修士論文

記述式小テストの解答を分類するための解答の特徴抽出に関する研究

平成18年度修了

三重大学大学院工学研究科

博士前期課程　電気電子工学専攻

大井　健太郎

目次

第1章　序論

近年，大学への進学率が上昇し続けている．これにより，さまざまな知識や興味を有する学生が大学に入学している[1]．また，学習意欲の低下が指摘されており，それにともなった学力低下が問題視されている[2]．その一方で，学習意欲の低下は講師の教え方にあるとする意見がある．大学の講師の関心は研究面に向けられ，学生の教育に対する責任が十分に意識されてこなかった[1]．学生の学力低下を抑えるため，講師の側にも，学生にとってわかりやすい講義を行うことが要求される時代になってきている．

講師がわかりやすい講義を行うためには，講義の内容をどれだけ学生が理解しているのかを把握することが必要である．学生の理解度を知るための手段にはさまざまなものがある．その一つに小テストがある．学生は講師の説明から理解したことをもとに小テストの解答を行う．そのため，小テストの結果から，講師は学生の理解が不十分な箇所を知ることができ，それに応じた追加説明などのフィードバックを行うことができる．フィードバックが早ければ早いほど学生の理解は深まる．したがって，フィードバックは，小テスト終了後すみやかに行うことが望ましい．しかし，小テストは解答の回収，集計，分析に手間がかかるため，すみやかなフィードバックは難しい．

近年，eラーニングが注目されている．計算機を使用することで，講義にさまざまな効果がもたらされる．eラーニングにおける小テストでは，解答の回収，集計，分析を素早く行うことができる．本研究では，解答の分析を行うことで，学生へ適切なフィードバックをすみやかに行えるようにすることを目標とする．

小テストにはさまざまな形式がある．そのうち，記述式小テストでは，解答中にさまざまな表現が現れるため，全体の傾向を即時につかむことは困難である．そこで，記述式小テストの解答を分類することにより，全体の傾向をつかみやすくすることを考える．分類には一般の文書分類に用いられる方法[3][4]が利用できる．しかし，記述式小テストの解答は，一般の文書分類が対象としている文書と，質，量ともに大きく異なる．そのため，一般の文書分類の手法をそのまま用いるのではなく，記述式小テストの分類に適した手法を用いる必要がある．

文書分類の一般的な手順は，まず，分類対象の文書を数値化し，ベクトル（特徴ベクトルと呼ぶ）として表現する．次に，特徴ベクトルの類似度に基づき分類を行う．ベクトル化がうまくできないと，いかなる分類手法を用いても適切な分類を行うことができない．そこで，本研究では前半の特徴ベクトルを作成する部分を対象とする．

記述式小テストを適切に分類するための特徴ベクトル作成手法として，以下の2点を提案する．

・数値化対象の品詞の拡大

小テストの解答にはそれほど多くの単語が含まれていない．その中から名詞のみを取り出して分類に用いるのでは，情報が不足する．そこで，名詞だけではなく動詞も分類に用いることにする．これは，文章の主な構成要素は主語（名詞）と述語（動詞）であるためである．ただし，動詞の活用形を考慮に入れると複雑になりすぎるので，語幹のみを扱うものとする．

・係り受けに基づく重み付け

これは，単語の出現回数以外に，解答中での文節の係り受けの関係に着目する方法である．具体的には，より多くの文節から修飾されている文節ほど，その解答で中心的な働きをしていると考え，それに含まれる単語を重要視する．その結果，ささいな修飾語の違いに惑わされず，その文で表したかった内容の特徴ベクトルを作成することができる．

本論文の構成を以下に示す．2章で分類対象である小テストについて述べる．3章では，一般的な文書分類に用いられる手法について説明する．4章では，記述式小テストの分類に際し，従来法では適していない点について検討し，それを解決する手法を提案する．5章では，実際に行われた記述式小テストの解答を対象に提案手法を適用し分類を行う．その結果から提案手法の有効性を確認する．最後に6章で本研究をまとめる．

第2章　小テスト

講師は，学生の理解度を把握する手段の一つとして，試験や演習を行う．学生の理解が不十分である箇所を知ることができれば，それに応じた補足説明などのフィードバックを行うことができる．

試験や演習にもさまざまな形式がある．本研究では，小テストについて考える．この章では，小テストの意義と，小テストの形式，また，eラーニングにおける小テストについて説明する．

2.1　学生の理解度を把握する手段

わかりやすい講義を行うためには，講師は学生の理解度を把握する必要がある．学生の理解度を把握するための手段にはさまざまなものがある．例えば，講師が「わかりましたか」「質問はありますか」など，学生に質問を促すことでも学生の理解度を把握できる．しかしながら，講義中に質問をすることは，学生にとって気恥ずかしさや緊張がともなう．したがって，講義中に講師が質問を促しても，自発的に質問をする学生は限られてしまう．そのため，一部の学生の理解度を把握することはできるが，学生全体の理解度を把握することはできない．

また，別の手段として，小テストがある．小テストでは，学生は実際に講師の説明を聞いて理解したことをもとに解答する．質問を促す場合と異なり，気恥ずかしさなどがないため，全学生がそれぞれの理解度に応じた解答をする．そのため，講師にとって，説明した内容をどの程度学生が理解できたかを把握する方法として有効である．そこで，本研究では小テストを対象とする．

しかしながら，小テストは効果的である反面，全ての解答に目を通すには時間がかかるという欠点がある．講義における学生の数は数十人に及ぶことも多く，講義中にフィードバックを行うことは困難になる．理想的には学生一人一人の理解度を把握して，それぞれに対応することが望ましい．しかし，講義中に学生一人一人に対応することは不可能である．講義中は学生全体に対してフィードバックを行い，個別の学生に対するフィードバックが必要な場合には講義外に応対するのが良い．本研究では，講義中に行うフィードバックのために，講師が学生全体の理解度を把握できるように補助することを考える．

2.2　小テストの形式

小テストの形式にはさまざまなものがある．以下に主な形式を挙げる．

・選択式問題

あらかじめ答の候補が挙げられており，候補の中から適切なものを選ばせる．

（問）E メールにおける subject とはどういう意味か？適切なものを選べ．

A. 問題　B. 科目　C. 主題　D. 主語

（答）C

・穴埋め式問題

文章などに空欄が用意されており，空欄に当てはまる解を入れることで解答させる．

（問）次の文章の空欄を埋めよ．

URL とは，(　　) 上に存在する情報資源（文書や画像など）の場所を指し示す記述方式である．

（答）インターネット

・記述式問題

問題に対する答を記述形式で解答させる．

（問）インターネットを利用するときに，情報倫理に関して気をつけなくてはいけないことは何か？説明せよ．

（答）インターネットを使えば容易に様々なデータを入手することができるが，どのようなデータに関しても利用する際には著作権が存在していることを忘れてはならない．

学生の理解度を把握する上で，最も有効な形式は記述式の小テストである．記述式小テストは，勘やうろ覚えで正しい解答をすることが難しく，学生が講義の内容を正しく理解しているかを適切に把握することができる．また，間違った解答であっても，学生がどのように考えて間違ったのかを把握できる．そのため，講師は追加説明などのフィードバックを行いやすい．

