

情報検索技術を用いた効率的な授業アンケートの分析

Efficient Analysis of Student Questionnaires
using Information Retrieval Techniques酒井 哲也[†] 伊藤 潤* 後藤 正幸[‡] 石田 崇* 平澤 茂一*[†](株)東芝 研究開発センター * 早稲田大学大学院 理工学研究科[‡]武蔵工業大学 環境情報学部 * 早稲田大学 理工学部

要旨

大学における授業の改善を最終目標として学生アンケートの自由記述を自動分析したのでその分析手法および分析結果について報告する。情報検索における擬似適合フィードバック技術の適用が、学生の考えを効率よく把握するのに有用であることを示す。

Abstract

This paper analyzes free texts from student questionnaires. Our ultimate goal is automatic knowledge acquisition for improving university lectures. We will show that application of Pseudo-Relevance Feedback is useful for efficiently grasping students' thoughts.

1 はじめに

大学教育は常に変革を求められている。通信基盤の整備により遠隔教育が可能になり、ワープロ・メールソフトの普及により教員が配布する教材や学生が提出する答案が電子化テキストとして蓄積されることは今や常識である。蓄積されたテキストデータから有用な知識を発掘する技術としてテキストマイニングが注目されており、この分野では、テキストの分類やクラスタリング、テキスト中の主要な概念の抽出、概念間の関係の抽出、およびこれらの視覚化などの研究 [3, 10] や製品化 [4, 11, 12] が行われている。また、自動要約の分野では複数テキスト要約の研究が行われており、これも大量テキストからの知識発見技術として将来的に役立つことが期待される [1]。

本研究では、大学の学部生を対象とした半年にわたる講義「コンピュータ工学」の初回および最終回に実施した学生アンケートの自由記述を自動分析したのでその分析手法および分析結果について報告する。我々は、個々の学生の自由記述と成績とを対応づけ、これらをトップダウンおよびボトムアップにグループ化し、情報検索における検索条件拡張技術である擬似適合フィードバックを用いて代表キーワード・代表センテンスを抽出した。このような比較的単純な方法により効率的な傾向分析が行えることを示す。我々の最終目標は大学教育における授業改善のための知識獲得であり、本研究はその第一歩である。

教育分野において情報検索技術を適用した先行研究としては例えば長坂・阿手による Interactive Education に関する取り組みがある [5]。彼らのシステムでは、記

述問題に対する学生の回答と模範回答との類似度を情報検索システムにより求めることにより成績付与支援を実現している。一方、本研究は学生の間・期末試験の成績をひとつの手掛かりとし、学生アンケートの自由記述から学生の要望や考えを抽出することを目的としている。各学生グループの代表センテンス抽出は、むしろ複数文書要約の立場に近い。

2 データの概要

早稲田大学理工学部経営システム工学科の学生を対象とした講義「コンピュータ工学」の初回および最終回において、選択式および自由記述式設問を含む記名式アンケートを実施した。学生には、アンケート用紙の Excel ファイルを Web からダウンロードし、記入を行ったファイルをメールに添付して提出するよう指示した。自由記述の設問は以下のとおりである：

初回 これから勉強するコンピュータについて、現在考えていることを自由に書いてください。

最終回 半年間勉強してきたコンピュータについて、最初の印象と変わりましたか。現在考えていることを自由に書いてください。

自由記述の長さは、0.7 ページ以上 1 ページ以下 (ただし 1 ページは 40 文字 × 36 行) に制限した。これにより、文書長が比較的均一なテキスト集合が得られた。初回・最終回アンケートともに期限内に提出し、かつ中間・期末試験を受験した学生は 132 人であり、これらを分析対象とした。各学生には、中間試験の点数 (m_{score}) および期末試験の点数 (f_{score}) をもとに以下の式で算出される「成績基準点」(c_{score}) が付与された。

$$c_{score} = \max\{f_{score}, 1/3 * m_{score} + 2/3 * f_{score}\} \quad (1)$$

アンケートは記名式であるため、各学生の自由記述と成績基準点とを対応づけて分析を行うことができた。なお、最終的な成績 (A, B, C, D, F) は、成績基準点に課題レポートの採点結果と出席状況を加味して決定される。

一方、選択式質問としてはコンピュータに関する興味・知識・経験の有無、また「ハードウェア」「ソフトウェア」などの特定分野に対する興味などを問うものを用意した。これらを利用した分析結果については別途報告予定である [2]。

