためNMFによるWSDは可能であると考えられる．

6.3.3 多義語に対する曖昧性の解消を伴った分類手法の概要

一単語を対象にしたNMFによるWSDを全ての単語を対象に拡張し，同時に文書分類を行う手法を考える．全ての多義語に対する曖昧性解消を同時に行うために，全ての多義語に対する特徴行列W_hを結合させた行列$Z \in \mathbb{R}^{s \times d}$を式(6.9)のように定義する．ここで，$s = \sum_h^H s_h$である．

$$Z = \begin{pmatrix} W_1 \\ \vdots \\ W_h \\ \vdots \\ W_H \end{pmatrix} = E_1 W_1^T + E_2 W_2^T + \cdots + E_h W_h^T + \cdots + E_H W_H^T \tag{6.9}$$

ここで，行列$E_h \in \mathbb{R}^{s \times s_h}$は多義語$h$に対応した単位行列$I_h \in \mathbb{R}^{s_h \times s_h}$と$h$以外の多義語$\bar{h}$に対応した零行列$O_{\bar{h}} \in \mathbb{R}^{s_{\bar{h}} \times s_{\bar{h}}}$を結合した行列である．

$$E_1 = \begin{pmatrix} I_1 \\ \vdots \\ O_H \end{pmatrix} \quad , \quad E_h = \begin{pmatrix} O_1 \\ \vdots \\ I_h \\ \vdots \\ O_H \end{pmatrix} \quad , \quad E_H = \begin{pmatrix} O_1 \\ \vdots \\ I_H \end{pmatrix} \tag{6.10}$$

その時，s_h 個の語義候補が一つの単語になっており，複数の要素の重ね合わせになっている．つまり，単語単位の文書行列 X より全多義語の語義候補が素性の文書行列 Z は低解像度データであるとみなせる．つまり，第 6.2 節と同じように単語を素性とした文書行列を高解像度とする．そして，各文書と各語義候補との関連度を表す特徴行列 W_h を低解像度とする．解像度が異なる 2 種のデータがあるため，節 4.5 で説明した多重解像度 NMF を利用する．

第一項を高解像度文書行列による文書分類，第二項を低解像度文書行列による文書分類として，第三項目に各多義語の曖昧性回避を行う目的関数 J_p として式 (6.11) を定義する．この多義語に対する曖昧性の解消を伴った多重解像度 NMF を MRNMF-WSD と呼ぶ．

$$J_p = ||X - U_1 V^T||_F + \alpha||Z - MU_2 V^T||_F + \sum_h^H ||J_h||_F \tag{6.11}$$

6.3.4 更新式

ラグランジュの未定乗数法を用いて式 (6.11) の J_p を最小にする基底行列 U，R_h，特徴行列 V，W_h の乗算型更新式を求める．

$$V_{ij} \leftarrow V_{ij} \frac{(X^T U + \alpha Z^T MU)_{ij}}{(VU^T U + \alpha VU^T M^T MU)_{ij}} \tag{6.12}$$

$$U_{ij} \leftarrow U_{ij} \frac{(XV + \alpha M^T ZV)_{ij}}{(UV^T V + \alpha M^T MUV^T V)_{ij}} \tag{6.13}$$

$$(R_h)_{ij} \leftarrow (R_h)_{ij} \frac{(Y_h W_h)_{ij}}{(R_h W_h^T W_h)_{ij}} \tag{6.14}$$

$$(W_h)_{ij} \leftarrow (W_h)_{ij} \frac{(Y_h^T R_h + \alpha VU^T M^T E_h)_{ij}}{(W_h R_h^T R_h + \alpha Z^T E_h)_{ij}} \tag{6.15}$$

式 (6.13) の初期値に観点行列を使用することで多義語に対する曖昧性の解消を伴った多重解像度 NMF による文書分類へ観点を導入する．

第7章　実験

この章では第5章の観点抽出手法で作成した観点行列を導入した第6章の分類手法に対する分類性能についての評価実験を行い，その考察について述べる．

7.1　実験概要

提案手法の分類性能についての評価実験として実験1〜3を行う．実験1では，第6.1節で提案したNMF-DVとNMF-IVを用いて分類実験を行う．実験2では，第6.2節で提案したMRNMF-Nを用いて分類実験を行う．実験3では，第6.3節で提案したMRNMF-WSDを用いて分類実験を行う．

各実験では第5.2節で提案したEM-1からEM-4の各手法で算出した観点行列を用いて観点抽出手法の比較も行う．

また，提案手法と次のつの手法と比較を行う．まず1つ目は第3.1節で述べた教師無しNMFである既存NMF，2つ目は第4.3項で述べた教師有りNMFのSSNMFを用いる．3つ目から5つ目の手法はNMF以外の分類手法を用いる．3つ目は第2.5.1項で述べたNaive Bayes，4つ目は第2.5.2項で述べたSVM，5つ目は第2.5.3項で述べたMLRを用いる．SVMはGaussカーネル，多項式カーネル，シグモイドカーネル，線形カーネルによる事前実験を行った．その結果，線形カーネルが良好であったため比較するSVMには線形カーネルを用いる．実験では，20種類の異なる初期値を準備する．それらの初期値に対して各手法で文書分類を行い平均評価値を調査した．SSNMFの結果に関しては論文[15]を参考に自作したプログラムによる結果である．

7.2 実験用文書データセット

7.2.1 実験 1 & 2 用

本論文ではシングルラベル文書とマルチラベル文書の文書データセットを用いて各手法の比較を行っている．シングルラベルの文書データは正解となるクラスが 1 つである文書データである．マルチラベル文書データは正解となるクラスが複数である文書データである．実験ではシングルラベルのみで構成されたデータとマルチラベルを合わせた混合データに対して分類を行う．シングルラベルのみで構成されたデータは CLUTO のサイト [32] で公開されている文書集合と，Web 朝日 [33] で公開されているラベルが一つである記事を収集し作成した文書集合を使用した．シングルラベルとマルチラベルのデータを合わせた混合データは Web 朝日 [33] から取得したシングルラベルの記事データに両方のラベルが付与されているマルチラベルの記事を合わせて作成した．各文書データセットの詳細は表 7.1，表 7.2 と表 7.3 に示す．

CLUTO のデータセットである k1a,k1b と wap は Yahoo!内の様々な web ページから構成され，re0 はロイターのニュースワイヤーから取得したニュース記事で構成されている．また，tr31，tr41 は TREC と呼ばれるワークショップが情報検索関連の研究分野のために作成したテスト用文書である．そして，fbis は米国の元政府機関である Foregin Broadcast Information Service (FBIS) が収集したニュース記事で構成されている．

表 7.1: CLUTO：シングルラベルデータ

Data	Docs	Terms	Class
k1a	2340	21839	20
k1b	2340	21839	6
re0	1504	2886	13
wap	1560	6460	20
tr31	972	10128	7
tr41	878	7454	10
fbis	2463	2000	17

Web 朝日のデータセットはマルチラベルの影響を比較し易くするために 2 クラスとしている．また，混合データの構成はそれぞれのシングルラベル文書と両ラベ

表 7.2: Web 朝日：実験 1 & 2 用シングルラベルデータ

Data	Docs	Sing Lelabel	Terms	Class
ps(政治，スポーツ)	80	40 + 40	2806	2
se(スポーツ，経済)	120	60 + 60	2776	2
et(経済，技術)	400	200 + 200	6277	2
it(事故，技術)	400	200 + 200	6855	2

ルが付与されたマルチラベル文書を混在させている．

表 7.3: Web 朝日：実験 1 & 2 用混合データ

Data	Docs	Sing Lelabel	Multi Label	Terms	Class	多義語数
ps(政治，スポーツ)	99	40 + 40	19	3149	2	1926
se(スポーツ，経済)	149	60 + 60	29	3317	2	2162
et(経済，技術)	500	200 + 200	100	7048	2	4372
it(事故，技術)	500	200 + 200	100	7374	2	4632