しかし，記述式小テストは他の問題形式と比べて全ての問題に目を通すには時間がかかる．講師がフィードバックを行いやすくするため，記述式小テストの解答を把握しやすくするすることを本研究の目的とする．

2.3　e ラーニングにおける小テスト

一般の教室で行う講義において，小テストは，配布，回収の手間がかかる．それが学生の理解度を把握することが遅れる原因になる．

近年，e ラーニングを利用した講義が増えている．e ラーニングでの講義ならば，小テストを配布，回収する手間は省くことができる．また，小テストの集計を計算機で行うことができるため，講師が解答の傾向を把握する助けになる．本研究では，e ラーニングにおける小テストを扱う．計算機を用いて解答を集計し，解答の傾向を把握しやすくなる形にして講師に提供する．

第3章　文書の自動分類

文書の自動分類は，広い意味でデータマイニングの分野に属している．データマイニングとは，明示されておらず今まで知られていなかったが，役立つ可能性があり，かつ，自明でない情報をデータから抽出する技術のことである．文書分類の分野においては，与えられた文書集合から，文書間の相似性を見出したり，文書を分類するための規則を抽出する技術のことを指す．

本章では，文書分類の概要とその必要性．また一般的な文書分類の手順について説明する．

3.1　文書分類とは

近年，計算機システムの高性能化とネットワーク化にともない，膨大な電子化された情報が計算機上でアクセス可能になってきている[4]．情報量が多いため，利用者が欲しい情報を探し出すのは時間と労力を消費する．そこで，計算機を用いることにより，利用者の負担を軽減しようという研究がなされている[5][6][7]．

文書分類は，多数ある文書からそれぞれの特徴を抽出することによって文書を分類し，利用者の情報検索を助ける技術である．文書分類における“文書”とは，Web ページ，患者のカルテ，顧客データなどを指す．多くの場合，分類対象とする各文書は少なくとも数十文，多いものでは数十ページにも及ぶ．また，文書の件数は数千から数万件になる．文書分類における“分類”とは，文書の内容に応じて文書をクラスタというグループに分けることである（クラスタリングと呼ぶ）．クラスタリングを行うことにより，どのような内容の文書があるかを把握しやすくなる．

文書分類をすることで，多くの文書の中から自分が欲しい内容の文書を探し出すのが容易になる．また，どのような内容の文書が多いかなどの傾向把握が容易にできるようになる．

3.2　文書分類の方法

文書分類を行うためには，まず，文書を計算機の内部表現に変換する必要がある[3]．内部表現の変換は，文書からその意味を抽出することによって行う．このような処理は自然言語処理などを用いて行われる．現在一般的に行われているのは，文書からその内容をよく表していると考えられる要素を抽出し，その集合で文書の内容を表現する方法である．このような文書の内容を表す要素はキーワード（keyword），または索引語と呼ばれる．すなわち，文書中から抽出したキーワードの集合でその文書の内容を表現するのである．

キーワードの集合を表現する手法はいくつか存在する．文書分類において用いられる主要な方法として以下のようなものがある．

- ベクトル空間モデル

  ベクトル空間モデルでは，文書をキーワードの重みベクトルで表現する．各ベクトルの関係から各ベクトルの類似度を計算することで分類をする．
- 確率モデル

  確率モデルでは，まず各クラスタについて，どのようなキーワードが現れやすいかを数値的に表したモデルを作成しする．そして，ある文書が各カテゴリに入る確率を計算する．
- 規則に基づくモデル

  規則に基づくモデルでは，文書を各クラスタに分類するための条件を記述した分類規則を用意し，それを用いて文書を各クラスタに分類する．

この中でも，最も一般的な表現方法はベクトル空間モデルで表現する方法である．本研究でも，ベクトル空間モデルを利用した文書分類を扱う．

ベクトル空間モデルを利用した文書分類の手順を示す．

1. 文書からキーワードを抽出する．
2. キーワードの重要度に応じて，キーワードの重み付けを行う．
3. 文書間の類似度を計算し，それをもとに分類を行う．

ステップ 1，2 でキーワードの重みベクトルが生成される．このベクトルは特徴ベクトルと呼ばれる．

3.3 特徴ベクトルの生成

ベクトル空間モデルを利用した文書分類では，特徴ベクトルの生成が適切に行われなければ，いかなる類似度計算手法や分類手法を用いても適切な分類を行うことができない．そのため，本研究では，特徴ベクトルの生成手法について検討する．

特徴ベクトルを生成する手順について，以降で詳しく説明する．

3.3.1 キーワードの抽出

キーワードの抽出ステップは，文書中からその文書を特徴付けるキーワードを漏れなく抽出することが目的である．

キーワードとして，どのような単位を考えるかはシステムの設計に依存する．多くの場合，キーワードには語や複合語を用いる．英語などのように，語と語の間に空白を置く言語では，空白を手がかりに語を認定する．しかし，日本語の場合は語の間に区切りを置かないため，語の認定には別の手段が必要である．

日本語における語の認定には形態素解析器（morphological analyzer）を用いることがで

きる．形態素解析器にはさまざまなものがある．その一つである MeCab[8]をとりあげる．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一般に公開されている形態素解析エンジンである．MeCab を用いて語の認定を行った例を図 1 に示す．ここでは「他人を誹謗中傷する書き込みをしない.」という文章を例として用いている．MeCab を用いると，文章が単語ごとに分割され，左に語の表層形，二番目に語の品詞，三番目以降に品詞細分類や活用形，発音などが出力される．

このように，形態素解析器を用いることで，文書中の語とその品詞を認定することができる．品詞を認定することができるため，文書を特徴付ける上であまり役に立たない品詞（助詞や助動詞など）を排除することが可能である．形態素解析の結果から，キーワードに用いる品詞の語のみを抽出し，文書の特徴ベクトルを生成する．

一般の文書分類が対象としている文書では，キーワードに用いる品詞は名詞のみである．文書の分量がある程度以上あれば，名詞だけでも十分に文書を特徴付けるベクトルを生成することができる．

（原文）他人を誹謗中傷する書き込みをしない.

```
他人        名詞，一般，＊，＊，＊，＊，他人，タニン，タニン
を          助詞，格助詞，一般，＊，＊，＊，を，ヲ，ヲ
誹謗        名詞，サ変接続，＊，＊，＊，＊，誹謗，ヒボウ，ヒボウ
中傷        名詞，サ変接続，＊，＊，＊，＊，中傷，チュウショウ，チュウショウ
する        動詞，自立，＊，＊，サ変・基本形，する，スル，スル
書き込み    名詞，一般，＊，＊，＊，＊，書き込み，カキコミ，カキコミ
を          助詞，格助詞，一般，＊，＊，＊，を，ヲ，ヲ
し          動詞，自立，＊，＊，サ変・スル，未然形，する，スル，スル
ない        助動詞，＊，＊，＊，特殊・ナイ，基本形，ない，ナイ，ナイ
.           記号，句点，＊，＊，＊，＊，．　，．　，．
```

