レポートのグループ化によるレポート採点支援に関する研究

広兼　崇博*　　能瀬　高明**　　森川　富昭***　　矢野　米雄****
*徳島大学工学研究科知能情報工学専攻　**徳島大学工学研究科情報システム工学専攻
***徳島大学病院医療情報部　****徳島大学工学部知能情報工学科
t_hirokane@is.tokushima-u.ac.jp

1. はじめに

現在, 大学では盛んにe-Learningを教育支援システムとして取り入れている[1]．徳島大学においても，レポート提出・授業資料閲覧などを支援するe-Learningシステムが導入されている．e-Learningは授業時間に限らない授業資料等の閲覧，メールや掲示板を利用した学生同士のコミュニケーション，教師へ容易にレポートの提出や質問ができるなど教師と学生の教育を支援する双方向なシステムである．一般的に，e-Learning システムでは，レポートの提出や質問など学生に対する支援は充実しているが，授業の準備や，研究，授業以外の仕事などにより多忙な教師に対する支援は充実しているとは言えない．大学などの高等教育機関における講義や演習では，授業を受けた学生に対して課題を出題し，レポートを提出させることが多い[2]．　しかし，教師にとってレポートの採点は，多くの時間を要し，負担となっている．

多くのレポートを採点していく中で，最初に設定した採点基準を最後まで一貫させることは困難であり，レポートの採点途中で採点基準を設定し直したいと感じる例が多くある．我々は，レポートの採点に多くの時間を割いている教師の負担を軽減することを目標とする．そこで，レポートのグループ化によってレポートの採点を支援するモデルを提案する．レポートをグループ化することで教師は，レポートの採点基準を決定しやすくなり，途中で採点をやり直すという事態を避けることができる．本研究の目的は，レポートをグループ化する手法を検証することである．

また，近年では電子データによるレポートや資料などが多いため，学生は他人のレポートや資料をそのままコピーして提出することが容易である．そこで，我々は類似性の高いレポートをグループ化することで, 不正なレポートを検出できるかを検証する．

本研究では，実際に学生から提出されたレポートを用いて, グループ化を行い, 検証を行った. また, レポートからキーワードを抽出する手法として, n-gram 解析を用いた．グループ化の手法は，クラスター分析を用いた．

2. n-gram 解析によるテキスト解析

本システムでは，レポートからキーワードを抽出する手法として，n-gram 解析を用いたテキスト解析を行う．

2.1 n-gram 解析とは

ある文章が与えられたとき，その文章の特徴を定量化するのは困難である．もっともわかりやすい評価法の一つは, 文字の出現頻度を調べる方法である．違う人が同じテーマで似た文章を書いたとしても，個人によって文章表現の癖や個性が現れるのはごく自然なことである．文章中でよく使用する単語や表現があれば，そうした単語や表現を記述するための文字が頻繁に出現する．文章を評価するためのアルゴリズムとしては，文章を構成する文字を先頭から順に数え上げれば良い．

このアルゴリズムを用いて，文字を順に数え上げるとき，1 文字ずつ抽出するのではなく，数文字についてその出現頻度を調べればより文章の特徴を明確に抽出することが可能である．n文字の並びをn-gramと呼ぶ．n の値をいくつにするかによって解析結果は異なるが，一般的に日本語ではn=3 程度が用いられる[3]．また, n=1, 2, 3 の場合をそれぞれ，unigram，bigram，trigramと呼ぶ. trigramの例を図 2.1 に示す.

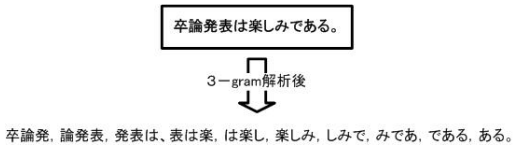

図 2.1　**n=3 の n-gram 解析の例**

2.2 本研究における n-gram の導入

本研究でテキスト解析の手法として n-gram 解析を用いた理由を述べる．

レポートの解析を行うとき，日本語や中国語などの場合，文章から単語を正確に抜き出して文章を評価する手法として形態素解析(morphological analysis)がある．形態素解析とは，文を適切な形態素に分割する処理のことである．この形態素とは，意味をもつ最小の要素のことで，文章はこの形態素で構成されている．形態素解析は，自然言語の分野では活発に研究開発が行われている[4].

