データマイニングによる
文書校正支援技術の開発

情報指導部　○上田芳弘　加藤直孝　林克明
金沢大学　木村春彦

１．目　的

データマイニングは，データベースなどに蓄積した大量データの中から有益な知識を発掘，発見するための技術である。本研究では，組織内で行われる文書校正に関するデータを蓄積し，データマイニングを用いて，文書校正に役立つルールを抽出する技術の開発を目的とする。以下，校正とは誤字や脱字の修正とともに文法や意味などの推敲を含むものとする。組織内では一般に，文書は情報周知をはじめ，知識の共有や組織の意思表現のために作成され，文書作成者とは別の校正者が様々な側面から校正を行っている。従って，校正結果は校正者の長年の業務経験が反映された重要な情報と考えられる。しかし，この校正作業は一般に紙面上で行われるため，校正結果の情報は組織内で蓄積，共有されていない。そこで，校正前後の文書ファイルの差分から校正データを生成し，このデータにデータマイニングの技術を応用して校正ルールを抽出する。この抽出時に適用度と呼ぶ新たな指標を導入し，ルールの抽出後に適用度を学習させて文書の種類ごとに異なる校正を可能とする。これにより，校正結果の情報を組織内で知識として蓄積，共有することを目指す。

２．内　容

2.1 データマイニングの概要

コンピュータ環境の急速な発展に伴い，組織内には大量のデータが蓄積されるようになった。データの大規模化は，人間の処理能力をはるかに越え，そこに内在する有益な知識の発見をこれまで以上に困難なものとしている。そこで，様々な業種でデータマイニングの技術への期待が高まっている。

例えば，小売業・流通業ではPOSの普及によってバスケット分析と呼ばれる商品の販売分析が行われている。すなわち，特定の商品の組み合わせが同時に売れることや，顧客の年齢・性別あるいは天候などによるトレンドの分析などが行われ，これまで人間が気づかなかった知識が販売戦略や商品陳列のレイアウトなどに直結している。さらに，リレーションシップマネージメントと呼ばれる顧客一人一人に対するきめ細かな分析が行われ，いわゆる顧客の囲い込みを目指している。

一方，製造業では修理伝票のデータベース化により，これまでよりも効率的な故障部品の特定や故障の発生予測などの分析が行われている。また，上述の小売業・流通業と同様な市場分析のほか，サービスセンターやコールセンターに寄せられる顧客からの質問やクレームの分析を行い，新製品開発や製品改良につなげている。

更に，医療分野や通信業においてもデータマイニングの応用が進められている．これらデータマイニングの応用の中で最も広く利用されている技術が相関ルールである。相関ルールとは，アイテムの集合(商品の集合)とデータベース(商品を購入した顧客のデータベース)の中から，X⇒Y（ある顧客が商品Xを購入するならば商品Yも同時に購入する)という

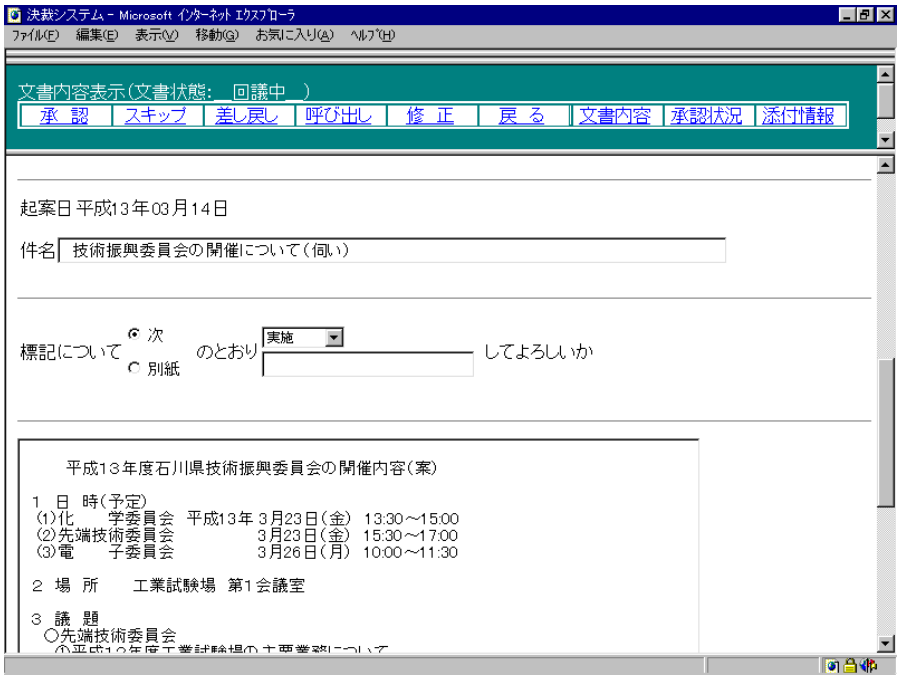

図１　文書決裁支援システムの承認画面

ルールを抽出するための技術である。一般に，この抽出のためのアルゴリズムは，ルールがデータベース全体で成立する率などの簡単な指標を用いることで大量データの中から高速に役立つルールを抽出できることを特徴としている。