（表層形，品詞，品詞細分類 1，品詞細分類 2，品詞細分類 3，活用形，活用型，原形，読み，発音の順）

図 1.MeCab の実行例

3.3.2 キーワードの重み付け

抽出したキーワードが文書をどれだけ特徴付けているかを表現するため，キーワードの重要度を用いる．抽出したキーワードにそのキーワードの重要度を表す尺度を与えることをキーワードの重み付けと呼ぶ．

キーワードの重みを利用することによって，同じキーワードを含む文書でも，そのキーワードの各文書中での重要度を考慮して，文書の特徴の違いを表現することが可能になる．

重み付けの手法は，tf・idf 法が一般的に用いられている．tf・idf 法は，ある一つの文書中に繰り返し出現するキーワードは重要なキーワードであるとし，重みを重くする．また，多くの文書で使われるキーワードは特徴がないと考えて重みを軽くし，逆に，少ない文書でしか使われていないキーワードは特徴があると考えて重みを重くする．tf・idf 法については 3.4 項で詳しく説明する．

3.4 tf・idf 法

tf・idf 法は，キーワードの重み付けをするための手法としてよく用いられる．tf・idf 法は，tf 値と idf 値を計算し，tf 値と idf 値の積からキーワードの重みを決定する手法である．

(a) tf 値

ある文書中に出現するキーワードの頻度が tf（term frequency）値である．キーワードの頻度に基づく重み付けの背景には「何度も繰り返し言及される概念は重要な概念である」[9] という仮定がある．

(b) idf 値

多くの文書に登場するキーワードは，分類するための特徴としては役に立たない．あるキーワードが，どの程度その文書に特徴的に現れるかという特定性を考慮するためには，他の文書中のキーワードの分布も考慮する必要がある．df（document frequency）値は，あるキーワードが全文書中のどれくらいの文書に登場するかを表す値である．df 値が大きいほどそのキーワードの特定性は小さくなるため，df 値の逆数を用いてキーワードの重みを小さくする．こうして算出される値が idf（inverse document frequency）値である．

idf 値は以下の式によって定義される．ここで，N は対象となる全文書数であり，$df(t)$ はキーワード t が出現する文書数である．N と $df(t)$ の比の対数をとるのは，文書全体の規模に対して，idf 値の変化を小さくするためである．

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

tf・idf 法によって重み付けをする例を示す．「電子メールを使うときの注意事項は何でしょうか」という問題に対する解答が 5 つあったとし，それぞれの解答の重み付けを行う．なお，この例では文書中の名詞をキーワードとしている．

表 1 は各解答におけるキーワードの tf 値である．「相手を傷つける内容のメールを送らない」という解答では，「相手」「内容」「メール」というキーワードが 1 回ずつ使われている

ため，それに対応する tf 値は 1 になる．また，「中傷」「宛先」「注意」「ウィルス」は使われていないため，それに対応する tf 値は 0 になる．一方，「メールの宛先に注意する．宛先を間違うと違う相手にメールが届くかもしれない」という解答では，「メール」「宛先」というキーワードが 2 回使われている．そのため，それに対応する tf 値は 2 になる．「相手」「注意」は 1 回なので tf 値はそれぞれ 1，「内容」「ウィルス」は 0 回なので，それぞれ 0 になる．

表 1. 各解答におけるキーワードの tf 値

文書	相手	内容	メール	中傷	宛先	注意	ウィルス
相手を傷つける内容のメールを送らない	1	1	1	0	0	0	0
メールを送るとき，相手を中傷する内容は避ける	1	1	1	1	0	0	0
メールの宛先に注意する．宛先を間違うと違う相手にメールが届くかもしれない	1	0	2	0	2	1	0
知らない相手からのメールはウィルスがあるかもしれないので注意する	1	0	1	0	0	1	1
ウィルスについて．あやしいメールはウィルスがあるかもしれないので開かない	0	0	1	0	0	0	2

次に，idf 値を求める．各解答におけるキーワードの idf 値は表 2 のようになる．「メール」というキーワードは，すべての文書に使われているため idf 値は最低の 1 となる．同様に，多くの解答に使われている「相手」の idf 値も低くなっている．一方，キーワード「内容」「中傷」「宛先」「注意」「ウィルス」は，使われている文書数が少ないため idf 値は高くなる．

表 2. 各解答におけるキーワードの idf 値

文書	相手	内容	メール	中傷	宛先	注意	ウィルス
相手を傷つける内容のメールを送らない	1.1	1.4	1.0	0	0	0	0
メールを送るとき，相手を中傷する内容は避ける	1.1	1.4	1.0	1.7	0	0	0
メールの宛先に注意する．宛先を間違うと違う相手にメールが届くかもしれない	1.1	0	1.0	0	1.7	1.4	0
知らない相手からのメールはウィルスがあるかもしれないので注意する	1.1	0	1.0	0	0	1.4	1.4
ウィルスについて．あやしいメールはウィルスがあるかもしれないので開かない	0	0	1.0	0	0	0	1.4

表 1 の tf 値，表 2 の idf 値から， tf・idf 法における各キーワードの重みを求めると，表 3 のようになる．tf・idf 法は，このようにして重み付けを行う．

表 3. tf・idf 法による重み付けの結果

文書	相手	内容	メール	中傷	宛先	注意	ウィルス
相手を傷つける内容のメールを送らない	1.1	1.4	1.0	0	0	0	0
メールを送るとき，相手を中傷する内容は避ける	1.1	1.4	1.0	1.7	0	0	0
メールの宛先に注意する．宛先を間違うと違う相手にメールが届くかもしれない	1.1	0	2.0	0	3.4	1.4	0
知らない相手からのメールはウィルスがあるかもしれないので注意する	1.1	0	1.0	0	0	1.4	1.4
ウィルスについて．あやしいメールはウィルスがあるかもしれないので開かない	0	0	1.0	0	0	0	2.8

3.5 文書分類の先行研究

文書分類の先行研究について，いくつか紹介する．

・森田らの研究

森田らは，本研究と同様，小テストを分類の対象としている[5]．解答中に同じキーワードが存在する場合には，その解答の内容が似ているであろうと考える．受講者の解答中に出現する重要な単語をキーワードとし，そのキーワードが解答中に含まれているか否かにより，解答を分類する．

また，講師が見たときに，グループ化された解答の類似性が妥当でない場合，講師が手動でキーワードの候補の重みの変更することによって，再度分類を行うことができる．

・黒橋らの研究

黒橋らは助詞に着目してキーワードの重み付けをする手法を提案している[7]．この手法は，助詞に着目して tf・idf 法で作った特徴ベクトルにさらなる重み付けをする．例えば，「は」が付属するキーワードの重みを一定倍することで主語にあたるキーワードを中心に分類することができ，「を」が付属するキーワードの重みを一定倍することで目的語にあたるキーワードを中心に分類することができる．このように，特定の助詞に付属するキーワードの重みを重くすることで重要視したい情報をもとに分類できる．

