WEB シラバスにおける初期クラスタ選択方法の違いによる安定性の比較

M2011MM006 青木康裕
指導教員：石崎文雄

1 はじめに

大学機関は講義内容を学生に把握してもらうために, シラバスを提供している. 近年, シラバスを WEB 上に掲載する形を用いる大学が増えており, 2010 年には, 大学は 44.5% の公開率となっている [1].

高校生の大学決定の指針の一つとなる WEB シラバスを対象としたクラスタリングによる分類が研究されているが, まだ研究途中の段階である. この研究が進み, 信頼できるクラスタリング結果を得ることができるようになることが重要である.

本研究において, 2 章では WEB シラバスのデータを対象としたクラスタリングを行っている関連研究について述べる. 3 章では本研究で用いる手法について説明する. 本研究においては, WEB シラバスデータ内の単語に対して TF-IDF 重み付けを行う手法を用いてからクラスタリングを行う. 4 章では本研究の手順と用いるアルゴリズムの実装内容について述べる. 5 章では実験の詳細と実験結果について述べる. 実験内容は 3 章で提案した各アルゴリズムによるクラスタリングを行う. 実験結果の評価方法として, 各アルゴリズムでのクラスタリング結果の比較を行い, 考察を行う. 比較対象として, 各クラスタ重心と分類された WEB シラバスデータとの平均距離, 分類精度の高さ, 計算回数, 初期配置からのデータ変更数の少なさを用いる. 6 章では 5 章の実験結果を踏まえて, まとめを行う.

2 WEB シラバスに関連する研究

2.1 クラスタリング

本節ではクラスタリングについての説明を行う. クラスタリングとはデータマイニング技術のひとつである. データマイニング技術とは, 大量に蓄積されたデータを対象としてデータを解析し, 項目間の相関関係やパターンなど直接明示されていない, 役立つ可能性があり自明でない知識を発掘する技術である. データマイニングは大きく分けると教師あり分類と教師なし分類がある. 教師あり分類は教師データ (訓練データ) と呼ばれるデータを用いて, 似た教師データを参考に入力データから出力データを予測する. 教師なし分類は教師あり分類のように最初から分類の基準を設けず, 何らかの基準を入力データから設け, それを最適化し, 出力データを求めるものである. クラスタリングは教師なし分類に属する手法である.

クラスタリングの使用目的は様々で, クラス定義するためのパターン分類, パターンの統計的構造の把握, データの分割などにも用いられる. 具体的な例として, WEB における情報検索のキーワードの分類のような利用方法がある.

2.2 先行研究との相違点

本節では先行研究の紹介と本研究との相違点を表 1 に示す. 野澤ら [2] の先行研究では WEB シラバスのデータに対してクラスタリング手法を用いたデータマイニングを行なっている. WEB シラバスデータへのクラスタリングアルゴリズムの一つとして k-means 法を利用している. しかし, 坂井ら [3] や, 小野田ら [4] の先行研究では k-means 法は初期配置が完全にランダムであるため, 初期配置に対して結果が大きく影響を受けてしまうという弱点があることを指摘しており, k-means 法の改良アルゴリズムとして KKZ 法, k-means++法や独自の提案手法を紹介している. また, 小野田らによって提案されたコサイン距離を用いた手法は, 小規模なデータセットを対象とした実験では良い性能を示していた.

本研究では WEB シラバスからの収集方法はシステムを用いた自動収集ではなく, 手動による収集で行う.

次に分類方法において, WEB シラバスのクラスタリングの先行研究では k-means 法を利用していたが, 本研究においては KKZ 法と k-means++法を選択する.

3 本研究内容の説明

3.1 データセットの入手方法の紹介

本節では本研究に用いるデータセットについての説明を行う. WEB シラバスのデータセットの入手方法は主に 2 種類あり, 一つ目の方法は WEB 上からの自動収集であり, 二つ目の方法は手動による収集である. 本研究では後者の手法を用いるものとする. これは WEB 上からの自動収集がまだ研究中の分野であり, その収集精度も年々上昇してきているが, 収集精度が 100% にまだ至っていないことから, 本研究の目的であるクラスタリングに用いるデータに不備がないようにするためである. 手動で入手する WEB シラバスのデータセットは南山大学情報理工学部, 三重大学理系学部のものである. それぞれ公式 HP に記載されているものをデータセットとして収集する対象とする.