2.2 校正データの生成

工業試験場では，官公庁における公文書などを対象とした文書決裁支援システムを開発し，運用している。本システムは図１に示すように，業務担当者が作成した文書について，役職順に組織内で承認を受け，最終的には組織長が決裁を行うことで，文書内容を様々な角度から検討し，文書を校正することを支援するシステムである。本システムを運用する中で，文書作成者や承認者が行った文書校正に関するデータ(以下，校正データと呼ぶ)を蓄積できるようになった。すなわち，これまで文書を印刷した紙面上で手書きにより校正内容を記していたのに対して，本システムでは校正による文書ファイルのバージョン管理を行っているので，どの部分がどのように校正されたかを解析することができる。また，本システムの使用を前提とせずに紙面上で校正作業を行う場合も，一般に作成者は文書をワープロソフトなどで作成しているので，文書ファイルのバージョン管理を行えば，校正データを蓄積することは可能である。

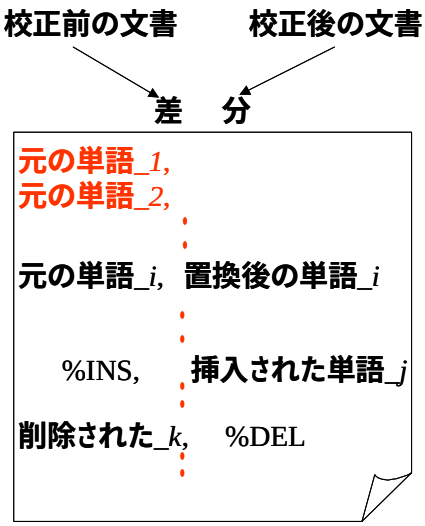

図２　校正データの生成

ここで，校正データは，図２に示すように蓄積した校正前後の文書ファイルの差分から生成する。ただし，この差分は，文書に対して形態素解析を行い，すべての文を単語に分割した後に取るものとする。この校正データは１行で単語が１つしか存在しないときは，その単語は校正されなかったことを表し，それ以外では

１行の左の単語が校正前，右の単語が校正後を表している。また，%INSは単語の挿入，%DELは単語の削除を表している。

2.3 適用度の導入と学習

上述の校正データから相関ルールの手法によって校正ルールを抽出する。このとき，文書の評価基準は人間の主観に多くを依存するため，対象業務や分野など(以下，ドメインと呼ぶ)によって好ましさが異なるものと考えられる。例えば，用語や用字などはドメインごとに暗黙的に統一されていることが多い。よって，ドメインごとに校正の視点が異なり，校正ルールも異なるものと考えられる。そこで，適用度を式(1)のように定義し，ルールはドメインごとに異なる適用度を持つものとする。

$$\text{適用度} = \frac{\text{実際に校正が行われた事例数}}{\text{ルールの条件部を満足する事例数}} \qquad (1)$$

ここで，適用度は，校正を完了した文書でドメインごとに学習するものとし，事例数とは学習に用いた文書内に出現する文の数を表す。あるルールが1に近い適用度を持つならば，対象とするドメインでは，このルールは確実に使用され，逆に0に近い値を持つならば，このルールは使用されない。この適用度の導入により，ドメインごとに異なった校正が可能になると考えられる。

2.4 抽出した校正ルールの事例

既存の文書データの中から校正があった文書を対象に，無作為に430ファイルを選択し，校正ルール抽出用の文書データとした。提案手法を用いて抽出した校正ルールの事例を図３に示す。この図のようにルールは，条件部として最初の{}内に校正される単語，次の{}内に校正される単語の付近に文書内で出現する単語を持つ。更に行動部として{RPL}(置換)，{INS}(挿入)，{DEL}(削除)のいずれかを取り，最後の{}内に置換または挿入される単語を持つ。

また，抽出したルールは(1)表記の統一，(2)文法や意味の不整合，(3)組織情報といった3種類の校正を表すルールに大別できた。(1)の表記の統一は，文法や意味的な誤りではないが，文書中で統一すべき用語，用字を指摘するルールで，例えば漢字とひらがなの表記を統一するルールが抽出できた。これらのルールの多くは，逆向きの置換を行うルールも抽出されていて，ドメインにより用語や用字が異なることが確認できた。(2)の文法や意味の不整合は，文法の誤りや意味の重複などを指摘するルールで，例えば<rule$_4$>は「・・・してる」という口語体の文章を文法的に正しい「・・・している」に校正することを意味している。また，<rule$_5$>と<rule$_6$>は，意味の重複を削除するルールである。(3)の組織情