第４章　記述式小テスト解答の分類

この章では，記述式小テスト解答分類の特徴，従来法で記述式小テスト解答を分類するにあたっての問題点，また，それを解決するための手法を提案法する．

4.1　記述式小テスト解答分類の特徴

3 章では文書分類の一般的な手法について説明した．しかし，一般的な手法が対象としている文書の分類に対して，本研究で対象とする記述式小テストの解答の分類は異なる．記述式小テストの分類は以下の 3 つの特徴を持つ．

第 1 に，分類を短時間で終える必要がある．講師がすみやかに学生にフィードバックするための手助けとして，解答を分類する．そのため，分類に時間がかかることは好ましくない．本研究では，分類をリアルタイムで行うことを目標とする．

第 2 に，分類対象は小規模なデータである．一般的な文書の分類では，件数の多いデータを対象とすることが多い．また，文章の長さは多いもので数ページにも及ぶ．一方の記述式小テストの解答は，件数が少なく，一般の文書分類の対象と比べて文章の長さが短い．例えば，「インターネットを利用するときに、情報倫理に関して気をつけなくてはいけないことは何か？」という問に対する記述式小テストに対する平均的な長さの解答は次のようになった．本研究では，学生の人数は 100 名程度，学生の解答の長さは 2，3 文程度を想定している．

（答）私が一番重要だと思うのはモラルです．匿名性が高いので他人を傷つけたりする人がいますが，現実世界と同じく傷つけられた人の気持ちを考えて，するべきではないと思います．

第 3 に，分類対象は特定の内容に偏っている．一般的な文書の分類は，多種多様な文章を対象とすることが多い．しかし，小テストの解答は，ある問に対する解答という意味では一つの内容である．本研究では，解答群中の主要な内容を把握できることを目標とする．

4.2　従来法で記述式小テスト解答を分類する問題点

記述式小テストの解答は小規模であるため，その解答に含まれる名詞の使用回数のみに着目するだけでは情報が足りないことが考えられる．少ない情報で特徴ベクトルを生成すると，異なった内容の文書が同じクラスタに分類されるなど，誤分類が発生する可能性が高まる．

また，1 つの文書に含まれる単語の数が少なく，重要である語と重要でない単語とでは出

現頻度がさほど変わらない．この場合，tf 値の「何度も繰り返し言及される概念は重要な概念である」という仮定が成立しない．したがって，tf・idf 法で適切な重み付けをすることができない．

3.5 項で紹介した先行研究の手法で分類する場合でも問題が生じる．森田らの手法では，多くの解答に含まれている単語が重要な単語でない場合，適切に分類することができない．また，記述式小テストでは，学生によってさまざまな表現がされるため，黒橋らの特定の助詞に注目する手法では適切な重み付けができない．

そこで，記述式小テストのような小規模のデータでも，文書の特徴を適切に抽出できる手法が必要である．

4.3 提案手法

4.2 で述べたように，一般的に行われている文書分類の手法では，記述式小テストの解答分類には適していない．そこで，次の二つの考え方に基づき解答の特徴ベクトルを生成する．

1. キーワード対象の品詞の拡大
2. 係り受けに基づく重み付け

以下でそれぞれについて詳しく説明する．

4.3.1 キーワード対象の品詞の拡大

小テストの解答にはそれほど多くの単語が含まれていない．その中から名詞のみを取り出して分類に用いるのでは，情報が足りないことが考えられる．そこで，解答から名詞以外の単語も抽出することで，特徴の量を増す．

抽出する品詞は名詞と動詞とする．これは，文章の主な構成要素は主語（名詞）と述語（動詞）であるためである．ただし，動詞の活用形を考慮に入れると複雑になりすぎるので，語幹のみを扱うものとする．

具体的な例を用いて説明する．「電子メールを使うときの注意事項は何でしょうか」という問題に対する二つの解答，

・むやみに知らない人にメールを送らない

・知らない人からのメールに注意する

について考える．これらの解答の内容は，送信時の注意と，受信時の注意なので，異なる特徴ベクトルを生成すべきである．これらの解答に含まれる名詞は「知らない人」と「メール」のみであり，いずれの解答にも含まれる．そのため，同じ特徴ベクトルを生成してしまう．しかし，動詞も含めることで，共通に含んでいない単語「送る（送らない）」,「注意する」も用いるようになる．そのため，異なる特徴ベクトルを生成する．

4.3.2 係り受けに基づく重み付け

係り受けに基づく重み付けは，単語の出現回数以外に，解答中での文節の係り受けの関係に着目する方法である．具体的には，より多くの文節からの修飾を受けている文節ほど，その解答で中心的な働きをしていると考え，それに含まれるキーワードを重要視する．その結果，ささいな修飾語の違いに惑わされず，その文で表したかった内容に沿ったベクトル化が行われる．

具体的な例を用いて説明する．「電子メールを使うときの注意事項は何でしょうか」という問題に対する解答，

・著作権を侵害するような内容のメールを送ってはならない

・著作権の侵害になる場合があるので，送られたメールの内容を公開してはならない

について考える．主要な内容は「メールを送ってはいけない」,「内容を公開してはいけない」と異なるので，異なる特徴ベクトルを生成するべきである．これらの解答からは，使用単語（名詞，動詞）の重複が多く，頻度もすべて 1 回なので，tf・idf 法は類似した特徴ベクトルを作成しまう．しかし，係り受け解析の結果，図 2 に示したように，文末の「メールを送ってはいけない」,「内容を公開してはいけない」の部分に多くの文節が修飾していることが分かる．そのため提案法では，それらの単語の重みが大きくなり，異なる特徴ベクトルを生成する．