3.2 専門用語リスト

本節では本研究において使用する専門用語リストを作成する. 専門用語リストは専門用語自動抽出システムの一つである『TermExtract』[1]を用いる. また, 形態素解析を行うために『茶筅』[2]を用いる.

3.3 データセットへの重み付けによる定量化

本節では本研究において使用する各シラバスデータに重み付けを行う. 作成した専門用語リストを用いて 各シラ

[1] http://gensen.dl.itc.u-tokyo.ac.jp
[2] http://chasen-legacy.sourceforge.jp

表 1 研究内容比較表

論文名 (文献番号)	目的	アルゴリズム
[2]	日本語 WEB シラバスのクラスタリング	k-means 法
[3]	アルゴリズムの改良	KKZ 法と k-means++法の改良
[4]	アルゴリズムの改良	KKZ 法と k-means++法の改良
本研究	日本語 WEB シラバスのクラスタリング	k-means 法, KKZ 法, k-means++法

バスデータに TF-IDF(Term frequency-inverse document frequency) で重み付けを行う. TF-IDF とは, ある文章中において特定の単語の出現率が高いが, 全文書中ではその単語の出現率が低い場合ほど重みが高くなる手法である.

全ての WEB シラバスにおいて専門用語とみなされた単語の数が n 個存在し, WEB シラバスデータの総数が N のとき, $i = 1, 2, 3, 4, 5, \cdots, n$ および $j = 1, 2, 3, 4, 5, \cdots, N$ とするとき, ある目的単語 i が出現した WEB シラバスの総数を df, ある目的単語 i が出現する WEB シラバス j における i の出現回数を tf としたとき, TF-IDF の値 w は次の式 (1) で求められる.

$$w_{ij} = tf_{ij} \cdot log(\frac{N}{df_i}) \tag{1}$$

3.4 k-means 法

本節では先行研究において使用されたアルゴリズムについて説明する. 使用されたクラスタリング手法は k-means 法と呼ばれるものである. これは分割最適化手法と分類されており, クラスタ毎の重心を用いて, 与えられたクラスタ数に分類することを目的としている.

全てのクラスタ数が k, $i = 1, 2, 3, 4, 5, \cdots, k$ となる i が存在するとき, あるクラスタ X に対してのクラスタの重心が $\bar{x}$, クラスタ X に属する WEB シラバスデータを x としたとき, 次の式 (2) という計算をすることで分類されるクラスタを表すことができる.

$$Err(X_i) = \sum_{i}^{k} \sum_{x \in X_i} (\|x - \bar{x}_i\|)^2 \tag{2}$$

k-means 法アルゴリズムは以下のステップで行われる.

1. 任意の k 個のクラスタ重心 C_i を一様にランダムに選択
2. 全てのデータを, 各データ点 x_j, $j \in 1, \cdots, n$ から最も近いクラスタ i に分類する
3. クラスタ毎の重心を次の式 (3) にしたがって求める.

$$x_i = \frac{1}{|X_i|} \sum_{x \in X_i} x \tag{3}$$

4. ステップ 2,3 をクラスタに変化がなくなるまで繰り返す
5. クラスタに変化がなくなったら終了

ステップ 1 では初期化を行う. これは WEB シラバスデータの集合をランダムに k 個に分割, 初期クラスタを生成するためのものである. ステップ 2 では各クラスタにデータを分類する. ステップ 3 ではクラスタ毎の重心を式 (3) したがって求める. ステップ 4 ではステップ 2, ステップ 3 をクラスタが変化することがなくなるまで繰り返す. ステップ 4 でクラスタに変化が生じなかった場合, 最適解が算出されたものとし, クラスタリングを終了する.

3.5 KKZ 法

本節では KKZ 法についての説明を行う. KKZ 法は k-means 法の初期配置が完全にランダムである点を変更したアルゴリズムである. その方法として, 最初に各データ間の最短距離を測定し, 最長となる 2 つのデータを初期クラスタとして選択する. その後, その 2 つのクラスタからの距離が一番遠いデータを次のクラスタに選択する. そして, クラスタに選択されたデータを含めて同様の作業を行い, k 個のクラスタを選択するまで繰り返すというアルゴリズムであり, 以下のステップで行う.