7.2.2 実験 3 用

実験 3 で調査する提案手法では，各多義語の語義推定を文書分類と同時に処理する．そのため，多義語の語数によって処理時間が大幅に長くなる．よって，実験 1 & 2 用の実験データよりも多義語が少ない実験データを利用する．

実験 3 の実験データは，実験 1 & 2 で使用した Web 朝日のデータの文書数を減少させたデータを利用する．文書データセットの詳細は表 7.4 に示す．

表 7.4: Web 朝日：実験 3 用文書データセット

Data	Docs	Sing Lelabel Docs	Multi Label Docs	Terms	Class	多義語数
ps	30	10 + 10	10	1491	2	892

7.3 評価方法

分類結果の評価値には Entropy，Purity，RandIndex，Precision，Recall 及び F 値 (Precision と Recall の調和平均）[20] を用いる．さらに最終的な分類性能値はこれらの F 値を除く五種類の評価値の調和平均 Hm により評価する．

Entropy は (7.1) 式より求める．Entropy は各クラスタにおける正解集合の分布度合を表しており，小さな値ほどクラスタリング結果が良好であることを意味している．ここで N は総文書数を示す．また，調和平均 Hm を算出する際には $(1-Entropy)$ として計算する．

$$Entropy = \sum_{i=1}^{k} \frac{|C_i|}{N} \times (-\sum_{h=1}^{k} P(A_h|C_i) \log P(A_h|C_i)) \tag{7.1}$$

Purity は結果クラスタに一番多く含まれている正解クラスタを用いて，結果クラスタに正解データが含まれている割合を示す指標である．クラスタリング結果の Purity は，各クラスタのデータ数による重み付き平均をとるように定義し，(7.2) 式に示す．

$$Purity = \frac{1}{N} \sum_{i=1}^{k} \max_{h} |C_i \cap A_h| \tag{7.2}$$

(7.1)，(7.2) 式において C_i はクラスタリング結果に対する i 番目の文書のクラスタであり，A_h は正解データに対する h 番目の文書のクラスタである．$A_h \cap C_i$ は正解データであるクラスタ A_h とクラスタリング結果のクラスタ C_i が共通している文書数である．

RandIndex はデータの各ペア同士の正解が同じクラスタならば同じクラスタになるかどうかの判定の正解率を表し (7.3) 式で求める．

$$RandIndex = \frac{TP + TN}{TP + FP + FN + TN} \tag{7.3}$$

ここで TP は同じ正解クラスタであるデータのペアが結果クラスタで同じクラスタである対の数，TN は異なる正解クラスタであるデータのペアが結果クラスタで異なるクラスタである対の数，FP は異なる正解クラスタであるデータのペアが結果クラスタで同じクラスタである対の数，FN は同じ正解クラスタであるデータのペアが結果クラスタで異なるクラスタである対の数を表す．

Precision はクラスタリング結果の中にどの程度正解が含まれているを表す．実際には以下の (7.4) 式で求める．

$$Precision = \frac{TP}{TP + FP} \tag{7.4}$$

Recall は正解データがどの程度結果クラスタで正しくクラスタリングされているかを表す．実際には以下の (7.5) 式で求める．

$$Recall = \frac{TP}{TP + FN} \tag{7.5}$$

7.4 事前実験

実験では混合データのマルチラベル文書の正解クラスを片方のクラスに偏らせることで擬似的な人の観点による分類を設定する．そして両方のパターンによる分類実験の平均を評価する．

しかし，実際の人の観点による分類との差異が大きい場合，正確な評価ができない．そのため，実際にマルチラベル文書を人が分類した正解データを使用した事前実験を行う．事前実験では擬似的な観点の正解と実際に人が分類した正解との比較を行う．人手による分類作業は負担が大きいため比較的サイズが小さい文書データ集合である表 7.3 の “ps” と “se” を使用した．人手による分類は学生と教員を含めた 6 人で行った．

7.4.1 実験結果

文書データ “ps” の分類結果を図 7.1，“se” の分類結果を図 7.2 に示す．

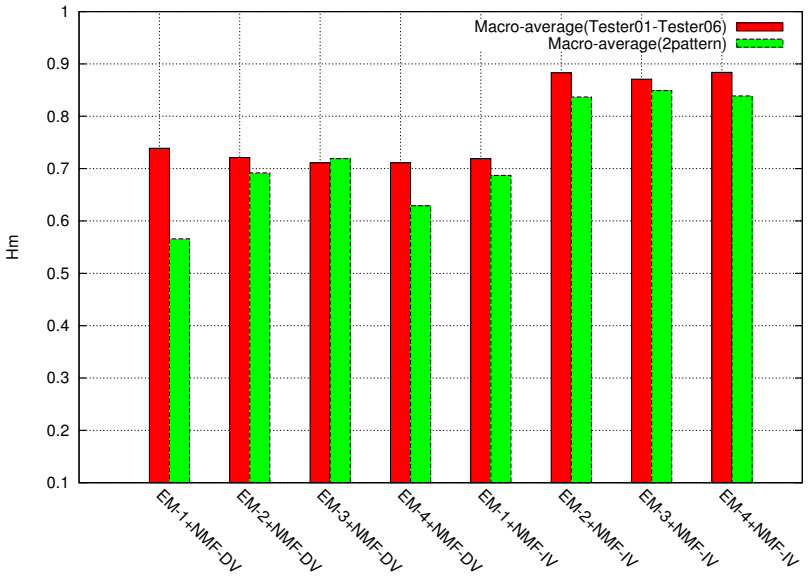

図 7.1: ps - 擬似的な観点の正解と実際の人の観点による正解との比較

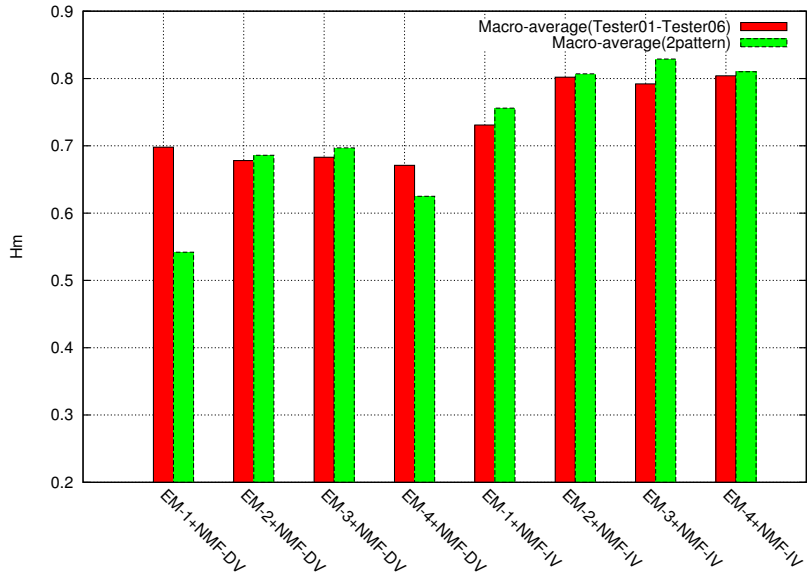

図 7.2: se - 擬似的な観点の正解と実際の人の観点による正解との比較

結果から擬似的な観点の正解と実際の人の観点による正解との比較する．図 7.1，図 7.2 において EM-1 と NMF-DV を組み合わせた分類手法では両データにおいて結果に大きな差ができた．EM-1+NMF-DV 以外の手法では大きな差は見られない．そのため，EM-1+NMF-DV 以外の手法については人手による分類が困難である大きなサイズの文書データに対して擬似的な観点の正解を使用しても分類手法の比較は可能であると思われる．EM-1+NMF-DV については，最も偏った観点である擬似的な観点の正解を使用した分類性能が低いため特異な観点を持つユーザに対して弱いという性質があると考えられる．