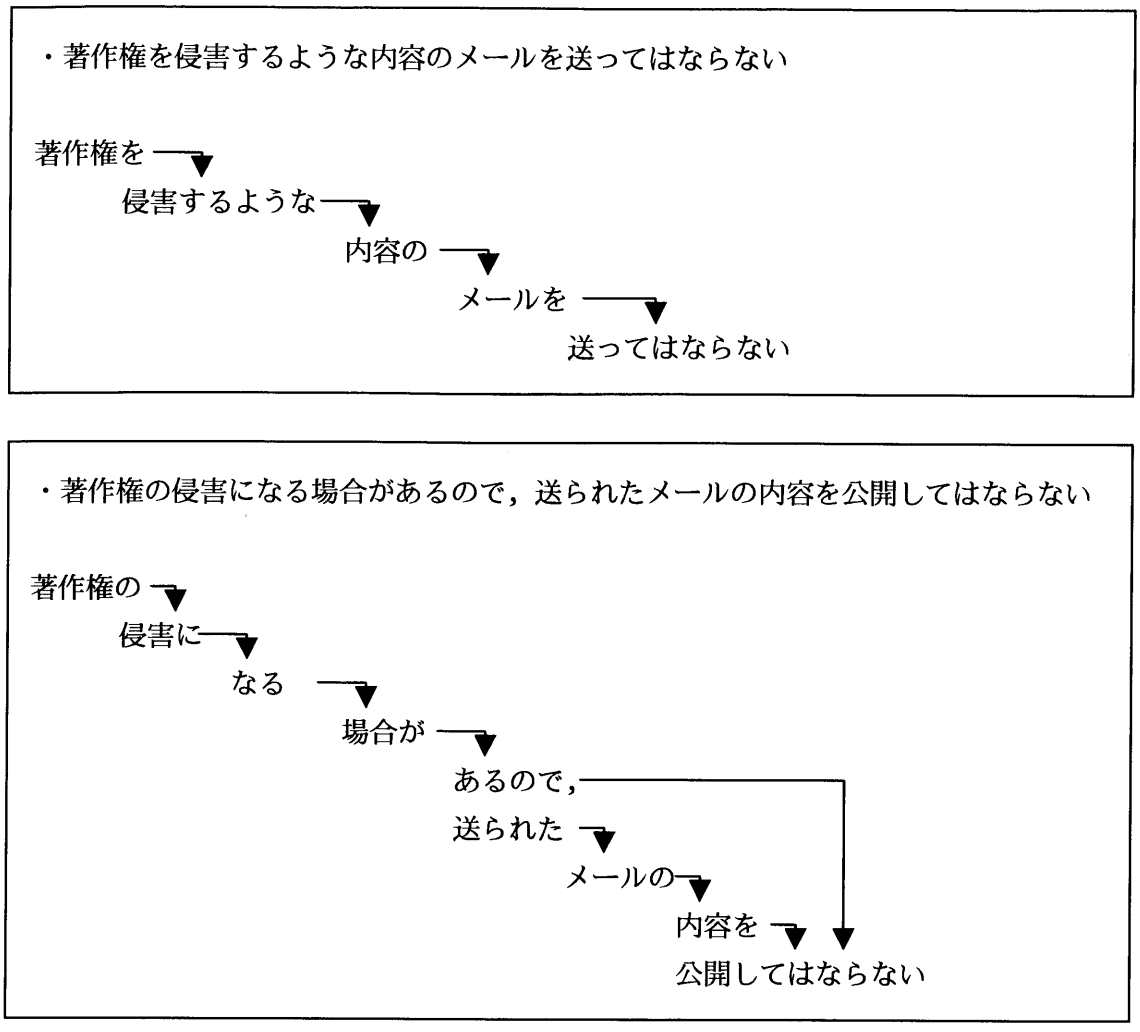

図 2. 係り受け解析の例

本研究では，係り受けの解析に Cabocha[10]という係り受け解析器を用いる．Cabocha は一般に公開されている日本語の係り受け解析器であり，高い解析精度をもつ．

本研究では具体的に，以下の手順に従い各キーワードの重みを計算する．

1. 各解答文を Cabocha で解析する．
2. 各文節について，被修飾回数を 1 とする．
3. 修飾されていない文節から順に，その文節の被修飾回数を，その文節が修飾している文節の被修飾回数に加算する．
4. 一文ごとに，各文節の被修飾回数を[0.5, 1]に正規化する．各キーワードの重みは，そのキーワードが属している文節の正規化した被修飾回数とする．

係り受けに基づく各キーワードの重み付けの例を図3に示す.各節の左側の数字が被修飾回数である.「著作権の」は「侵害に」に係っているため,「著作権の」の被修飾回数である1を加算し,「侵害に」にの被修飾回数は2になる.「侵害に」は「なる」に係っているため,「侵害に」の被修飾回数を加算し,「なる」の被修飾回数は3になる.以下同様にして,図3の修飾回数が求められる.ここから,各節の被修飾回数を[0.5, 1]に正規化し,キーワードの重みを節の修飾回数として重み付けを行う.重み付けの結果は表4のようになる.

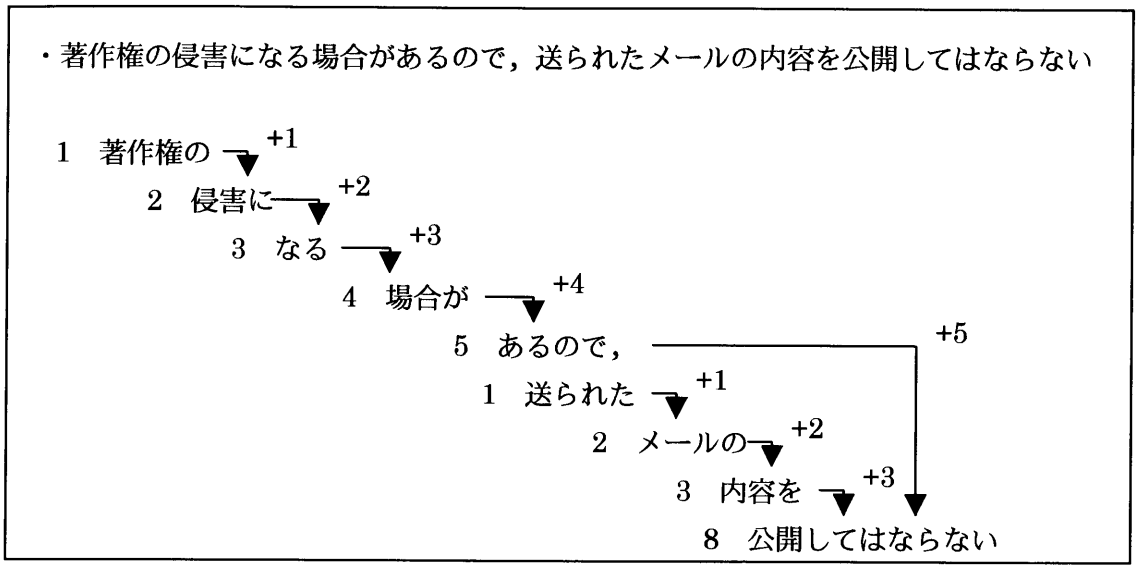

図3. 係り受け解析による重み付けの例

表4. 係り受け解析による重み付けの結果

著作権	侵害	なる	場合	ある	送る	メール	内容	公開
0.5	1.0	1.5	2.0	2.5	0.5	1.0	1.5	4.0

前述の通り,分類にはあまり時間をかけることはできない.しかし,小テスト解答は文書が短いため,係り受けを解析しても分類に要する時間はそれほど長くならない.