1. データ間距離 dis が最長の 2 データを C_0, C_1 に選択
2. 決定したクラスタと各データとの $D(x)$ の和 $MD(x)$ を計算
3. ステップ 2 の結果, $MD(x)$ が最大となるデータ x' をクラスタ C_i, $i \in 2, 3 \cdots, n-1$ に選択
4. 全ての n 個のクラスタが決定するまで 2,3 を繰り返す

3.6 k-means++法

本節では k-means++法についての説明を行う. k-means++法は KKZ 法と異なり, 初期配置における一つ目の初期クラスタのみをランダムに選択し, 残りのクラスタ重心位置を, 測定した距離を利用して選択していく手法である. その方法として, 一つ目の初期クラスタをランダムに選択した後, そのクラスタから最短距離が最長となるデータを次のクラスタに選択する. そして, クラスタに選択されたデータを含めて同様の作業を行い, k 個のクラスタを選択するまで繰り返すというアルゴリズムであり, 以下のステップで行う.

1. ランダムに選んだデータを C_0 に選択
2. 決定したクラスタと各データとの距離 $D(x)$ を計算
3. $\frac{D(x')^2}{\sum_{x \in X} D(x)^2}$ を最大にするデータ x' を次のクラスタ C_i, $i \in 1, \cdots, n-1$ に選択

4. 全ての n 個のクラスタが決定するまで 2,3 を繰り返す

4　実験の説明

4.1　実験の手順の説明

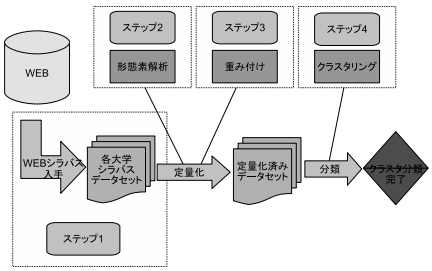

図 1　実験の流れ

本節では本研究で行う実験の手順についての説明を行う．手順としては, 図 1 のような流れで以下のステップで進める．

1. 対象とする WEB シラバスを選択・収集
2. WEB シラバスの内容を茶筅を用いて形態素解析
3. 各データに対する重み付けを tf-idf を用いて実行
4. 各アルゴリズムによるクラスタリング

4.2　データセットの構築

本節では本研究で使用するデータセットについて, 使用項目数や種類などを説明する．データセット内容の使用項目は「科目名」,「授業概要」,「学修目標」,「授業計画」,「教科書」の 5 項目である．

対象とする WEB シラバスデータとして, 南山大学情報理工学部から合計 306 科目分, 三重大学理系学部から合計 910 科目分をそれぞれ使用するものとする．

4.3　WEB シラバスデータセットの定量化

本節では 4.3 節で作成した専門用語リストを用いて, 各 WEB シラバスデータに TF-IDF による重み付けを行うことで WEB シラバスデータの定量化を行う．実行した結果, 重み付けによるデータの分布は値が 1000 を越えるデータが南山大学は 9 個, 三重大学は 99 個存在した．今回は上位 300 までのデータを用いてクラスタリングを行う．

5　クラスタ分類実験

5.1　変更がなかったデータ数の比較

本節では南山大学, 三重大学それぞれの WEB シラバスデータを用いた際の初期配置によって分類されたデータ数と, その後の計算によって分類されたデータ数に着目する．図の x 軸は実験回数とし, 100 回目までの結果を示している．y 軸は実験毎の変化しなかったデータ数を割合で示している．これは南山大学と三重大学の WEB シラバスデータ数が異なるため, 比較できるようにしたものである．図 2, 図 3, 図 4 では k-means 法, KKZ 法, k-means++法それぞれの実験結果をまとめて示す．

