2005 年度卒業研究概要

情報検索技術を用いたアンケートデータの分析手法に関する研究

指導教授
承認印

後藤 正幸 研究室
0232250 渡辺 智幸

1. 研究の背景と目的

近年の IT 技術の発展に伴う情報の電子化により、膨大な量のデータを電子データの形で収集することが容易になった。これらの情報を有効に活用するには、必要な情報を効率よく見つけ出すための手段が必要である。この要求を満たすものが情報検索技術であり、教育分野においてもこの技術が応用され始めている。

大学を始めとする教育機関において、授業改善等を目的とした学生アンケート調査は従来から行われてきた。アンケートの設問方式には、選択式と自由記述式の二通りが考えられる。選択式回答の分析は様々な統計的手法を用いることができ、集計は比較的容易である。一方、自由記述式回答については、自由記述のテキストを自動的に分析する方法が確立しておらず、人手による分類や意見の列挙などの手間がかかるのが実情である。情報検索技術を用いてこれを効率的に処理できれば、アンケート分析において有用な手段となろう。

本研究は自由記述式回答の分析に焦点を当てる。従来から研究されてきた自由記述の分析手法として重要文抽出がある。大量の文書中から重要な部分だけを自動的に抜き出し、要約文書を作成する手法である。その際、文書間の類似度情報を用いるが、類似度の大きさに基づいて抽出すると内容の類似した文ばかりが抽出され、網羅性が低下する[1]。通常、この手法は新聞記事等の要約に用いられるが、この場合は抽出結果において特に問題はない。新聞記事は文書全体を通して起承転結の流れがあるので、類似した表現が少なく、網羅性の低下による影響を受けにくいからである。一方、自由記述アンケートデータは新聞記事と異なり、類似意見が多数存在する可能性が高い。よって、同手法を適用すると、網羅性低下の影響を強く受けてしまい、回答者全体の傾向を把握するのに充分な結果を得られない。本研究は、従来の重要文抽出法に類似度閾値 γ を新たに導入することで、自由記述アンケートデータの分析に適応した網羅性の高い重要文抽出手法の提案を目的とする。さらに、提案手法を計算機上に実装し、実際の自由記述アンケートデータに適用して、その評価を行う。

2. 研究内容

2.1 従来手法 ―類似度を用いた重要文抽出アルゴリズム―

亀田の手法[2]における、文書間の類似度を用いて重要文を抽出する基本的な考え方を示す。

[定義1] (**単語ベクトル**) 文書 x における各単語の出現回数を要素とするベクトルを単語ベクトル v_x とする。

[定義2] (**類似度**) 文書 x, y の類似度 $sim(x, y)$ を単語ベクトル v_x, v_y の余弦で定義する。

$$sim(x, y) = \frac{v_x \cdot v_y{}^T}{\sqrt{(v_x \cdot v_x{}^T)(v_y \cdot v_y{}^T)}} \quad (1)$$

多くの文書と類似度の大きい文書は、文書集合全体の多くの内容を含んでいると考えられる。よって、類似度の平均により文書の重要度を与える。

[定義3] (**重要度**) 文書 x の重要度 $imp(x)$ を次式で定義する。

$$imp(x) = \frac{1}{m-1} \sum_{y \neq x} sim(x, y) \quad (2)$$

ただし、m は文書の総数である。これらを用いて以下のアルゴリズムで文書を順序付けする。

[類似度を用いた重要文抽出アルゴリズム]

Step1 要約したい文書集合の形態素解析を行って単語を抽出し、各文書を単語ベクトルで表す。
Step2 文書集合中の各文書間の類似度 $sim(x, y)$ を式(1)より計算する。
Step3 各文書の重要度 $imp(x)$ を式(2)より計算する。
Step4 文書の重要度 $imp(x)$ の大きい順に文書を抽出する。

重要度の大きい文書との類似度が大きい文書は、これも重要度が大きくなる傾向があると考えられる。従って、この手法では内容の似た文書ばかりが重要文として抽出される傾向がある。

2.2 提案手法

従来手法では抽出される文書の内容が偏ってしまう。しかし自由記述意見の分析においてこれは望ましくなく、意見を網羅的に抽出したい。そこで、本研究では網羅性を向上する重要文抽出手法を提案する。類似度閾値 γ を用いて、重要文が 1 つ抽出されるたびに、その文書との類似度が γ 以上の文書を除外し、残った文書の中から次の重要文を抽出する。これにより、内容の似た文書が抽出されにくくなる。除外された文書は類似文書として扱い、その数を表示する。(γ の値はユーザにより与える。ただし、$0 < \gamma \leqq 1$ とする。)