第5章　実験

ベクトルを生成する手法として，提案法と tf・idf 法で同じ記述式小テストの解答の特徴ベクトルを生成し，それをもとに分類を行う．その結果から，提案法の有効性を確認する．

5.1　実験条件

同じ記述式小テストの解答を提案法と tf・idf 法で分類する．そして，分類結果の違いを生成された特徴ベクトルの違いから検証する．分類対象は三重大学工学部電気電子工学科 1 年生を対象に開講された講義「計算機基礎 I 及び演習」で行われた記述式のテストとする．使用した計算機は Pentium4 2.4Hz であり，OS は Windows XP Professional である．また，提案手法での重み付けに MeCab と Cabocha を用いており，類似度の計算と出力はオープンソースの統計解析システムである R[11]を使用している．

分類手法はデンドログラムを用いる．デンドログラムでは，より内容が近いと判断された解答ほど，近い距離（木の下のほう）で各解答に対応した枝が交わる．したがって，低いところで枝が交わっているほど，それらの枝に対応する解答が類似していることを意味する．

「インターネットを利用するときに，情報倫理に関して気をつけなくてはいけないことは何か？」に対する，86 名分の解答を分類する．

デンドログラム中の数字は解答の番号であり，先頭の A〜E のアルファベットは，解答の内容に応じて筆者が分類し，ラベル付けしたものである．解答の内容は表 5 のように対応している．また，解答をした人数も同様に示す．

表 5. デンドログラムのラベル，解答の内容，解答人数の関係

A	プライバシー，個人情報	10 人
B	著作権，肖像権	35 人
C	誹謗・中傷	8 人
D	情報の信頼性	27 人
E	複数の内容，少数解答	6 人

なお，全ての解答を入力してからデンドログラムを作成するまでに要した時間は，約 5 秒であった．これは，小テストの分類はリアルタイムで行うという要件を十分に満たす．

5.2 実験結果

デンドログラムの出力を図4に示す．tf・idf法の結果では，右端の枝にBの解答が集まっている．しかし，その枝から漏れているBの解答が14個存在するうえ，右端の枝にはBの内容でない解答が5個存在する．

一方，提案法では，左端にBの内容である枝ができた．そこに漏れたBの内容である解答数は14個で，tf・idf法と変わらないが，左端の枝の中にはBの内容以外の解答は存在しない．このように評価した場合，tf・idf法よりも，提案手法の方が優れていると言える．

デンドログラムの結果を定量的に評価することは難しいので，デンドログラムの一部分を詳細に検討する．

tf・idf法において，最も低い高さにある誤分類した解答をとりあげる．ここで，誤分類とは，低い高さで交わった二つの枝に対応する解答について，それらの人手による分類結果が異なっている状況を意味する．これは，tf・idf法では図4(a)のA-186とD-114（図中のX1，X2）が該当する．また，これらの具体的な解答は以下に示す通りである．

表6. A-186とD-114の解答の内容

A-186 (X1)	最も重要だとおもうものはプライバシー．
D-114 (X2)	インターネットを使って、何かを調べている時には、すべての情報が信頼できるわけではないので、簡単に信じてはいけないということ。

tf・idf法の結果，A-186はDが多く含まれる枝の中に分類されており，D-114が最も近い距離であると判断されている．しかし，Aの内容であるこの解答がこの位置に分類されるのは好ましくない．一方の提案手法では，A-186はAが多く含まれる枝に分類されており，最も近い解答はA-113である．いずれも人手による分類結果と一致しており，好ましい．

この結果を，各手法で抽出したキーワードに基づいて分析する．

A-186からは，「重要」「プライバシー」がtf・idf法により抽出されている．また，tf・idf法で誤って似ていると判断したD-114からは，「インターネット」「すべて」「情報」「信頼」「簡単」が，抽出されている．単語の重複が少ないため，異なる解答とみなすことができそうだが，この中で，「インターネット」と「情報」については，半数近くの学生の解答に用いられており，idf値が小さい．そのため，分類に際して重要視されない．また，「簡単」はD-114以外の解答には含まれていないため，今回は分類に用いていない．結局，tf・idf法におけるA-186とD-114差は，実質4単語分となる．図4より，この差は，tf・idf法のデンドログラム上での高さにして，5に満たない近さとなる．これは他の解答間の差と比べると低い高さであり，結果として，近い解答と誤判断してしまった．

これに対して，提案手法では，動詞も分類に加え，修飾語の重みを小さくしている．その結果，各解答の「おもうのはプライバシー」「信頼できるわけではないので、」の部分に含まれる単語を重要視して分類する（抽出される単語は下線を引いた部分）．そのため，これらの解答を異なるものとして判断できた．

表 7. A-186 と D-114 の特徴ベクトル

(a) tf・idf 法

A-186	重要	プライバシー		
(X1)	1.9	2.7		
D-114	情報	インターネット	信頼	すべて
(X2)	0.7	0.8	3.1	2.9

(b) 提案手法

A-186	重要	プライバシー	おもう				
(X1)	1.3	2.7	3.4				
D-114	情報	インターネット	信頼	すべて	調べる	使う	信じる	すべて
(X2)	0.4	0.4	3.1	2.9	1.1	1.0	1.6	1.4

また，同様に tf・idf 法において，低い高さにある誤分類した解答をもう 1 つとりあげる．図 4(a)の D-119 と B-148（図中の X3，X4）に注目する．これらの具体的な解答は以下に示す通りである．

表 8. D-119 と B-148 の解答の内容

D-119 (X3)	私が最も重要だと思うことは、インターネット上の情報をすべて鵜呑みにしないということである。自分にとって大事な決断をする時人の情報を参考にすることはよくあり、また大事なことでもある。しかし、インターネット上の情報がすべて本当だとは限らない。そういう時それを鵜呑みにして後悔しても遅いからきちんと、見極めることが大事だと思う。
B-148 (X4)	情報倫理に関して気をつけなければいけないことで，自分が最も重要だとおもうのは「著作権」だと思います．例えば自分がある情報についてのレポートを作っているときその情報が書かれた本をまるまる写したら著作権の侵害になります．そうならないためにも，自分の文書で書き，参考にした本はしっかりと書くべきです．

D-119 と B-148 には，共通して使われている名詞が多く存在する（表中の下線を引いた部分）．それは「重要」「情報」「自分」「参考」が該当する．キーワードが名詞のみである tf・idf 法では，似た特徴ベクトルが生成され，同じ解答であると分類された．しかし，解答の内容が異なっているため，これは誤分類である．

これに対して，提案手法では，「限る」「書く」などの動詞が抽出されているため，2 つの解答の違いがより表現できている．また，修飾関係による重み付けにより，「鵜呑み」などのキーワードの重みが tf・idf 法と比べて重くなっている．また，「重要」「情報」「自分」「参考」のキーワードの重みは tf・idf 法と比べて軽くなっている．そのため，これらの解答を異なるものとして判断でき，適切なクラスタに分類できた．

表 9. A-186 と D-114 の特徴ベクトル

(a) tf・idf 法

D-119	重要	情報	自分	参考	鵜呑み	インターネット
(X3)	1.9	2.1	0.9	3.1	3.3	3.6

B-148	重要	情報	自分	参考	著作権	レポート	気	侵害	文書	情報倫理
(X4)	1.9	1.4	2.6	3.1	2.0	1.9	1.8	1.7	3.1	3.1

(b) 提案法

D-119	情報	思う	する	参考	ある	自分	重要	鵜呑み	すべて	限る	本当	インターネット
(X3)	1.4	2.0	2.3	1.7	1.8	0.5	1.2	3.6	3.0	3.8	2.2	2.0

B-148	情報	思う	する	参考	ある	自分	重要	つける	著作権	書く	作る	レポート
(X4)	0.9	1.1	0.6	1.7	0.6	1.4	1.5	1.2	1.5	3.2	1.4	1.2

気	なる	侵害	文書	情報倫理	おもう
1.0	2.2	1.2	1.8	1.7	3.4

以上の結果から，tf・idf 法で誤分類した例の一部を正すことができたことがわかる．抽出するキーワードを増やし，適切な部分を重要視して分類を行うことで，期待した分類が行われることがわかる．