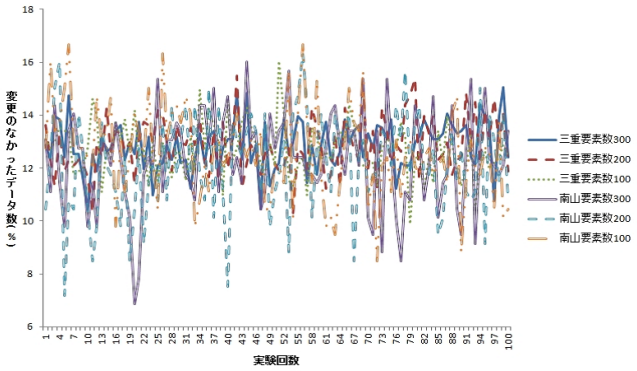

図 2　変更がなかったデータ数の比較 (k-means 法)

図 2 より, k-means 法を用いた場合に変更がなかったデータ数は, ともに実験を行う度に変化しており, 不安定な結果を示していることがわかる．また, 南山大学の場合が 7〜17%, 三重大学の場合が 10〜16% となっていることも分かり, 南山大学の場合は約 12.4%, 三重大学の場合は約 12.8% という平均値をとっており, ほぼ変化がないことがわかる．

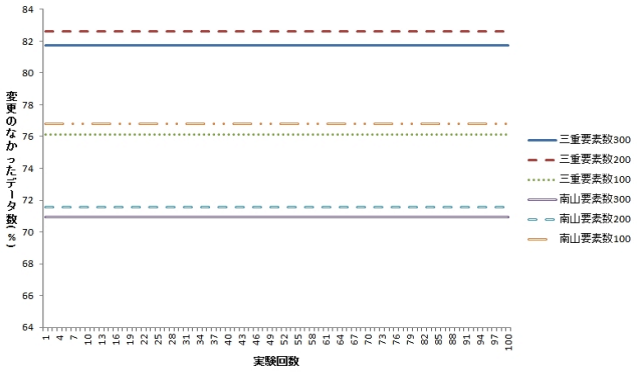

図 3　変更がなかったデータ数の比較 (KKZ 法)

図 3 より, KKZ 法を用いた場合に変更がなかったデータ数は, ともに実験を何度おこなっても一定の値を示しており, 安定した結果を示していることがわかる．また, 南山大学の場合が 70〜77%, 三重大学の場合が 76〜83% となっていることも分かり, 南山大学の場合は約 73.1%, 三重大学の場合は約 80.2% という平均値をとっている．そして, 南山大学の要素数 100 や, 三重大学の要素数 100 の場合が他の要素数の場合から大きく異なる値を示していることから, 要素数が増えるに連れて結果が等しくなりやすい傾向にある．

図 4 より, k-means++法を用いた場合に変更がなかったデータ数は, 南山大学の場合は要素数が増えると安定していたが, 三重大学の場合は実験を行う度に変化しており, 不安定な結果を示していることがわかる．また, 南山大学の場合が 97〜100%, 三重大学の場合が 61〜98% となっていることも分かり, 南山大学の場合は約 99.6%, 三重大学の場合は約 94% という平均値をとっている．

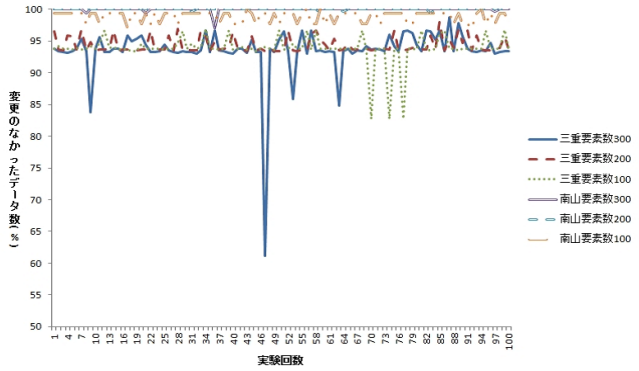

図 4　変更がなかったデータ数の比較 (k-means++法)

5.2　クラスタ精度の比較

本節ではクラスタリングによってクラスタ分類されたデータがどの程度の精度で分類されたかの比較を行う．図 2 では南山大学の WEB シラバスデータの場合の要素数 300 での結果を示す．y 軸はクラスタ精度を示す．また，表 2, 表 3, 表 4 に k-means 法, KKZ 法, k-means++法それぞれの実験結果を示す．示した数値は南山大学と三重大学で WEB シラバスデータ数が違うため, 割合で示したものである．