【提案アルゴリズム】

Step1 要約したい文書集合の形態素解析を行って単語を抽出し、各文書を単語ベクトルで表す。

Step2 文書集合中の各文書間の類似度 $sim(x, y)$ を式(1)より計算する。

Step3 未抽出文について、文書の重要度 $imp(x)$ を式(2)より計算する。

Step4 未抽出文のうち重要度の最も大きい1文書を抽出する。

Step5 Step4 で抽出した文書との類似度が γ 以上の文書を除外し、それを類似文書とする。類似文書数を数えて抽出文書の付随情報として付加する。

Step6 残り全ての文書が除外された場合は終了。

Step7 Step4 に戻り、残った文書の中から再び重要度の高いものを抽出する。

なお、従来手法は提案アルゴリズムにおいて $\gamma = 1$ と設定することと等価である。

3. 評価実験内容

ある大学（以下 A 大学とする）で実施された学生アンケート調査について実際に分析を行うことで、提案手法の有効性を検証する。このアンケートは A 大学ポータルサイト上で全学生を対象に実施されたものであり、選択式項目と自由記述式項目で構成されている。

3.1 アンケートデータ概要

評価実験に使用するアンケートデータの内容は以下の通りである。

【実証実験後の Web 科目登録システムの性能と利用環境に関するアンケート】
設問 3：Web 科目登録システムを利用した感想をお聞かせください。（自由記述）
調査期間：2004 年 8 月 2 日（実証実験日）　有効回答数：605 名

3.2 実験方法

提案アルゴリズムに基づいて文書の順位付けを行い、重要文を抽出する。その際、類似度閾値 γ を 1, 0.0001 の 2 種類に設定する。$\gamma = 1$ では従来手法と同様に全文書（605 件）が抽出され、$\gamma = 0.0001$ では抽出文書数が 32 件に絞られる。これら 2 つのパターンにおける重要文抽出結果の上位 30 件を比較することで、提案手法の網羅性の向上を検証する。形態素解析は「茶筌」[3]で行い、明らかな間違い箇所は人手で修正した。

4. 評価実験結果

4.1 重要文抽出結果（一部抜粋）

1 簡単にできて、良かった。　類似意見数0件
2 特に重くなく、快適に操作できました。　類似意見数0件
3 はやく申請できたが、本当に申請できたかどうか不安になる。　類似意見数0件
4 簡単でよかった。　類似意見数0件
5 今回はとてもスムーズにできたと思います。　類似意見数0件
6 今回は早かった。　類似意見数0件
7 わざわざ大学まで行かなくても科目登録でき，しかも夜などでも申請できるので，大変よい。　類似意見数0件
8 全然つながらない。　類似意見数0件
9 自宅のPCから簡単に利用できるのはうれしい。　類似意見数0件
10 わざわざ大学の混雑に巻き込まれなくてすむのでよい。　類似意見数0件
11 スムーズに処理が行え、登録の間違いもなかったのでぜひ来年度からこのシステムを利用したいです。　類似意見数0件
12 操作が簡単で、わかりやすかったと思います。　類似意見数0件
13 科目の検索がイマイチやりにくい。学科、学年で検索できるようにしてほしい。　類似意見数0件
14 何度か制限画面が表示されたものの、短時間で終わるし、家で科目登録できるのでとても便利だと思う。　類似意見数0件
15 説明を見ながら作業したので全体の時間は少し多めにかかったが、検索、取り消し、申請などのページ
移動は思ったよりもスムーズで快適に利用することが出来た。　類似意見数0件
16 前回より格段に快適になった。おそらく応答時間が早かったせいだと思う。　類似意見数0件
17 時間も早く満足のいく結果だった。　類似意見数0件
18 大学まで登録に行くのが面倒だし、パソコンの方が選択ミスも気付きやすいと思うから　類似意見数0件
19 前回より改善されていたし、実用化は可能だと思う。　類似意見数0件
20 今回は非常にスムーズでした。　類似意見数0件
21 前回よりずっと使用時の躊躇は軽かったです。ただ、科目選択で５０件を越えたからといってまったく
表示してくれないのでは使い勝手がいいとは言えません。　類似意見数0件
22 楽チン　類似意見数0件
23 非常に便利なので、ぜひ採用して欲しい。　類似意見数0件
24 これまでの実験よりも、スムーズに行うことができた。　類似意見数0件
25 前回とは異なり利用者数が表示されていたので混み具合がわかり、どの位の時間がかかるかという目安
になったので非常に良かったです。ただ今回は予想していたよりもかなり参加数が少なかったので早く
終わりました。　類似意見数0件
26 申請方法を載せたページを見ながら行ったのですが、そのページの説明がわかりにくかったので少々
時間がかかった。もっとわかりやすく作るとなお良いと思います。類似意見数0件
27 12:30ちょうどにアクセスし始めたのでスムーズに登録できました。実際の科目登録でもこれくらいの
軽さで出来るなら、ぜひWeb登録をしたいです。　類似意見数0件
28 思った以上にスムーズに科目登録が行えたので正直びっくりしている。けど、この調子なら来年の
科目登録はうまくいくのではと思った。　類似意見数0件
29 登録申請をしてエラーになった科目がそのまま登録されていました。　類似意見数0件
30 以前の実験にも参加したが、今回のはとてもスムーズだと感じた。　類似意見数0件