(1) 表記の統一

＜rule$_1$＞ {等}{}⇒{RPL}{など}

＜rule$_2$＞ {出来る}{}⇒{RPL}{できる}

＜rule$_3$＞ {さらに}{}⇒{RPL}{更に}

(2) 文法や意味の不整合

＜rule$_4$＞ {る}{て} ⇒{RPL}{いる}

＜rule$_5$＞ {最も}{最適解} ⇒{DEL}{}

＜rule$_6$＞ {など}{例えば} ⇒{DEL}{}

(3) 組織情報

＜rule$_8$＞ 地域活性化事業⇒{RPL}専門家派遣事業
{地域}{活性化,事業}⇒{RPL}{専門家}
{活性化}{地域,事業}⇒{RPL}{派遣}

＜rule$_9$＞ 日本小型自動車振興会⇒{RPL}日本自転車振興会
{小型}{日本,自動車,振興会}⇒{RPL}{自転車}
{自動車}{日本,小型,振興会}⇒{DEL}{}

＜rule$_{10}$＞
NEDO ⇒{RPL} 新エネルギー・産業技術総合開発機構(NEDO)
{NEDO}{}⇒{RPL}{新}
{}{NEDO} ⇒{INS}{エネルギー}
{}{NEDO} ⇒{INS}{・}
{}{NEDO} ⇒{INS}{産業}
{}{NEDO} ⇒{INS}{技術}
⋮

図３　校正ルールの事例

報は，組織内における固有の情報に基づいた校正を表すルールで，例えば<rule$_8$>と<rule$_9$>は，工業試験場が補助金を申請する事業や申請先の変更にともなう指示を意味している。また<rule$_{10}$>は共同研究機関名を略称だけではなく正式名と略称で表すルールである。

以上，抽出したルールによってユーザは文の表記や文法，意味に関する指示を得られるだけではなく，組織情報に基づいた指示を得られるものと考えられる。

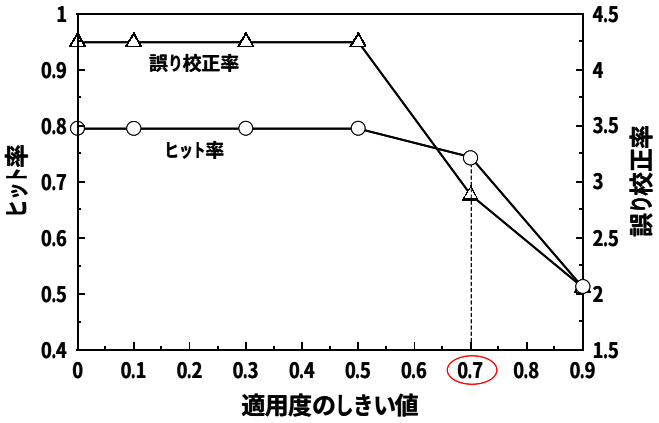

図４　校正者とルールによる校正の比較

2.5 校正者との比較実験

ドメインをコンピュータ関連に限定して，校正を完了した330ファイルを無作為に選択し，ルールの適用度の学習を行った。更に，これらとは別の文書で，コンピュータ関連の校正されていない50ファイルを用いて，実際の校正者が行った校正結果とルールによる校正結果を比較した。その結果，図４に示したように，まず校正者が行った校正のうち，ルールで校正できた割合を表すヒット率では，最大で80%がルールにより校正可能であることが分かった。また，適用度のしきい値を変化させてヒット率を求めた。すなわち，しきい値を超える適用度を持つルールを対象のドメインで有効とした。その結果，同図のように適用度のしきい値を0.5より大きく設定すると，低い適用度のルールは適用されないため，適用できるルール数は減少し，ヒット率は低下することが分かった。

また，ルールによる校正のうち，校正者が行わなかった校正，すなわち，ルールによる誤り校正率を評価した。その結果，同図のように適用度のしきい値が0.5以下では，対象のドメインで使用されないルールが適用されるため，誤り校正率は4.2という高い値を示した。しかし，しきい値が大きくなるに従って，この率は減少する。以上のように適用度のしきい値は0.5以下の値に設定した方が，実際の校正者が行う校正とのヒット率が高い校正が行えるが，その反面，誤り校正が多くなる。本評価実験ではこの両側面を考慮して，適用度のしきい値は0.7とすることが妥当であるといえる。

以上，データマイニングにより実際の校正者が行う校正にほぼ適合した校正ルールを抽出でき，校正結果から抽出した知識を組織内で共有化することの有効性を示すことができた。

３．結　果

データマイニングの代表的な手法である相関ルールを用いて，文書校正に役立つルールを抽出する手法を提案した。提案手法の有効性を評価するために実験を行い、以下のことが分かった。

(1)提案手法により，文の表記や文法，意味レベルの校正とともに，組織固有の情報に基づいた校正を表すルールが抽出できた。

(2)校正者による校正結果と抽出したルールによる校正結果を比較したところ，校正者による校正の 80%がルールにより校正できた。

(3)ルールによる誤り校正率は，ルール適用時に適用度のしきい値を設け，適切な値に設定することにより減少可能であった。