7.5 実験 1：観点行列を用いた行列分解による分類

7.5.1 実験結果

CLUTO のシングルラベルの教師データ数は各クラス 5 個使用している．Web 朝日のシングルラベルは各クラス 10 個使用している．混合データは各クラス 10 個使用している．CLUTO のシングルラベル文書の分類実験結果を図 7.3 に示す．Web 朝日のシングルラベル文書の結果を図 7.4 に示す．混合文書の実験結果は図 7.5 に示す．

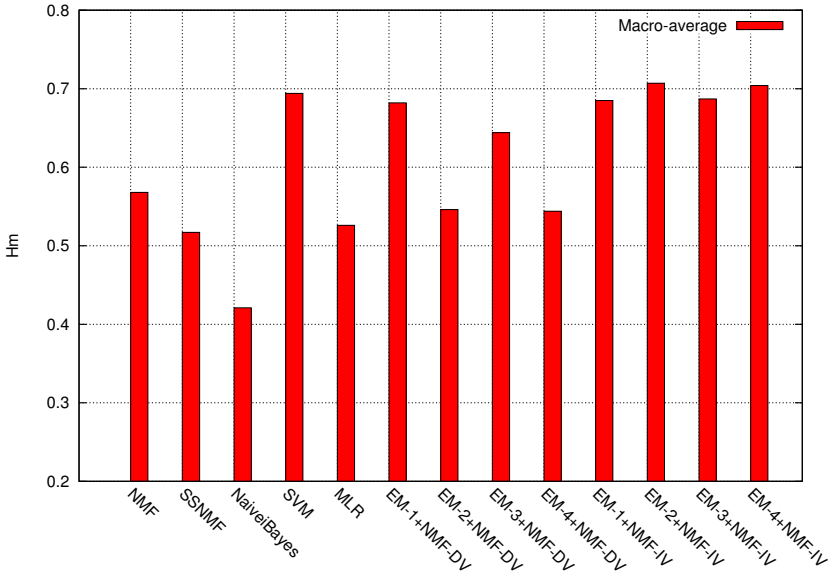

図 7.3: CLUTO のシングルラベルデータに対するマクロ平均

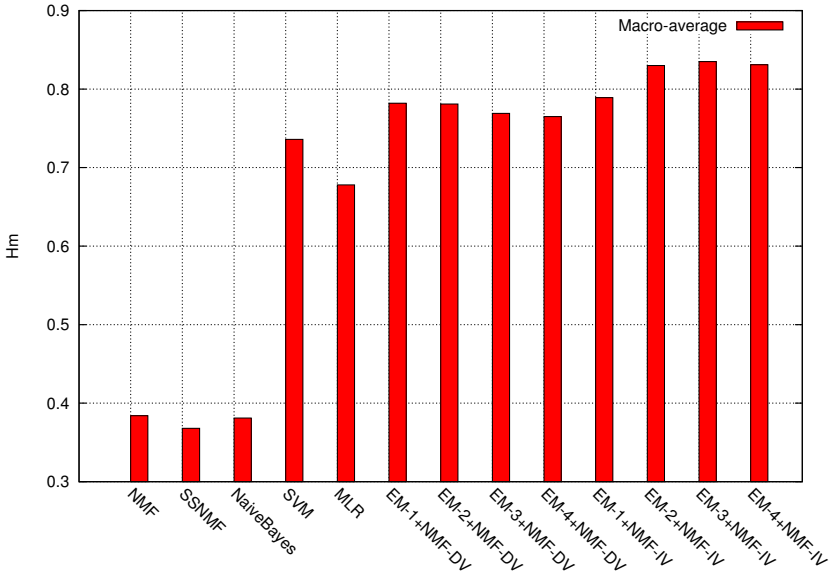

図 7.4: Web 朝日のシングルラベルデータに対するマクロ平均

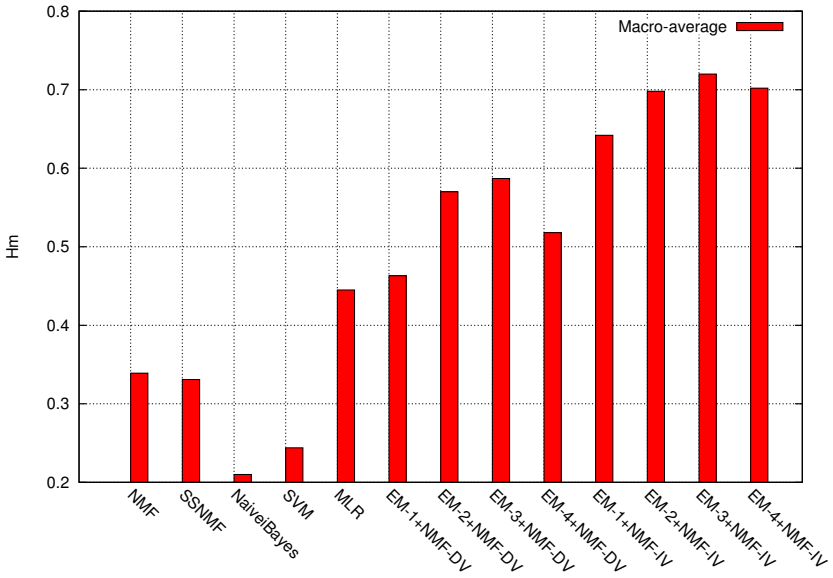

図 7.5: Web 朝日の混合データに対するマクロ平均

7.5.2 考察

7.5.2.1 分類手法について

まず，CLUTO のシングルラベル文書に対する分類結果について考察する．図 7.3 では，SVM，EM-1+NMF-DV と EM-1〜4+NMF-IV が同等の評価値である．NMF-DV と当該クラスと他クラスの比を寄与度としない EM-1 と EM-3 の組み合わせが良かった理由としては，NMF-DV は教師データから作成した観点行列への依存性が高く，シンプルに EM-1 や EM-3 のように当該クラスの特徴のみで十分分類できたためと考えられる．

表 7.5: 実行時間の比較 [sec]