一方で，提案手法で誤分類をした例をとりあげる．図 4(b)の D-124 と C-131（図中の X5，X6）に注目する．これらの具体的な解答は以下に示す通りである．

表 10. D-124 と C-131 の解答の内容

D-124 (X5)	何かを調べる時、インターネットを利用する場合、みんなが見てるからといって、正しい情報とはかぎらないといううことを頭にいれて調べる必要がある。
C-131 (X6)	電子掲示板で自分の顔が見えないからといって、相手を中傷する言葉などを使ってはいけない。

D-124 と C-131 は,「いう」というどの解答にも現れる可能性があるキーワードの重みが特に重くなっている．また，他には重くなっているキーワードがない．したがって，2 つの解答の距離が近くなり，誤判断された．また，D-124 には「正しい情報とはかぎらない」という表現が使われているが,「かぎる」というキーワードを使用した解答は D-124 のみであるため，今回は分類に用いていない．これがもし「限る」と表現していた場合，情報の信頼性について書いている解答によく登場するキーワードであるため，分類に使用することができ，誤分類を起こさなかったと考えられる．

このように，重みが重くなっているキーワードが少ない解答では，誤分類が起きやすくなっていた．また，変換ミスや表現の違いによって別のキーワードであると判断されることがあり，そのような解答は誤分類されやすかった．このような解答は，提案手法に限らず，tf・idf 法でも同様に誤分類をすることが多い．本来は，このような解答であっても適切に分類することが望ましい．したがって，このような解答を分類できるようにすることが今後の課題といえる．

表 11. D-124 と C-131 の特徴ベクトル

D-124	いう	見る	調べる	情報	ある	利用する	インターネット	いれる
(X5)	4.2	1.1	2.4	0.4	1.0	1.0	0.4	2.3
C-131	いう	見える	言葉	自分	使う	中傷する	相手	
(X6)	2.6	2.7	2.5	0.5	1.7	2.2	1.7	