3 分析方法

初回および最終回アンケートのデータを別個のデータセットとして扱い、各データセットに含まれる学生132人分の<自由記述,成績基準点>の対を以下の2通りの方法でグループ化した。

トップダウン分析 成績基準点が80点超,70点超80点以下,60点超70点以下,50点超60点以下,50点以下の5つのカテゴリに分類。

ボトムアップ分析 情報検索システムBRIDGE [9]を用いて各学生の自由記述をターム(索引語)のビットベクトルに変換し,上記ベクトル群をもとに群平均法によりクラスタツリーを作成,ビットベクトル間の類似度係数としてはJaccard係数を用いた [7],クラスタ数は30に設定。

なお,ここでいうタームとは形態素解析結果のうち付属語などを除いたものである。初回・最終回アンケートにおけるタームの異なり(ベクトルの次元)はそれぞれ約2,300および約2,000であった。

各学生グループの代表キーワード/センテンス抽出には,情報検索における検索条件拡張技術である擬似適合フィードバック(Pseudo-Relevance Feedback: PRF) [8]を応用した。以下に,本研究で採用した確率検索モデル [6]に基づくPRFについて説明する。

確率検索モデル型のPRFでは,まずユーザの入力した検索要求を用いて第1回目の検索を行い,初期検索結果(ラングづけされた文書のリスト)を得る。次に,上位 R 件の文書を「適合文書」と仮定し,各適合文書から新たなタームを以下で定義される選出基準(ow)の高い順に T 個選出する:

$$rw_t = \log \frac{(r_t + 0.5)(N - n_t - R + r_t + 0.5)}{(n_t - r_t + 0.5)(R - r_t + 0.5)} \quad (2)$$

$$ow_t = r_t * rw_t \quad (3)$$

ここで, r_t は上記 R 件のうちターム t を含む文書数, N は文書の総数, n_t はこの N 件のうちターム t を含む文書の総数である。 n_t は文書頻度(document frequency)とも呼ばれる。上式により,「適合文書」に出現する確率が高く,かつ「非適合文書」に出現する確率が低いタームが選出される。情報検索では,このようにして得られた T 個の新たなタームを検索条件に付加して再検索を行うことにより,検索有効性が向上する場合が多いことが知られている [8]。

上記の手法を特定のテキスト集合の代表キーワード抽出に応用することは容易である。すなわち,対象となるテキスト集合のサイズを R とし, N 件の全テキストとの比較により T 個の代表キーワードを抽出すればよい。また,テキスト中の各センテンス(s)に対して以下のスコア($SCORE(s)$)を算出することにより代表センテンス抽出を容易に行うことができる。

$$SCORE(s) = \sum_{t \in s \cap e} rw_t \quad (4)$$

ここで e は代表キーワードの集合($|e| = T$)である。上記の式は rw_t の単純な和であるため長い文に高いスコアを与えるが,本研究のように各テキスト集合の典型的な文を読むという用途にはこれで十分であると考えられる。長い文の選出が好ましくない場合はセンテンス長による正規化を施せばよい。

4 結果および考察

4.1 トップダウン分析

今回は,成績基準点に応じてトップダウンに設けた5カテゴリのうち80点超のカテゴリと50点以下のカテゴリに着目し,両カテゴリの自由記述の比較を初回および最終回アンケートについて行った。代表キーワード/センテンスの一部を表1および2に示す。

表1から,初回アンケートについて以下のような傾向が読み取れる。

80点超:「コンピュータの基本を理解する」という姿勢で授業に臨んでいる学生が多い。

50点以下:コンピュータを「インターネットやパソコンソフトを使用するための道具」としてしかとらえていない。代表キーワードとして出現している「使い」「使用」「機能」「便利」というターム以外にも,代表センテンス中には「活用」「利用」という類義のタームが出現している。

成績の良い学生と悪い学生とで初回アンケートの記述内容に傾向の違いが見られることから,上記のような分析結果を学習データとして用い,来年度の学生に対して初回アンケートを実施した時点で各自の成績を予測し,これに応じて授業内容や進め方を変更したり,将来2つのクラスに分割するための検討資料とするといった応用が考えられる。また,本研究では例えば形態素解析時に半角の「Excl」と全角・大文字の「EXCEL」を同一タームと見なすなどの異表記吸収処理を行っているのみで,「概念レベル」の分析を行っているわけではないが,表1「50点以下」のカテゴリには「EXCEL」と「エクセル」「WORD」と「ワード」という異表記吸収処理の範疇外の同義語が他のタームとの共起を通じて得られている。すなわち,このような分析において,「概念レベル」の処理は必要不可欠ではないと考えられる。