	NMF-DV	NMF-IV
ps	0.391	16.303
se	0.353	17.189
et	0.675	24.758
it	0.944	25.948

また，図 7.6 からシングルラベル文書のみに対しては教師データ数が少なくなると SVM よりも NMF-IV の方が有利であることがわかっている．

次に，Web 朝日のデータに対する分類結果について考察する．シングルラベル

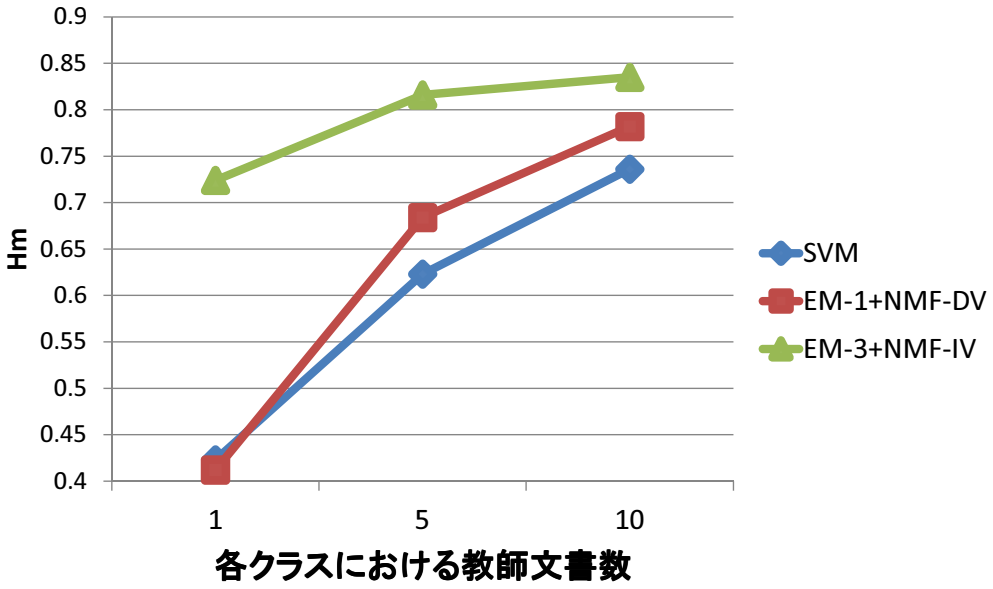

図 7.6: Web 朝日のシングルラベルに対する教師データ数の推移

文書のみの場合，図 7.4 から比較した手法の中では EM-3+NMF-IV が最も高い評価値であった．マルチラベル文書を混合した文書集合の結果について考察すると，表 7.5 でわかる通りに NMF-DV の方が NMF-IV よりも実行時間は短いが，分類性能を比較した図 7.5 では，EM-3+NMF-IV が最も高い評価値となっている．特に，NMF-DV 系よりも NMF-IV 系の手法の方が全体的に評価値が高い．その理由としては以下のように考えられる．マルチラベル文書は両方のクラスにおける特徴的な単語が出現するため良質な教師データとしての観点行列を作成することが困難であり，NMF による基底行列の最適化を行わす教師データに対する依存性が高い NMF-DV での分類は難しいため混合データに対する NMF-DV の評価値が低くなったと考えられる．シングルラベルのみと混合データの結果を比較すると，両方において，EM-3+NMF-IV が他の手法よりも高い評価値である．混合データは分類が困難なマルチラベル文書を混ぜているためどの手法でも分類性能は下がっている．しかし，NMF-IV は約 13%しか低下していない．他の手法の NMF-DV と MLR は約 23%低下してる．特に，SVM はシングルラベルのみの時は高い分類性能であったが，約 49%も低下している．

7.5.2.2 観点行列について

擬似的な観点に対して各観点抽出手法で算出した観点行列の比較と，NMF による基底行列に対する最適化処理前の観点行列と後の基底行列を比較する．観点行列と基底行列は単語とクラスとの関連度を表していると考えられる．そのため，各クラスにおいて関連度が高い上位 10 位の単語を求め観点の推定が機能しているかを調査する．代表して文書データセットの “ps” を図 7.6，図 7.7 に示す．残りの文書データセットである “se”，“et” と “it” の結果は付録に載せる．

表 7.6: ps:政治寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	政治	スポーツ	政治	スポーツ	政治	スポーツ	政治	スポーツ
1	スポーツ	ＰＫ	スポーツ	ＰＫ	浅田	田原	浅田	田原
2	浅田	W杯	浅田	田原	球団	ＰＫ	球団	ＰＫ
3	首相	田原	首相	帝京大	山下	帝京大	トロフィー	帝京大
4	球団	行進曲	球団	行進曲	トロフィー	行進曲	%	行進曲
5	%	ペア	%	ペア	%	ペア	山下	ペア
6	氏	帝京大	氏	綱	狩野	早大	狩野	綱
7	トロフィー	綱	トロフィー	早大	横断幕	綱	横断幕	早大
8	設置	早大	羽生	カ月	羽生	本田	羽生	みき
9	羽生	中日	設置	本田	文化	ロナルド	文化	本田
10	経済	—	経済	中日	女性	みき	女性	ロナルド

表 7.7: ps:スポーツ寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	政治	スポーツ	政治	スポーツ	政治	スポーツ	政治	スポーツ
1	認定	スポーツ	認定	スポーツ	%	浅田	認定	浅田
2	経済	浅田	アベベ	浅田	認定	山下	層	球団
3	%	五輪	都知事	球団	経済	球団	アベベ	山下
4	アベベ	球団	細川	山下	アベベ	トロフィー	単位	トロフィー
5	首相	%	%	トロフィー	層	%	%	狩野
6	氏	山下	単位	五輪	単位	狩野	支持	横断幕
7	大統領	トロフィー	大統領	狩野	支持	横断幕	決議	羽生
8	市長	選手	層	羽生	東国	羽生	東国	文化
9	単位	羽生	中国	設置	戦争	文化	島国	%
10	細川	狩野	東国	庁	甘利	スポーツ	戦争	スポーツ

表 7.8: ps:政治寄り観点の上位単語 (NMF-IV 後)

	EM-1		EM-2		EM-3		EM-4	
Rank	政治	スポーツ	政治	スポーツ	政治	スポーツ	政治	スポーツ
1	氏	位	氏	位	氏	位	氏	位
2	首相	目	首相	目	首相	目	首相	目
3	都知事	W杯	都知事	W杯	都知事	W杯	都知事	W杯
4	選	メートル	選	メートル	選	メートル	選	メートル
5	経済	決勝	経済	決勝	経済	決勝	経済	決勝
6	安倍	—	安倍	—	安倍	—	安倍	—
7	細川	戦	スポーツ	戦	スポーツ	戦	スポーツ	戦
8	スポーツ	4	細川	4	細川	4	細川	4
9	大統領	勢	大統領	勢	大統領	勢	大統領	勢
10	東京	ソチ	東京	回	%	回	東京	回

表 7.9: ps:スポーツ寄り観点の上位単語 (NMF-IV 後)

	EM-1		EM-2		EM-3		EM-4	
Rank	政治	スポーツ	政治	スポーツ	政治	スポーツ	政治	スポーツ
1	氏	位	氏	五輪	氏	位	氏	五輪
2	首相	W杯	都知事	位	都知事	五輪	都知事	位
3	都知事	五輪	首相	W杯	首相	W杯	首相	W杯
4	選	目	選	目	選	目	選	目
5	細川	メートル	細川	選手	細川	メートル	細川	選手
6	大統領	選手	移設	メートル	大統領	選手	移設	メートル
7	経済	決勝	大統領	スポーツ	移設	決勝	大統領	決勝
8	移設	スポーツ	経済	決勝	経済	スポーツ	経済	スポーツ
9	%	4	原発	4	原発	4	原発	4
10	原発	首相	%	首相	市長	戦	%	首相

まず，各観点抽出手法の比較を行う．表 7.6 は，政治とスポーツのマルチラベル文書を政治寄りの観点による分類を行った際に算出した観点行列における上位単語である．そのため，EM-1〜4 で作成した観点行列の政治クラスでは上位にもスポーツに関する単語が見られる．表 7.7 は，政治とスポーツのマルチラベル文書をスポーツ寄りの観点による分類を行った際に算出した観点行列における上位単語である．表 7.6，表 7.7 の両方において EM-1 と EM-2，または EM-3 と EM-4 のランキングは似ている．この事から他のクラスとの比率を使用しても上位単語にはあまり影響が少ないと考えられる．しかし，図 7.3 においてシングルラベル文書の NMF-DV の結果ではクラス間比率を考慮しない方が，NMF-IV ではクラス間比率を考慮した方が良い結果であった．一方，図 7.5 では，クラス間比率を考慮しない EM-3 が NMF-DV，NMF-IV 共に良い結果であった．この事からシングルラベルのようにクラス間の差が大きい文書集合ではクラス間比率を考慮した EM-2 や EM-4 が有効であると考えられる．そして，混合データに対しては，現在は 2 クラスでの実験であるため，3 クラス以上の際や特に似たクラスの場合における調査が必要である．また，EM-3 と EM-4 では，一文書内で多く出現する単語の寄与度が高くなるため，EM-1 や EM-2 よりも個人名が上位になる傾向がみられる．