しかし，先述したように，日本語には単語間に明示的な区切りを入れる習慣がないため，単語の定義が明確ではない．そのため，形態素解析によるテキスト解析は，使用した形態素解析システムに依存する．また，同一システムで同じような文字列でも，文脈によりことなった分割が行われる可能性があるという問題点がある[4]． そこで，本システムでは形態素解析を使用せずにキーワードを抽出する手法としてn-gram解析を用いた．

さらに，レポートのグループ化にn-gram解析を用いると, 単なる文字列比較では検出できないような, レポートの類似性を検出できる場合がある. これは, n-gram解析による文字の出現傾向は，対応するレポートの文字列の傾向であり，レポートの特徴となるからである[2].

英語の形態素解析のシステムとしては，Brill's Taggerがある[5]．日本語の形態素解析のシステムとしては，「茶筌」がある[6].

3. システム概要

本研究で用いたシステムは，Webブラウザで使用するe-Learningシステムを想定して，PHPとMySQLで構築を行った．システムをモジュール毎に分類すると“Txt変換モジュール”, “n-gram解析モジュール”, “クラスター分析モジュール”となる．図 3.1にシステムのフロー図を示す.

次に各モジュールの説明を行う．

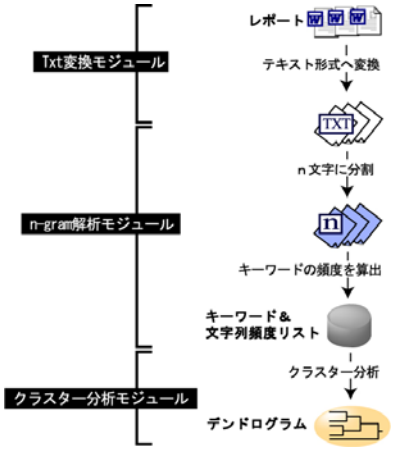

図 3.1　**システム概要**

3.1 Txt 変換モジュール

本システムでは，学生から提出されるレポートの形式の Word ファイル Doc 形式を対象とする．

DocファイルからTxtファイルに変換するために，wvwareというツールを使用した．wvwareとは，WordのDocファイルをHTMLのデータやText，PDFなどに変換／コンバートできるツールである[6].

3.2 n-gram 解析モジュール

システムは，各レポートの文章をn-gram解析によりキーワードを抽出する．今回は，日本語文章の特徴を考慮してn=3 とする[3].

n-gram 解析モジュールは，最初に，学生から提出されたレポートに対してそれぞれ trigram 解析を行い，キーワード毎に分割する．それらのキーワードを MySQL のテーブルに登録していく．ただし，既に登録されているキーワードは重複して登録しない. このとき同時に，キーワードの頻度は 0 に初期化しておく．最後に，レポートに対するキーワードの頻度を算出する．まず，もう一度 n-gram 解析を行ったキーワードを順番に，登録したテーブルのキーワードと比較する．そして，テーブルに登録されているキーワードの頻度を増やし，テーブルを更新する．