表2からは,成績の良い学生でも「半年ですべてを理解することは無理」と感じていることが読み取れるし,成績の悪い学生の代表センテンスからは「友人に教わる」という姿勢が伺える。また,表2「50点以下」のカテゴリには「情報処理技術者試験」と「国家公務員試験」の両方に興味をもっている学生と,後者のみに興味をもっている学生がいることがわかる。テキストマイニングの分野では,トピックと述語の対(例えば「情報処理技術者試験」が{好き,嫌い})に基づく分析を行うアプローチがあるが,タームの集合に基づく代表センテンス抽出という我々の単純なアプローチでも「同一トピックに関する異なる記述」を収集することはある程度可能であることがわかる。

4.2 ボトムアップ分析

図1および2は初回および最終回アンケートをそれぞれ階層的にクラスタリングした結果と成績基準点との対応を表している。各点は個々の学生に対応する。グラフの横軸は,1学生1クラスであった学生が初めて他のクラスに吸収された時点での両クラス間のJaccard係数であり,クラスタリングは横軸の右から左へ進行する。すなわち,左にプロットされている学生ほどユニークな語彙を用いていたことになる。

図1における3つの異常値は、他人のアンケートの丸写しを行った学生を表している。最初の数文のみ差し替えてあったため、人手による調査ではこのことを見落とし可能性がある。両グラフにおいて、試験の得点により算出される成績基準点とアンケート自由記述のユニーク性ととの間に相関はないが、自由記述自体に対して主観評価を行う際に、上記ユニーク性をひとつの参考値とすることは可能である。実際、教員が記述問題を評価する際にユニーク性を重視することが比較的多いという調査結果もある [5]。

表3および4に初回および最終回アンケートのクラスタリング結果(50クラスタ)に対して代表キーワード/センテンス抽出を行った結果の一部を示す。クラスタは大きさの降順になっている。すなわち、第1クラスタからは多数派の話題・意見が得られ、最後尾に近いクラスタからはユニークな話題・意見が得られる。群平均法はクラスタサイズが均等になりにくいという性質をもつが [7]、これはユニークな文を抽出するタスクにおいては必ずしも欠点ではない。

表3の第1クラスタから、「コンピュータ」よりも「経営・管理」に興味があるという学生や両方に興味があると言っている学生が少なくないことが把握できる。一方、表4の第1クラスタから、講義を受けた上で「コンピュータ工学は経営に必要」という学生とこれに必ずしも賛成しない(コンピュータは道具として使えばよい)という学生がいることがわかる。すなわち、多数派意見は初回・最終回ともにコンピュータと経営との関係について述べたものが多い。

一方、ユニークなクラスタの例としては表4の第4および第13クラスタのように、「黒板に書いた文字」について触れたものがある。「文字が見にくい」という意見と「文字が見やすい」という意見の両方が得られており、ここでもまた「同一トピックに関する異なる記述」が自動的に得られている。

5 まとめ

本研究では、情報検索技術を適用した学生アンケートの自由記述の分析手法および分析結果を報告した。トップダウン分析は、成績上位の学生と下位の学生が考えていることを比較して授業改善につなげるという使い方に適している。一方、ボトムアップ分析は、多数派の意見やユニークな意見に焦点をあてることが可能であり、両手法は相補的である。ビットベクトルによるクラスタリングや表層レベルのタームに基づく代表キーワード/センテンスの抽出と言った比較的単純な手法により、効率良く自由記述の分析が行えることがわかった。本手法では、同一トピックに対する異なる記述を対比させて表示することが比較的容易である。今後、これらの手法を発展させ、実際の授業改善や採点の客観性向上につなげていきたい。

参考文献

- [1] Document Understanding Conferences: <http://www-nlpir.nist.gov/projects/duc/>
- [2] 後藤正幸, 酒井哲也, 伊藤潤, 石田崇, 平澤茂一: 選択式・記述式アンケートからの知識発見. 2003 PCカンファレンス [発表予定] (2003).
- [3] Hearst, M.: Untangling Text Data Mining. *ACL '99 Proceedings*, pp. 3-10 (1999).