次に NMF-IV による観点行列の最適化後の比較を行う．表 7.8，表 7.9 では，EM-2，EM-3 と EM-4 のランキングがとても似ている．これは，NMF による基底行列の最適化により近い値に収束したためと考えられる．そのため，図 7.5 において EM-2〜4+NMF-IV の結果が同等になったと考えられる．また，表 7.8，表 7.9 のランキングも似ている．つまり，同様の値に収束していることから観点の違いによる NMF への影響が少ないと考えられる．そのため，初期値のみならず局所解の制約としても観点行列を導入するなど観点の影響を更に強めるような手法が必要である

と考えられる．

7.6 実験 2：N-gram モデルを低解像度とした多重解像度 NMF による分類

7.6.1 実験結果

N-gram モデルを低解像度とした多重解像度 NMF である MRNMF-N の結果を図 7.7 に示す．図 7.7 では実験 1 の中で最も良い結果を示した EM-3+NMF-IV との比較も行っている．

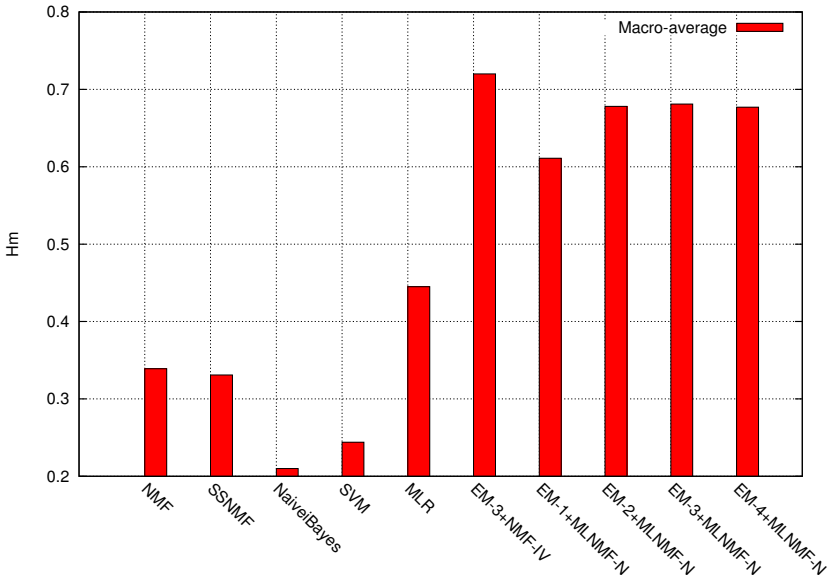

図 7.7: Result of MRNMF-N

7.6.2 考察

7.6.2.1 分類手法について

MRNMF-N 系の手法の中では，EM-2〜4+MRNMF-N が良い評価値である．しかし，EM-3+NMF-IV と比較すると，EM-3+NMF-IV の方が約 4% 評価値が高いという結果が分かる．

更に，EM-3+NMF-IV と EM-3+MRNMF-N の結果について，第 7.3 節で説明した 5 種の評価値を解析した所，他の評価値は下がっていたのに対して Recall の評価値は上がっていた．また，分類後に出来た各クラスのクラスタサイズにおいて，

NMF-IV よりも少し片方のクラスに偏っていた．つまり，MRNMF-N は NMF-IV よりも一つのクラスタに偏った分類をしてしまうという性質が確認された．しかし，使用する教師データによっては 5 つ全ての評価値が約 2% 程度上がったケースが見られた．これにより，単語よりも熟語が重要となる文書も存在すると考えられる．MRNMF-N の結果はあまり良いとは言えないが，熟語を考慮する必要がある文書も存在するため熟語に対応した MRNMF-N は重要であるが基底行列の制約や α，β や γ 等の重みパラメータの組合せなどの改良の必要がある．

7.6.2.2 観点行列について

第 7.5.2.2 項で調査したように，MRNMF-N を実行した後の基底行列 U_1 における各クラスにおいて関連度が高い上位 10 位の単語を求める．結果を表 7.10 と表 7.11 にまとめる．

表 7.10: ps:政治寄り観点の上位単語 (最適化後)-U_1

	EM-1		EM-2		EM-3		EM-4	
Rank	政治	スポーツ	政治	スポーツ	政治	スポーツ	政治	スポーツ
1	首相	位	首相	位	首相	位	首相	位
2	氏	目	氏	目	氏	目	氏	目
3	都知事	W杯	都知事	W杯	都知事	W杯	都知事	W杯
4	経済	メートル	選	メートル	選	メートル	選	メートル
5	選	—	経済	—	経済	—	経済	—
6	安倍	決勝	安倍	決勝	安倍	決勝	安倍	決勝
7	スポーツ	4	スポーツ	4	スポーツ	4	スポーツ	4
8	大統領	戦	大統領	戦	大統領	戦	大統領	戦
9	細川	5	細川	回	細川	回	細川	回
10	東京	回	東京	ソチ	東京	5	東京	ソチ

表 7.11: ps:スポーツ寄り観点の上位単語 (最適化後)-U_1

	EM-1		EM-2		EM-3		EM-4	
Rank	政治	スポーツ	政治	スポーツ	政治	スポーツ	政治	スポーツ
1	氏	位	氏	五輪	氏	位	氏	五輪
2	首相	W杯	都知事	位	都知事	W杯	都知事	位
3	都知事	五輪	首相	W杯	首相	五輪	首相	W杯
4	選	目	選	目	選	目	選	目
5	細川	メートル	細川	選手	細川	メートル	細川	メートル
6	大統領	—	大統領	メートル	大統領	—	大統領	選手
7	経済	決勝	移設	—	移設	選手	移設	—
8	移設	選手	経済	スポーツ	経済	決勝	経済	決勝
9	%	4	%	決勝	%	スポーツ	女性	スポーツ
10	市長	スポーツ	原発	4	市長	4	%	4

NMF-IV の時と MRNMF-N の時の最適化後における基底行列の重要上位単語を比較する．そのため，表 7.8 は表 7.10 と，表 7.9 は表 7.11 と比較する．比較結果として，全体的にランキングは似ているが，異なる所として，政治クラスでは “首相” や “大統領” などのランキングが上がり，スポーツクラスでは “首相” や “スポーツ” のランキングが下がっている．“首相” や “大統領” に関しては単語としてよりも” 安部首相”，“森首相” や “韓国大統領” といった熟語として多く出現するため低解像度文書行列である bigram の効果で政治クラスでは重要度が上がり，スポーツクラスでは下がったと考えられる．一方で，“スポーツ” は熟語よりも単語による出現が多いため，MRNMF-N では重要度が下がったと考えられる．

7.7 実験 3：多義語の曖昧性の解消を伴った多重解像度 NMF による分類

7.7.1 事前実験

MRNMF-WSD では all-words WSD と文書分類を同時に処理している．MRNMF-WSD の all-words WSD 部分は第 6.3 節で解説したように式 (6.11) の目的関数 J_p における第 3 項目である．そして，ある一単語 h のみの WSD は式 (6.7) を目的関数とした NMF により語彙判断を行うことができる．MRNMF-WSD は一単語を対象とした WSD の精度に依存する．そのため，事前実験として NMF による一単語を対象にした WSD の性能調査を行う．

実験には SemEval-2007[34] で公開されている語義付きコーパスである “English Lexical Sample” を用いる．このコーパスは WordNet の辞書を利用している．コーパスには 100 単語の多義語が準備され，各多義語が出現する数十から数百の文書が準備されている．各文書に出現する多語義は WordNet に基づいてアノテーションされている．NMF による語義判定の結果を表 7.12 に示す．

表 7.12 から語義判定を NMF で行うと高い精度で判定できる多義語もあるが判定精度が低い多義語もあることが分かる．判定性能である F 値が 0.7 以上である多義語は 4 単語あった．また，0.3 以下である多義語は 13 単語であった．ここで，0.7 以上の 4 単語は “attempt”，“avoid”，“disclose” と “remove” である．