図 1. $\gamma = 1$ の抽出結果（従来手法）

1 簡単にできて、良かった。　類似意見数419件
2 時間も早く満足のいく結果だった。　類似意見数25件
3 大学まで登録に行くのが面倒だし、パソコンの方が選択ミスも気付きやすいと思うから　類似意見数38件
4 これまでの実験よりも、スムーズに行うことができた。　類似意見数4件
5 12:30ちょうどにアクセスし始めたのでスムーズに登録できました。実際の科目登録
でもこれくらいの軽さで出来るなら、ぜひWeb登録をしたいです。　類似意見数21件
6 前回、前々回に比べて一連の作業がとてもスムーズであったと思います。　類似意見数14件
7 特にトラブルなく、スムーズに登録できたのでよかったと思う。　類似意見数0件
8 マークシートより格段に楽。　類似意見数2件
9 ３月に行った科目登録と同じようにスムーズに行えたように思います。（理工学部）　類似意見数8件
10 大変良いシステムだと思います。１年年からこの方法で登録できたら本当に夢のようです。
便利で楽しくて時間で面倒な手間がなくて文句なしです。　類似意見数3件
11 各学部ごと、曜日ごとなどに分類されていて、科目検索が非常に分かりやすかったです。　類似意見数8件
12 大学まで足を運ばずにすむから。　類似意見数2件
13 レスポンスが早かったので、特に問題は感じなかった。　類似意見数1件
14 今回は正常に作動したが、今までの経緯から使用したいとは思わない。　類似意見数7件
15 実用化されればすごく便利だと思う。　類似意見数0件
16 保健体育科目の場所が多少わかりにくかったです。　類似意見数4件
17 おもってたよりスムーズにできた。登録も難なくできたが、実験実証の手順の説明がわかり
にくくて時間がかかった。PDFをひとつにまとめてほしい。　類似意見数4件
18 時間と交通費の節約になる。24時間申請可能なシステムならば時間がとても有効に使えてる
ようになるから。　類似意見数0件
19 動作はいたって快適だった。ただ、検索が可能となる条件が厳しく（特に件数）、細かい指定
をしないと検索結果が得られない点が、時には不便となりうる気がした。そのあたりは負荷が
大きくなり過ぎない程度に改善して欲しい。あと、同じ時限に設置されている科目を複数個
選択しても問題なかった点が、誤解を招きやすいと思った。　類似意見数0件
20 以前よりはるかに早い。　類似意見数0件
21 他学部の科目がわかりやすい　類似意見数2件
22 今回は快適に申請が行えた。　類似意見数0件
23 シナリオの後、時間が余ったのでいろいろやってみた。実験だからなのかもしれないが、時間が
重複する科目が何個でも申請できた。また、既習の科目も申請できた。例えば、商学部英語選択A
は４６個申請できたし、基礎数学も申請できてしまった。商法Aは２個申請できた。結局、２３６
単位分申請できた。やろうと思えば１０００単位でも申請できる。　ロシア語Ⅱで検索すると、
フランス語Ⅱが検索結果として出た。フランス語Ⅱで検索すると、フランス語Ⅱが検索された。
24 科目の検索等もスムーズで非常に使い易かった。類似意見数3件
25 自宅で科目登録できるので便利だと思う。待ち時間がほとんどなかったので来年度から十分に使えると思う。　類似意見数1件

図 2. $\gamma = 0.0001$ の抽出結果（提案手法）

4.2 考察

$\gamma = 1$（従来手法）と $\gamma = 0.0001$（提案手法）の結果を比較すると、重要度上位 30 件において違いが見られた。例えば、「スムーズ」という単語が含まれるのはそれぞれ 8 件と同数であった。その 8 件を見ると、従来手法では図 1 の 5, 20, 24, 30 からわかるように、表現が少し違うだけで趣旨が同じ意見が複数抽出されており、アンケート集計において有益な情報とは言えない。一方、提案手法では同じ「スムーズ」という単語を使っていても全体としての趣旨は異なる意見が抽出されているため、従来手法よりも得られる情報量が多くなる。さらに、30 件全体を見渡しても、特に類似した意見が複数抽出されていることはなく、網羅性は向上していると言える。

また、提案手法では長い意見が多く抽出される傾向が見られた。長い意見は含まれる単語の種類の多さから他の意見と類似しにくく、提案アルゴリズム中の除外対象になりにくいことが原因と考えられる。

5. 結論と今後の課題

文書間の類似度から重要文を抽出する際、類似度閾値 γ を基準に抽出対象を選り分けることで網羅性を高めることができた。同時に、γ によって除外された文書を類似意見として数え上げることで各意見にどの程度の人数が分布しているかがわかる。大量の自由記述文書を少量に絞り込んで読みやすくし、なおかつ全体の傾向も把握できるこの手法は、自由記述アンケートデータを効率的に処理する上で有効である。

今後の課題として、分析対象とする文書集合の性質による抽出結果の変化や、γ の与え方による抽出結果の変化をより深く調べていくことが求められる。

参考文献

[1]伊藤潤、石田崇、後藤正幸、平澤茂一“文間の単語共起類似度を用いた重要文抽出法”2002
[2]亀田雅之 ”擬似キーワードによる重要キーワードと重要文の抽出”1996
[3]日本語形態素解析システム「茶筌」, http://chasen.aist-nara.ac.jp/hiki/ChaSen/