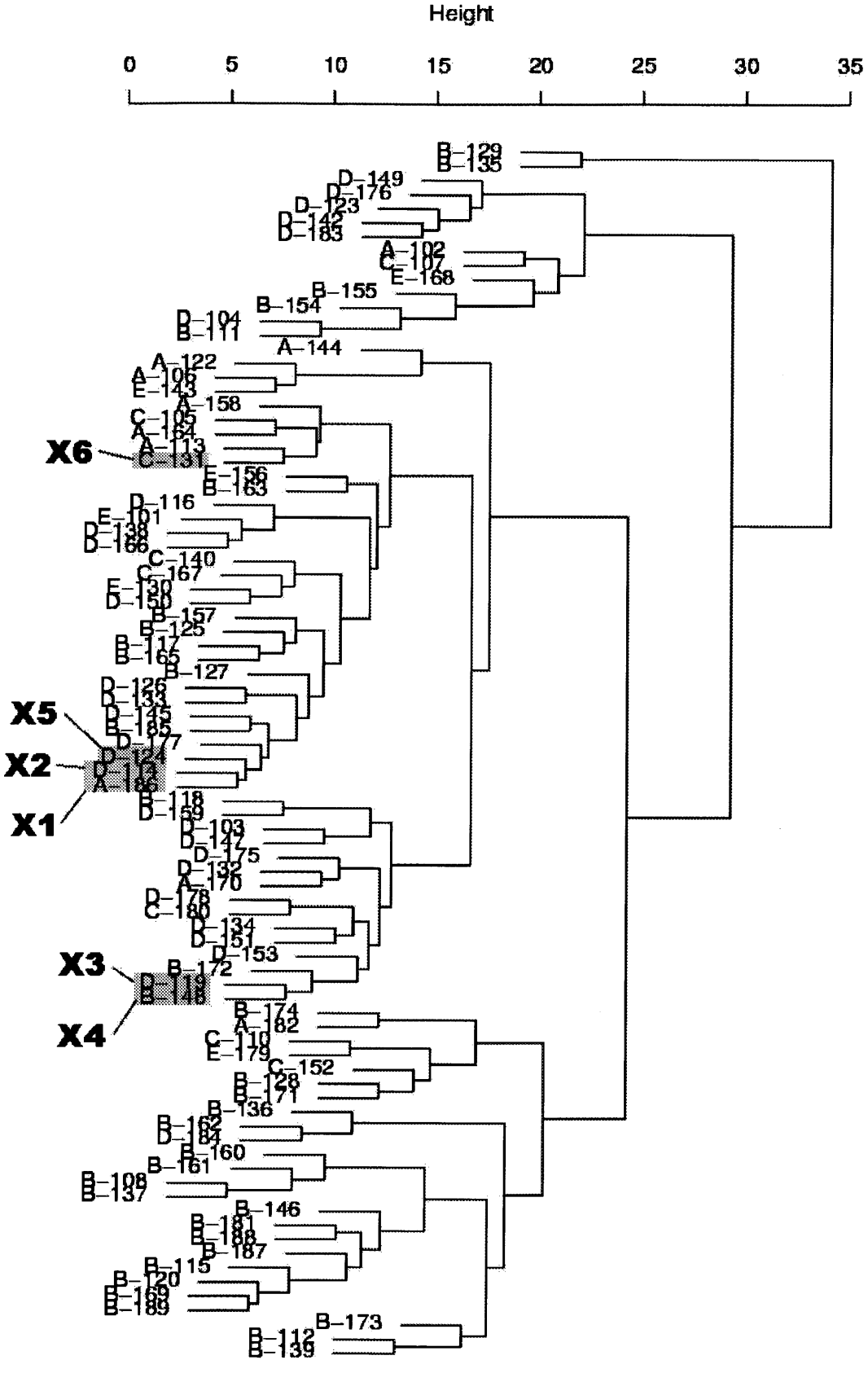

図 4(a). tf･idf 法によるデンドログラムの出力

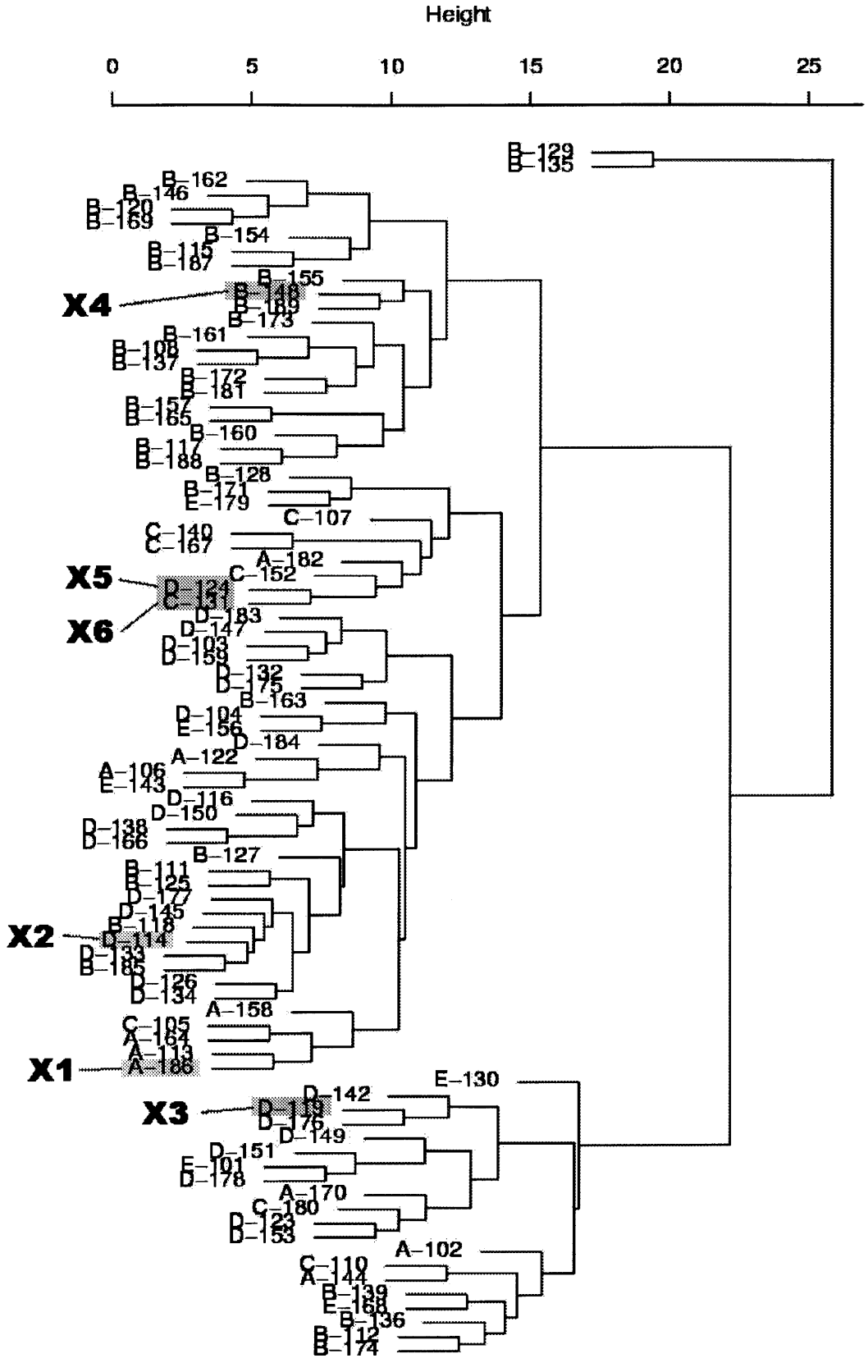

図 4(b). 提案手法によるデンドログラムの出力

第6章　まとめ

わかりやすい講義を行うためには，講師は学生の理解度を把握する必要がある．学生の理解度を把握する有効な手段の一つに記述式小テストがある．本研究では，記述式小テストの解答を分類して講師に提供することで，理解度把握のための助けを行うことを目的とした．

文書分類には特徴ベクトルの生成が重要である．しかし，従来の特徴ベクトル生成法は，記述式小テストの解答を分類するには適していない．そこで新たな手法を提案した．その手法は以下の方法で特徴ベクトルを生成する．

・キーワード対象の品詞の拡大

・係り受けに基づく重み付け

実際の記述式小テストの解答を用いて，提案手法で分類を行った．そこで，提案手法の有効性を確認した．

参考文献

[1] 文部科学省の Web ページ,「大学における学生生活の充実方策について（報告）－学生の立場に立った大学作りを目指して」(2007 年 2 月現在)
http://www.mext.go.jp/b_menu/shingi/chousa/koutou/012/toushin/000601.htm

[2] 和田秀樹, 西村和雄, 戸瀬信之 :「数学軽視が学力を崩壊させる」, 株式会社講談社 (1999)

[3] 徳永健伸 : 情報検索と言語処理, 東京大学出版会 (1999)

[4] 北研二, 津田和彦, 獅々掘正幹 : 情報検索アルゴリズム, 共立出版 (2002)

[5] 森田直樹, 北英彦, 高瀬治彦, 林照峯 : 記述式の解答を即時に講師が把握するためのシステム, FIT 情報科学技術フォーラム (2002)

[6] 神嶌敏弘 : データマイニング分野のクラスタリング手法(1) クラスタリングを使ってみよう！, 人工知能学会誌, Vol.18, No.1, pp.59-65 (2003)

[7] 黒橋禎夫, 中村順一, 長尾真 : 構文情報を利用した電子ニュース記事のクラスタリングシステムの作成と評価, 電子情報通信学会技術研究報告, 言語理解とコミュニケーション, Vol.98, pp.15-22 (1998)

[8] 中川 哲治, 工藤 拓, 松本 裕治 : Support Vector Machine を用いた形態素解析と修正学習法の提案, 情報処理学会論文誌, Vol.44, No.5, pp.1354-1367 (2003)

[9] Luhn, H. P.: A statistical approach to mechanized encoding and searching of literary information, IBM Journal of Research and Development, Vol. 1, No. 4, pp. 390–317 (1957)

[10] 工藤拓, 松本裕治 : チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol.43, No.6, pp.1834-1842 (2002)

[11] 統計解析システム R 言語の Wiki のページ, 「RjpWiki」(2007 年 2 月現在)
http://www.okada.jp.org/RWiki/

発表論文

[i] 大井健太郎, 高瀬治彦, 森田直樹, 北英彦, 林照峯：記述式小テストシステムにおける解答の自動分類に関する一考察, 平成 17 年度三重地区計測制御研究講演会講演論文集, pp.B10-1～B10-4 (2005)

[ii] 大井健太郎, 高瀬治彦, 森田直樹, 北英彦, 林照峯:記述式小テストの解答自動分類のための特徴抽出に関する一考察, 2006 PC カンファレンス論文集, pp.449-452 (2006)（2006 PC カンファレンス優秀学生論文賞受賞）

[iii] 大井健太郎, 高瀬治彦, 北英彦, 林照峯:小テスト解答の自動分類のために抽出する解答の特徴抽出に関する一考察, 平成 18 年度三重地区計測制御研究講演会講演論文集, pp.P-21-1～P-21-6 (2006)

[iv] 高瀬治彦, 大井健太郎, 森田直樹, 北英彦, 林照峯：記述式小テストの解答を自動分類するための特徴抽出, CIEC, Computer & Education Vol.21（採録決定）

謝辞

本論文は，著者が三重大学大学院工学研究科博士前期過程に在学中に行った研究をまとめたものである．本研究を進めるにあたり，懇切丁寧なご指導とご督励を賜った三重大学林照峯教授，北英彦助教授，高瀬治彦助手に深く感謝いたします．また，日頃熱心に討論して頂いた計算機工学講座の皆様方にお礼申し上げます．

そして，貴重な時間を割いて本研究論文の査読をして頂いた情報処理研究室の鶴岡信治教授に深く感謝いたします．

最後に，本論文をまとめるにあたり，助言，討論，その他お世話になったすべての方々に感謝いたします．