表 7.12: NMF による語義判定の結果

	平均値	中央値	最大値	最小値
Entropy	0.273	1.696e-06	0.969	2.465e-33
Purity	0.391	6.283e-07	0.992	6.163e-34
RandIndex	0.487	0.493	0.775	3.140e-05
Precision	0.589	0.571	0.988	5.547e-33
Recall	0.326	0.276	0.939	6.163e-34
F-measure	0.448	0.453	0.786	0.123

7.7.2 実験結果

多義語の曖昧性の解消を伴った多重解像度 NMF である MRNMF-WSD の結果を図 7.8 に示す．実験 3 では，使用する文書データが実験 1 と異なる．そのため，NMF-IV との比較を可能にするために，実験 3 で使用した文書データによる NMF-IV の結果も図 7.8 に載せている．

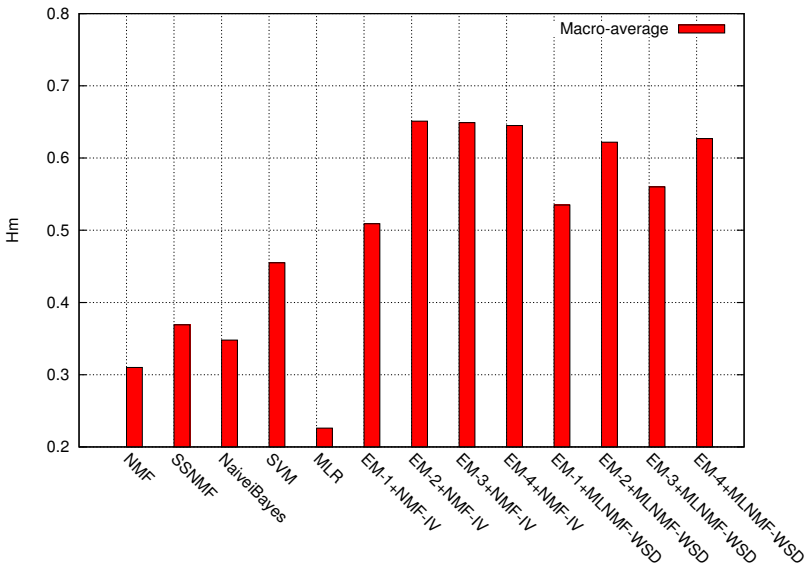

図 7.8: Result of MRNMF-WSD

7.7.3 考察

図 7.8 において MRNMF-WSD 系の中では，EM-2+MRNMF-WSD と EM-4+MRNMF-WSD が良い評価値である．しかし，EM-2〜4+NMF-IV と比較すると，EM-2〜4+NMF-IV の方が約 3% 評価値が高いという結果が分かる．

また，図 7.5 における NMF-IV 系の手法や図 7.7 における MRNMF-N 系の手法では EM-2 や EM-4 よりも EM-3 の方が有効であったのに対して，図 7.8 では EM-2 や EM-4 に MRNMF-WSD を組み合わせた手法の方が EM-3+MRNMF-WSD よりも良好である．これは，EM-2 と EM-4 は両クラスで多く出現する単語の重みを小さくするため，WSD の精度の影響が小さかったのではないかと考えられる．そのため，WSD の結果が大きく反映される EM-1 や EM-3 の評価値は減少したということは WSD の精度が低いと考えられる．また，対象とするデータが異なるが第 7.7.1 節で行った事前実験では WSD の精度が比較的に良好な多義語も存在したが，全体的な精度は高いとは言えない結果であった．

よって，WSD 部分のみの精度に対する改善が必要である．

第8章　おわりに

この章では本研究のまとめと今後の課題について述べる．

8.1　まとめ

本論文では，ユーザの観点を抽出し観点行列として表現し行列分解を用いた文書分類手法に導入することで，ユーザの観点に沿った文書分類を目指し，そのためのいくつかの手法を示した．

まず，明示的に表現することが困難な観点を算出する手法として EM-1，EM-2，EM-3 と EM-4 の 4 手法を提案した．作成される観点行列において寄与度が高い上位 10 単語ランキングを確認すると，十分各クラスタの特徴を捉えている単語が上位にきた．

次に，観点を反映した分類手法として，比較的シンプルに観点行列を用いる NMF-DV，NMF-IV を提案した．Web 朝日の記事データを使用した結果の場合，シングルラベルデータのみの時とマルチラベルデータを混ぜた混合データの時の両方で EM-3 + NMF-IV が他の手法よりもよい実験結果であった．他の手法の中では SVM が最も良い結果であった．特に，シングルラベルデータのみの際には SVM も高い分類性能を示した．しかし，教師データを少なくした場合や分類が困難なマルチラベルデータを混ぜた場合には SVM よりも EM-3 + NMF-IV の方が有効であった．例えば，教師データ数を各クラス 1 個にした場合は SVM より EM-3+NMF-IV の方が評価値は約 30%も高かった．更に，マルチラベルを混ぜた場合には SVM よりも EM-3 + NMF-IV の方が評価値は約 48%も高かった．EM-3+NMF-IV が有効であるということは，ユーザの観点に沿った分類を行うには教師データから算出する特徴としてクラス内における文書間での共通の特徴よりも各文書における独自の特徴を用いることが重要であると考えられる．

以上のことから，マルチラベル文書も含まれる一般の文書に対してユーザの観点

に沿った文書分類を行うには EM-3+NMF-IV が有効である．

また，単語のみではなく単語の共起具合を考慮した MRNMF-N と，多義語に対する曖昧性の解消を伴う手法として MRNMF-WSD も提案した．

NMF-IV と MRNMF-N を比較すると，実際の分類された各クラスの構成と 5 種の評価値の中で Entropy が最も大きく悪化していることから，MRNMF-N は多くのテストデータで NMF-IV の際よりも一方のクラスに少し偏りやすい傾向がみられた．実験での結果は期待したほどの性能向上は見られなかった．しかし，使用する教師データによっては MRNMF-N の方がよい分類結果である場合も確認できた．これは，例えば，“韓国” や “大統領” と “韓国大統領” のように単語よりも単語列の方が各文書間での共起数が少なくなるため教師文書と分類する文書間で同じ単語列が共起している場合は bigram を素性とした低解像度文書行列がプラスに働くと考えられるが，単語の共起特徴を十分に捉えるには教師データが不足していたと考えている．

MRNMF-WSD の結果をまとめる．まず，一単語を対象にした NMF による WSD は最近の従来手法と同等の性能であった．しかし，文書分類に用いるには WSD 部分の精度が十分ではなかったため，実験では期待したほど WSD の効果による分類の改善は見られなかった．マルチラベル文書はクラスごとに多義語の語義が異なる場合があると考えられるため高精度の MRNMF-WSD は特にマルチラベル文書の分類には必要であると期待される．しかし，all-words WSD を行うため計算コストが他の手法よりも掛かるという問題も他の all-words WSD タスクと同様に存在している．

提案手法は記事の分類を対象にしたが，色々な応用が考えられる．例えば，企業内の部署によって観点が異なる場合は，その部署の観点に沿った書類の分類に使用できる．また，実用的なシステムへの適用を考えると，最初はユーザが通常の検索結果を使用して閲覧した記事にラベルを付けてもらうことで教師文書を獲得する．その教師文書から観点の特徴を算出し EM-3+NMF-IV で観点に沿った文書分類ができると考えられる．ユーザがシステムを使用していくにつれ教師文書を増やす機能を載せることで良質な教師データの獲得を行うことが望ましい．方法としては，ユーザが閲覧した文書のラベル付けを初回以外も行う方法や，システムが定期的に分類結果が観点に沿っているか尋ねて，正しいという回答が得られたら，その時の

分類結果を教師文書に追加する方法などが考えられる．教師データが少ない初期頃はEM-3+NMF-IVによる分類が有効であるが，教師データの数が増えたら，計算コストが小さいEM-3+NMF-DVによる分類が実行時間が短くなるため有効であると考えている．