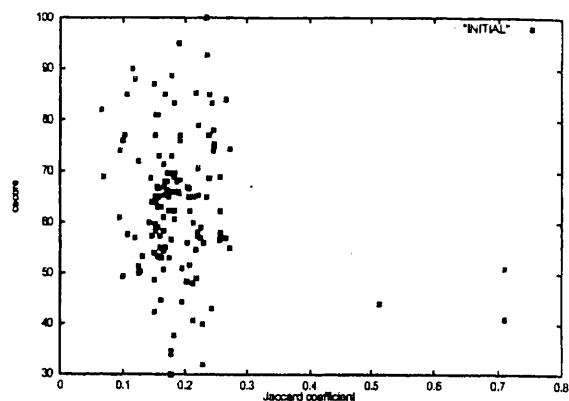


図1: 成績基準点とクラスタリングの関係 (初回アンケート)

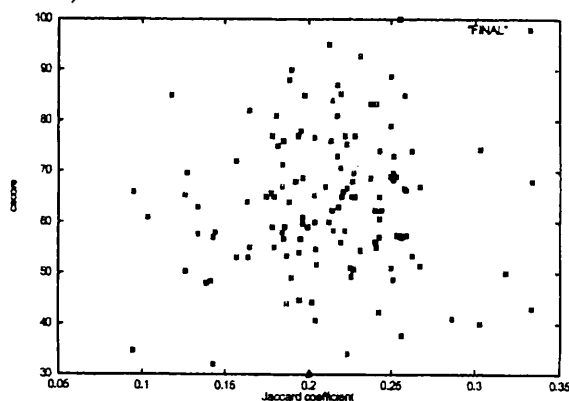


図2: 成績基準点とクラスタリングの関係 (最終回アンケート)

- [4] 東芝 KnowledgeMeister: http://cn.toshiba.co.jp/prod/km2/index_j.htm
- [5] 長坂悦敬, 阿手雅博: 記述問題の自動評価を目指した教育支援情報システムによる Interactive Education. *情報教育方法研究* 第3巻, 第1号, pp. 37-42 (2000).
- [6] Robertson, S. E. and Sparck Jones, K.: Simple, Proven Approaches to Text Retrieval. Computer Laboratory, University of Cambridge (1997).
- [7] Romesburg, H. C. (西田・佐藤共訳): 「実例クラスタ分析」. 内田老鶴圃 (1992).
- [8] 酒井ほか: 確率モデルに基づく日本語情報フィルタリングにおけるフィードバックによる検索条件展開および検索精度評価. *情報処理学会論文誌*, Vol. 40, No. 5, pp. 2429-2438 (1999).
- [9] 酒井哲也, 小山誠, 鈴木優, 真鍋俊彦: 意味役割解析に基づく高適合英語文書の検索. *情報科学技術フォーラム (FIT2002) 情報技術レターズ*, pp. 67-68 (2002).
- [10] 城塚音也: テキストマイニングにおける概念関係視覚化方式の検討. *情報処理学会 情報学基礎研究会 FI-62-18*, pp. 137-144 (2001).
- [11] NEC SurveyAnalyzer: http://www.labs.nec.co.jp/DTmining/products/s_analyzer/
- [12] クオリカ VextSearch: <http://www.qualica.co.jp/develop/vxtsc/index.html>

表 1: 80 点超 vs 50 点以下の比較 (初回アンケート)

80 点超 (代表キーワード: コンピュータ 非常 考え 授業 理解 登場 雑誌 賢真 一見 基本)	
順位	代表センテンス
1	コンピュータ関係の雑誌を読んだりもしたのだが、基本となる知識に乏しいため非常に理解に苦しんでしまった覚えがある。
3	最低限の目標としては、新聞や日経ビジネスなどの雑誌の中で出てくる...コンピュータネットワーク関係のカタカナの用語も、浅くではなく人に聞かれたとき、しっかり説明できるくらいちゃんと理解しておきたいです。
4	しかし自分がコンピュータに出会ってからの10年間を思い出してみると、コンピュータの発達の実質的な速さにはただただ驚かされる。
7	そこで、私自身もコンピュータの基本的なことを理解すべく、この「コンピュータ工学」の授業でいろいろ学びたいと思っている。
50 点以下 (代表キーワード: インターネット 社会 使い 経営 使用 機能 本格 一人 システム 便利)	
順位	代表センテンス
1	使い始め当初はワードやインターネットしか使う機会がなかったが、経営システム入門実験...の講義を受けることで C 言語や Excel について学ぶことができ、今ではとてもパソコンを活用できていると思います。
4	自分は理解できずにパソコンを使用している一人であると言える。
5	自分たちのような学生は EXCEL や WORD などの機能を利用して...使用し、非常に役に立っている。
11	インターネットを使用すれば、...ありとあらゆることを調べるができます。
12	「ワード」、「エクセル」、「パワーポイント」、「インターネット」...の使い方を習ったという記憶があります。