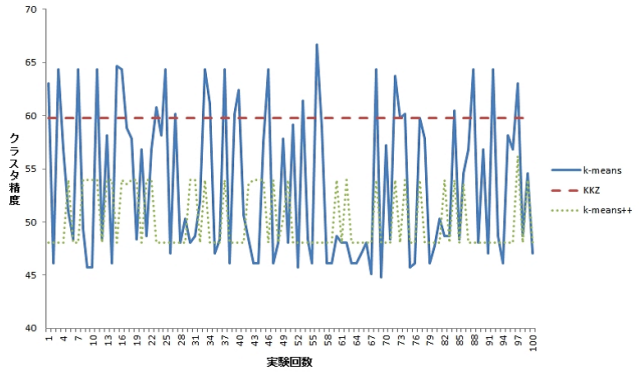

図 5　クラスタ精度の比較 (南山大学:要素数 300)

図 2 より, KKZ 法が 100% で常に同じ値を示しており, 安定していることがわかる．

表 2　k-means 法でのクラスタ分類結果の比較

k-means	南山大学	三重大学
要素数 300	52.95	58.29
要素数 200	54.63	58.73
要素数 100	54.72	57.52

表 2 より, k-means 法は WEB シラバスデータ数, 要素数に限らず 50〜60% 程度の精度であることがわかる．

表 3 より, KKZ 法は南山大学の場合に約 59% , 三重大学の場合に 72〜78% 程度であることから, WEB シラバスデータ数, 要素数が多い場合のほうが精度が高くなることがわかる．

表 3　KKZ 法でのクラスタ分類結果の比較

KKZ	南山大学	三重大学
要素数 300	59.80	77.58
要素数 200	59.80	77.58
要素数 100	59.48	72.86

表 4　k-means++法でのクラスタ分類結果の比較

k-means++	南山大学	三重大学
要素数 300	50.08	89.89
要素数 200	49.16	90.42
要素数 100	47.58	89.81

表 4 より, k-means++法は南山大学の場合に 47〜50% , 三重大学の場合に 89〜90% 程度であることから, WEB シラバスデータ数, 要素数が多い場合のほうが精度が高くなることがわかる．

6　おわりに

本節では本研究に対するまとめを行う．まず, クラスタリング前とクラスタリング終了後での分類されたクラスタの変更がなかったデータ数が 70〜80% であり, その結果を安定して得られていることから KKZ 法, 93〜99% の変更がなかった k-means++法が優秀であることがわかる．次にクラスタ分類精度の結果から, KKZ 法, k-means++法がデータ数が多い場合のほうが精度が上がることがわかる．そして, KKZ 法は常に 100% で同じ結果を示すことができている．

以上より, KKZ 法が k-means 法, k-means++法に対して優れているという結果となった．よって, KKZ 法が k-means 法, k-means++法に比べて WEB シラバスに対するクラスタリングにおけるクラスタの初期配置方法として適切である．

参考文献

[1] ウェブシラバスを公開している大学, 短大, 高専, WEBS（Web Education Books Software）, http://powercampus.jp/top/stat/ウェブシラバス公開機関一覧.htm （accessed 2012.6）

[2] 野澤 孝之, 井田 正明, 芳鐘 冬樹, 宮崎 和光, 喜多 一, “シラバスの文書クラスタリングに基づくカリキュラム分析システムの構築,” 情報処理学会論文誌 46(1), pp.289-300, 2005.

[3] 坂井 美帆, 山田 誠二, 小野田 崇, “独立成分分析による k-means 法の初期値設定手法の提案,” 人工知能学会全国大会論文集 (CD-ROM), pp.ROMBUNNO.2G2-OS9-2(論文番号) 2010.

[4] 小野田 崇, 坂井 美帆, 山田 誠二, “k-means 法の様々な初期値設定によるクラスタリング結果の実験的比較,” The 25th Annual Conference of the Japanese Society for Articial Intelligence, 2011 1J1-OS9-1.