8.2 今後の課題

本論文では，観点行列の導入方法としてNMF-DVとNMF-IVの2通りの手法を提案したが，NMFにおける目的関数には反映されていない．そのため，NMF-IV後には観点の影響が小さくなってしまっていると考えられる．よって，観点行列をNMFの局所解への制約として導入する必要がある．

また，単語の共起特徴を十分に反映させるために，MRNMF-Nの改善が必要である．特に，基底行列に対する制約部分に対して改善の余地があると考えられる．また，教師文書の量を変更するなどの更なる実験が必要である．そして，MRNMF-WSDはWSD部分だけでの精度向上が必要である．さらに，MRNMF-WSDはall-words WSDを行っているため，他の手法よりも計算コストが大きい．そのため，計算コストを減らすような改善が必要である．

謝辞

本研究を行うにあたり，御多忙にもかかわらず御指導頂いた大学院情報工学研究院知能情報工学研究系の中村貞吾准教授、永井秀利助教に心からお礼申し上げます．

また，本研究をまとめるにあたり，適切な助言を戴くとともに本論文の細部にわたり御指導を頂いた大学院情報工学研究院知能情報工学研究系の竹内章教授，乃万司教授，また，大学院情報工学研究院情報創成工学研究系の梅田政信教授に感謝致します．

そして，研究室の方々にはいろいろと相談にのっていただき，ありがとうございました．

参考文献

[1] Google，http://www.google.co.jp

[2] Yahoo!，http://www.yahoo.co.jp

[3] Koji Eguchi，Keizo Oyama，Emi Ishida，Noriko Kando，Kazuko Kuriyama，“Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure”，IEICE，Transactions on Information and Systems，Vol.E86-D，No.9，pp1804-1813，2003.

[4] Sergey Brin，Lawrence Page，R.Motwani，and Teryy Winograd，“The PageRank citation ranking: Bringing order to the Web”，Technical Report 1999-0120，Computer Science Department，Stanford University，1999.

[5] Amy N.Langville，Carl D.Meyer，岩野和生訳，黒川利明訳，黒川洋訳，“Google PageRank の数理 最強検索エンジンのランキング手法を求めて”，共立出版，2009.

[6] Jon Kleinberg，“Authortative sources in a hyperlinked environment”，Journal of the ACM，46，1999.

[7] 安形輝，石田栄美，久野高志，野末道子，上田修一，“WWW ページの自動分類 : NDC 分類体系と Yahoo のカテゴリを使った分類”，情処研報 FI-54，1999.

[8] 松村亮介，福田直樹，横山昌平，石川博，“SearchLife : 単語の特徴量を考慮した多視点クラスタリング検索エンジン”，情報処理学会論文誌，データベース，Vol.3，No.2，pp123-137，2010.

[9] 成田宏和，太田学，片山薫，石川博，“Web 文書検索のための非排他的クラスタリング手法の提案”，DBWeb2003，2003.

[10] 城市広大，三好力，“ベクトル空間法とファジィ推論を用いたWEB検索結果自動分類システム”，日本知能情報ファジィ学会誌，知能と情報，Vol.18，No.2，pp184-195，2006.

[11] 金子大輔，高山毅，池田哲夫，長内亘，“Web文書のページタイプを用いた適応的分類と試作システムの評価”，日本知能情報ファジィ学会誌，知能と情報，Vol.18，No.2，pp319-336，2006.

[12] D.D.Lee , H.S.Seung “Algorithms for Non-negative Matrix Factorization”, NIPS , pp.556-562 , (2000).

[13] W.Xu，X.Liu，and Y.Gong，“Document clustering based on non-negative matrix factorization”，in Proc.ACM SIGIR Conf.Research and Development in Information Retrieval (SIGIR)，Toronto，ON，Candada，2003.

[14] 伊東翼，太田圭輔，村山正宣，青西享，“多重解像度非負値行列因子分解によるカルシウムイメージングデータ解析”，IEICE Technical Report，Vol.115，No.514，pp131-136，2016.

[15] H.Lee , J.Yoo , S.Choi “Semi-Supervised Nonnegative Matrix Factorization”, IEEE SIGNAL PROCESSING LETTERS , Vol.17 No.1 , pp.4-7 , JANUARY 2010.

[16] Roberto Navigli，Universita di Roma La Sapienza，Rome，Itary，“Word Sense Disambiguation: A Survey”，ACM Computing Surveys，Vol.41，No.10，pp10:1-10:69，2009.

[17] Ronen Feldman，James Sanger，辻井潤一監訳，IBM東京基礎研究所訳，“テキストマイニング ハンドブック”，東京電機大学出版局，2010.

[18] Philip J.Hayes，Steven P.Weinstein，“Consture-TIS: A System for Content-Based Indexing fo a Database of News Stories”，IAAI，pp49-64，1990.

[19] 田中穂積監修，“自然言語処理ー基礎と応用ー”，社団法人 電子情報通信学会，株式会社コロナ社，1999.

[20] C.D.Manning，P.Raghavan，H.Schutze，岩野和生訳，黒川利明訳，濱田誠司訳，村上明子訳，“情報検索の基礎”，共立出版株式会社，2012.

[21] 新納浩幸，佐々木稔，“NMF とリンクベースの修正法によるピンポン型文書クラスタリング”，情報処理学会，自然言語処理研究会報告，Vol.2007,no.47,p.7-12.

[22] 新納浩幸，佐々木稔，“Mcut + NMF による文書クラスタリング”，言語処理学会年次大会発表論文集，Vol.13，pp,558-561,(2007).

[23] 堀田政二，宮原末治，“Non-neagtive Matrix Factorization の初期値の設定法とその応用”，電子情報通信学会技術研究報告，Vol.102，No.652，pp.19-24，2003.

[24] 丸田要，永井秀利，中村貞吾，“文書分類のための教師制約を用いた非負値行列因子分解”，IAS2013，情報アクセスシンポジウム 2013，pp.14-21，(2013).

[25] Carrot2，http://search.carrot2.org/stable/search

[26] もりきよし原編，“日本十進分類法 新訂 10 版”，日本図書館協会，2014.

[27] C.Ding,T.Li,W.Peng : “On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing”, Computational Statistics and Data Analysis 52 , 3913 - 3927 (2008).

[28] D.Zhang，Z.Zhou，S.Chen，“Nonnegative Matrix Factorization on Kernels”，PRICAI 2006: Trends in Artificial Intelligence Lecture Notes in Computer Science Vol.4099，pp.404-412，2006.

[29] Ted Pedersen，Satanjeev Banerjee，Siddharth Patwardhan，“Maximizing Semantic Relatedness to Perform Word Sense Disambiguation”，Research Report UMSI，pp1-34，2002.

[30] Jordan Boyd-Graber，David M.Blei，Xiaojin Zhu，“A Topic Model for Word Sense Disambiguation”，Procccdings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computaional Natural Language Learning，pp.1024-1033，2007.

[31] WordNet，http://wordnet.princeton.edu/

[32] CLUTO，http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download/

[33] Web 朝日，http://asahi.com

[34] SemEval-2007，http://nlp.cs.swarthmore.edu/semeval/

[35] Michael Lesk，“Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream conc.”，In Proceedings of the 5th Annual International Conference on Systems Documentation，SIGDOC，pp.24-26，1986.

[36] Eneko Agirre，Oier Lopez de Lacalle，Aitor Soroa，“Random Walks for Knowledge-Based Word Sense Disambiguation”，Computational Linguistics，Vol.40，No.1，pp.57-84，2014.

[37] 萩野広樹, 吉田哲也, “トピックグラフに基づく NMF を用いた転移学習”, IPSJ SIG Technical Report, Vol.2011-MPS-82 No17.