3.3 クラスター分析モジュール

システムは n-gram 解析モジュールで解析された文字列頻度のデータから式(1)を用いてユークリッド距離を求める．

$$\text{距離}(x, y) = \sqrt{\sum (x_i - y_i)^2} \qquad (1)$$

求めた距離をウォード法により分析する．分析結果は，図 3.2のような形式で出力する．

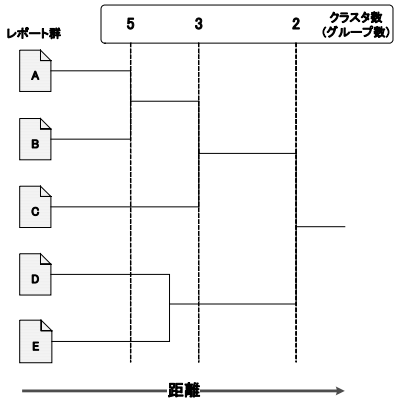

図 3.2　**クラスター分析によるデンドログラム**

4. 評価実験

本研究では評価実験を 2 回実施した．

評価実験 1 として，2004 年度前期，徳島大学医学部･歯学部附属病院 歯学部 1 年生の「医療情報処理」の授業で課されたレポート「電子カルテについて」に対して行った．49 人分のレポートを対象とした．レポートの量は, A4 レポート用紙に 1 枚から 2 枚で，文字数は 1000 文字前後のレポートである．

評価実験 2 では, 2004 年度後期, 徳島大学医学部・歯学部附属病院 医学部保健学科の 2 年生と編入生の 3 年生の「情報処理・統計学演習」の授業で課されたレポートで行った．レポート課題は「将来，統計解析や根拠に基づく医療と知識が，どのような場面で活躍できて, 役立てることができるかについて」である．また，評価実験 2 では，任意に選択した 5 つのレポートに対して，文の入れ替え・複文を短文に変換・語調を変換などの，いわゆる不正なレポートを作成して，同時に解析を行った．全部で 90 人分のレポートに対して解析を行った. レポートの量は, A4 のレポート用紙に 1 枚から 2 枚程度で, 文字数は 1000 文字前後のレポートである．

実験手順は，前章で述べたシステムの流れに従って行った．解析結果をデンドログラムで出力し，そのグループ分けについての考察を行った．

4.1 実験結果

評価実験 1 と評価実験 2 の結果で出力されたデンドログラムをそれぞれ図 4.1，図 4.2に示す．

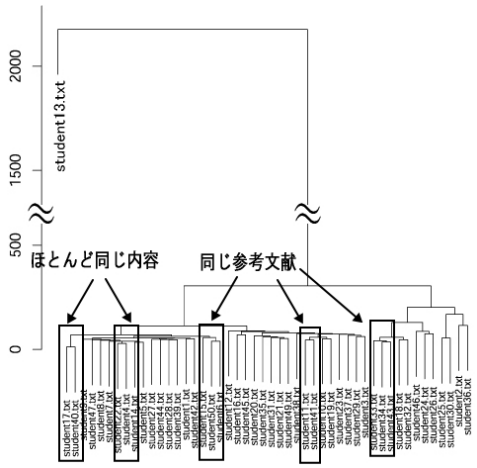

図 4.1　**評価実験 1 の結果**

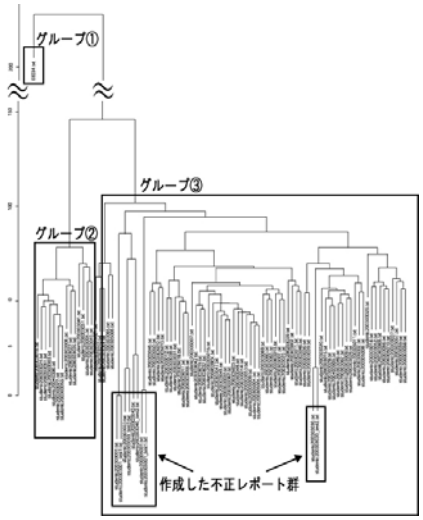

図 4.2　**評価実験 2 の結果**

5. 考察

5.1 評価実験 1 における考察

評価実験 1 では，全体的にレポートの内容については，大きな違いは見られなかった．これは，レポートの課題が「電子カルテについて」という，調査したものをレポートに書くという，いわゆる調べ物のレポートであったことが原因であると考えられる．図 4.1から，最も近い距離でグループ化されているレポート群を見てみると，レポート内容が全く同じで，感想を述べている部分だけが異なるといった，明らかに他人のレポートをコピーしたと考えられるレポートを見つけることができた．その他にも，同じ参考文献から引用したと考えられるレポートをグループ化することができた．中には，同じ参考文献から引用したと思われる部分が見られたが，グルー

プ分けが異なるというレポートも存在した. これは, 同じ参考文献を用いた部分が, 全体の文章からすると小さい割合であったためであると考えられる.