表 2: 80 点超 vs 50 点以下の比較 (最終回アンケート)

80 点超 (代表キーワード: 理解 時間 少し 部分 歴史 時代 進め 考え 無理 機会)	
順位	代表センテンス
2	...半年間で行われる授業でコンピュータのすべてを理解する事は到底無理な話であると思うが、さわりの部分だけでも、どんな仕組みで動いているのか、またどのような技術を使って普段私たちが利用している情報処理ができるかなどを知るいい機会になった。
5	実際に、コンピュータ工学を受講しある程度その歴史から基礎的なコンピュータの動作理論を学んだわけですが、...特に用語を理解することができませんでした。
7	こうして振り返ると、...時代がらということもあって、コンピュータについて知るいい機会であったなと思いました。
6	そのほかにハードウェア、ソフトウェアの分野も授業を受けて興味を持った部分がいくつかあったが、やはり半年という短い期間ですべてを習得することははっきりいって無理であると思う。
50 点以下 (代表キーワード: 自分 友達 システム 知能 工学 全体 入力 公務 情報 人工)	
順位	代表センテンス
2	...友達が各学科にいたので、情報学科の友達にほとんど毎日ラウンジでパソコンについて習い、...ほとんど自分で出来るようになった。
4	情報処理技術者試験や国家公務員試験 (情報分野) に関する情報をつめ、簡単なものを自分の技術習得の道しるべとするために、合格することを目指したい。
5	情報技術者試験は...自分に向いていない分野だとは思いますが、国家公務員は興味があるので公務員の情報関係のことは興味がある。
6	自分は将来人工知能の勉強がしたいと思っています。
16	しかし、情報セキュリティやネットワーク技術は友達にも学んでいる人がいて、少し説明してもらったら、少し興味をわいた。

表 3: 各クラスターの代表キーワードおよび代表センテンスの例 (初回アンケート)

第 1 クラスター (サイズ: 52) (成績基準点平均: 63.4) (代表キーワード: 自分 知識 コンピュータ 分野 将来 興味 技術 大学 管理 授業)	
順位	代表センテンス
1	将来的な視点で見ると、今現在の段階では、コンピュータ技術よりも、経営・管理の技術や手法のほうに興味があり、...
3	将来自分がどんな職業に就きたいか、どんな道に進みたいかなどもまだ全然決まっていないので、今の時点ではコンピュータそのものにも興味があるし経営管理の分野にも興味があるので、まだ分野を絞らずいろいろなことを学んでいきたいと思う。

表 4: 各クラスターの代表キーワードおよび代表センテンスの例 (最終回アンケート)

第 1 クラスター (サイズ: 31) (成績基準点平均: 65.3) (代表キーワード: 工学 内容 経営 授業 意味 少し 考え 半年 中間 システム)	
順位	代表センテンス
1	従来までの経営システム工学の授業では、コンピュータに入っているソフトの使い方に重点がおかれており、経営問題を解決するためのツールとしてコンピュータを使うというこのほうが重要という考えも少しあったように思えます。
3	コンピュータの応用を、利用したり、考えるには、中味を知っていることは必要不可欠であると考えてるので、このコンピュータ工学の授業は、経営システム工学に必要である、と思う。
6	コンピュータ工学の授業は現在経営システム工学では必修となっていますが、経営システム工学の生徒でもコンピュータの中の構造を専門的に知りたいといった生徒はわずかしらないと思う。
第 4 クラスター (サイズ: 5) (成績基準点平均: 70.1) (代表キーワード: 進数 教科 部分 実際 板書 途中 分野 月曜 内容 期間)	
順位	代表センテンス
1	...板書の文字の大きさが小さかったために見えなかったところがあったので、もっと大きめに文字を書く、...
8	しかし、板書が見にくい部分が多くそれで困ることがよくあったので、もう少し見やすい板書にしてもらいたいと感じた。
第 13 クラスター (サイズ: 2) (成績基準点平均: 60.3) (代表キーワード: 多く 暗号 文字 大学 最初 ネットワーク 機会 セキュリティ 黒板 量子)	
順位	代表センテンス
3	黒板の文字も見やすかったし、書くスピードもあまり遅くなかったので、とてもノートが取りやすかったです。
4	それから黒板の文字も小さかった。