[38] Y.Chen，M.Rege，M.Dong，J.Hua，“Non-negative matrix factorization for semi-supervised data clustering”，Knowl，Inform. Syst.，vol.17，pp.355-379，2008.

[39] F.Wang，T.Li，C.Zhang，“Semi-supervised clustering via matrix factorization”，in Proc. SIAM Int. Conf. Data Mining (SDM)，Atlanta，GA，2008.

[40] C.Ding，T.Li，W.Peng，H.Park，“Orthogonal nonnegative matrix tri-factorizations for clustering. ”，In Proceedings of ACM SIGKDD，pp.126-135，2006.

[41] L.Baker，A.McCallum，“Distributional clustering of words for text classification. ”，In Proceedings of ACM SIGIR，1998.

[42] Y.Xue，C.S.Tong，W.S.Chen，W.Zhang，Z.He，“A modified non-negative matrix factorization algorithm for face recognition”，in Proc. Int. Conf. Pattern Recognintion (ICPR)，Hong Kong，pp.495-498，2006.

[43] 亀岡弘和，ルルージョナトン，“Frobenius ノルム規準の非負値行列因子分解における乗法更新式に関する一考察”，日本音響学会講演論文集，pp.709-712，2009.

[44] 菅原拓夢，笹野遼平，高村大也，奥村学，“単語の分散表現を用いた語義曖昧性解消”，言語処理学会，第 21 回年次大会発表論文集，pp648-651，2015.

[45] 藤井洋一，鈴木克志，辻秀一，“段落内共起情報を利用した文書自動分類方式”，情報処理学会誌，Vol.42，No.3，pp495-506，2001.

[46] A.Agresti，“An Introduction to Categorical Data Analysis”，John Wiley & Sons，2006.

[47] 佐々木悠人，古宮嘉那子，森田一，小谷善行，“周辺語義モデルによる日本語の教師無し語義曖昧性解消”，研究報告自然言語処理，Vol.218，No.3，pp1-14，2014.

付録

A　各観点抽出手法での重要語ランキング一覧

表 9.1: se:スポーツ寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	スポーツ	経済	スポーツ	経済	スポーツ	経済	スポーツ	経済
1	選手	日経	選手	日経	レース	日経	レース	日経
2	岡田	円	岡田	銭	西武	銭	西武	銭
3	錦織	銭	錦織	取引	ザック	ウルグアイ	ザック	ウルグアイ
4	さん	取引	さん	ウルグアイ	ジャガー	炉	ジャガー	炉
5	スポンサー	炉	スポンサー	炉	部員	ウー	部員	腰
6	日本	ウルグアイ	W杯	株価	遼	生命	遼	生命
7	W杯	株価	試合	値上がり	キャプテン	腰	キャプテン	ウー
8	試合	コイン	レース	生命	セット	コイン	セット	工場
9	五輪	値上がり	五輪	腰	岡田	工場	岡田	バー
10	レース	生命	監督	米	本塁打	株価	本塁打	株価

表 9.2: se:経済寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	スポーツ	経済	スポーツ	経済	スポーツ	経済	スポーツ	経済
1	鈴木	岡田	鈴木	岡田	内村	レース	内村	レース
2	台風	選手	内村	円	鈴木	ザック	鈴木	ザック
3	内村	円	台風	錦織	オリックス	ジャガー	オリックス	ジャガー
4	オリックス	錦織	オリックス	さん	台風	遼	台風	遼
5	受賞	さん	ヤクルト	スポンサー	受賞	キャプテン	受賞	キャプテン
6	体操	スポンサー	受賞	レース	偉業	セット	偉業	セット
7	ヤクルト	レース	体操	選手	久保	尼崎	久保	尼崎
8	パラリンピック	W杯	パラリンピック	W杯	真中	ハンドボール	パラリンピック	ハンドボール
9	久保	契約	久保	契約	パラリンピック	岡田	真中	日経
10	小川	五輪	小川	ユニ	安打	日経	安打	セール

表 9.3: et:経済寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	経済	技術	経済	技術	経済	技術	経済	技術
1	銭	北朝鮮	銭	北朝鮮	ＡＶ	北朝鮮	ＡＶ	北朝鮮
2	日経	ロボット	日経	出願	コイン	ロボット	コイン	館
3	取引	館	取引	館	Ｗｉｉ	Ｗｉ	Ｗｉｉ	出願
4	買収	出願	買収	踏切	通貨	館	日経	踏切
5	平均	Ｗｉ	平均	Ｗｉ	日経	踏切	通貨	Ｗｉ
6	円	踏切	コイン	モン	拠点	出願	拠点	郵便
7	コイン	モン	株価	サイバー	イー	郵便	イー	モン
8	通貨	受験	通貨	受験	支店	モン	ＮＦＣ	受験
9	株価	サイバー	ビット	スト	ＮＦＣ	受験	支店	スト
10	ビット	郵便	Ｗｉｉ	郵便	音	スト	音	ロボット

表 9.4: et:テクノロジー寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	経済	技術	経済	技術	経済	技術	経済	技術
1	・	コイン	銭	コイン	日経	ＡＶ	銭	ＡＶ
2	銭	買収	日経	買収	銭	Ｗｉｉ	日経	Ｗｉｉ
3	ドル	携帯	前日	携帯	増税	通貨	増税	コイン
4	増税	ＡＶ	増税	ＡＶ	ドル	コイン	支出	イー
5	円	通貨	ＴＰＰ	Ｗｉｉ	支出	イー	原油	通貨
6	ＴＰＰ	Ｗｉｉ	平均	端末	原油	ルモンド	クッキー	ルモンド
7	日経	ゲーム	首相	ビット	クッキー	拠点	報酬	拠点
8	協議	ビット	報酬	通貨	報酬	ＮＦＣ	民営	ＮＦＣ
9	安	端末	クッキー	ゲーム	復興	動物	ロシア	動物
10	首相	ＮＦＣ	支出	ＮＦＣ	ロシア	音	復興	買収

表 9.5: it:事故寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	事故	技術	事故	技術	事故	技術	事故	技術
1	不正	コイン	片山	コイン	会員	コイン	片山	コイン
2	会員	工場	不正	工場	片山	工場	会員	工場
3	流出	米	流出	ビット	ＰＷ	ポイント	ＰＷ	衛星
4	東芝	ポイント	会員	衛星	島野	衛星	島野	小保
5	片山	衛星	東芝	小保	条項	鶴岡	児童	負担
6	情報	小保	ＬＩＮＥ	鶴岡	児童	負担	条項	鶴岡
7	ＬＩＮＥ	鶴岡	島野	観測	東芝	小保	東芝	録画
8	件	観測	条項	負担	Ｓｕｉｃａ	観測	Ｓｕｉｃａ	ＯＳ
9	被告	・	情報	%	爆竹	録画	爆竹	%
10	島野	負担	被告	区間	慶応大	・	慶応大	観測

表 9.6: it:テクノロジー寄り観点の上位単語 (最適化前)

	EM-1		EM-2		EM-3		EM-4	
Rank	事故	技術	事故	技術	事故	技術	事故	技術
1	アンネ	不正	アンネ	片山	アンネ	会員	アンネ	片山
2	さん	流出	猪瀬	流出	猪瀬	片山	猪瀬	会員
3	判決	片山	船	不正	被爆	コイン	船	コイン
4	猪瀬	東芝	認定	コイン	船	島野	被爆	島野
5	船	会員	さん	サイト	医師	ＰＷ	脱税	ＰＷ
6	被爆	コイン	被爆	会員	弟	条項	弟	条項
7	少年	サイト	弟	東芝	脱税	ミク	談合	ミク
8	医師	情報	脱税	ＬＩＮＥ	少年	慶応大	医師	ポイント
9	認定	ＬＩＮＥ	同校	アクセス	認定	東芝	同校	慶応大
10	弟	島野	医師	島野	疾患	爆竹	認定	爆竹