このように, 調べ物のレポートに対しては, 同じ参考文献を用いたレポート, 他人のレポートをコピーしたと思われるレポートをある程度グループ化することができた.

5.2 評価実験 2 における考察

評価実験 2 では, 作成した不正レポート群が他のレポートと異なり, かなり近い距離でグループ化されていることがわかる. このことから, 今回のレポートでは, 他人のレポートをコピーしたようなレポートは存在しないということがわかる. これは本実験が, 評価実験 1 のような単純な調べ物のレポート課題ではなかったためであると考えられる.

評価実験 2 のレポートのデンドログラムから, 図 4.2のグループ①は, 他のレポートと比べて, 全て自分の言葉で書いてあり, レポート課題とは無関係な話題を書いている部分もあった. つまり, このレポートは, 他のレポート群とは異なる独自性の強いレポートであると言える. このことから, グループ①のレポートは他のレポート群とはかなり遠い距離でグループ化されたものであると考えられる. 図 4.2 のグループ②とグループ③のグループ分けでは, 自分の考えを多く述べているグループ②と, そうでない参考文献から調べたと見られる内容が多く述べてあるグループ③に分けることができると考えられる. これは, グループ②には,「考えられる」等の自分の考えを述べるときに用いられる単語が多用されているレポートが多く含まれているためである. グループ③の中には, 同じ文献を参考にしたと考えられるレポートをグループ化することができた.

また, 検査技師, 看護師を目指す学生が多いグループが存在するという結果も得ることができた. これは, 評価実験 2 で対象となった学生は, 看護師や検査技師を目指す学科の学生であり, 将来の目標の違いによって, レポート課題のとらえ方が異なったためであると考えられる.

このように, 評価実験 1 とは違い, 幅広い意味を持ったレポート課題に対しても有効なグループ分けができることがわかった. 評価実験 1 のときのような不正なレポートは存在しなかったが, 自分の考えを多く述べているレポート群や幅広い意味を持ったレポート課題であったにもかかわらず, 似たような分野について述べているレポートをグループ化することができた. このような結果から, 評価実験 2 の結果は n-gram 解析を用いたテキスト解析の性質である, レポートの文章の特徴によるグループ化を行うことができた.

6. 結論

本研究では, n-gram 解析を用いたレポートのグループ化の手法についての検証を行った.

評価実験の結果から, n-gram 解析とクラスター分析を用いたグループ化において, 意味を持ったグループ化を行うことができた. また, 評価実験 1 では特に, 不正なレポートを検出することもできた. 評価実験 2 では, n-gram 解析を用いたグループ化の性質である, レポートの特徴によるグループ化を顕著に示した結果となった. 今後は, 形態素解析を用いたグループ化や形態素解析と n-gram 解析を組み合わせたグループ化などとも比較検討していきたいと考えている.

参考文献

[1] 先進学習基盤協議会(ALIC) : “e ラーニング白書”

[2] 小高知宏, 村田哲也, 高建斌, 諏訪いずみ, 白井治彦, 高橋勇, 黒岩丈介, 小倉久和 : “n-gram を用いた学生レポート評価手法の提案”, 電子情報通信学会論文誌, Vol.J86-D-I, No.9, pp.702-705, 2003 年 9 月

[3] 小倉久和, 小高知宏 : “人工知能システムの構成基礎からエージェントまで”

[4] 北研二, 津田和彦, 獅子堀正幹 : “情報検索アルゴリズム”

[5] Brill's Tagger:http://www.cs.jhu.edu/~brill/

[6] 茶筅 : http://chasen.naist.jp/hiki/ChaSen/

[7] wvware : http://2php.jp/install/wvware.html