コサイン類似度を用いた部活動推薦システムの構築

2017SC005 福谷駿人
指導教員：河野浩之

1　はじめに

近年文部科学省によると 20 代以上の人が定期的に運動している割合が増加傾向にある [6]. 最低でも週に一回体を動かす人が増えているが, 大学生が体育会系部活動に所属する人数は増えた記録がない.

新入生が部活を検討するにもほとんどの大学が体育会系部活の種類が 30 種類以上あるので, 新入生もすべてを検討するのは難しい. そこで学生が希望する条件に合っている部活動を推薦して学生が部活動を検討する支援をする.

2　推薦システムの先行研究

部活動を推薦するうえで本研究では就活生を対象にした就職先推薦システムの構築を行った研究を参考にした [1].

表 1　先行研究の使用したアルゴリズム

著者	使用したアルゴリズム
何ら [1]	内容ベースフィルタリングと協調フィルタリング
湯本氏 [2]	ラフ集合と決定ルール

湯本氏は企業の集合をラフ集合として特徴を抽出する方法で就職先推薦システムを開発した. これは求職者の個人情報を用いずにサンプルに対し評価をひととおり行うだけで嗜好を分析する. また各企業の情報からのユークリッド距離によって企業を推薦している.

何らは数ある企業の情報を数字型データ, 単一選択肢型データ, 複数選択肢型データの 3 つの種類に数値化を行った. また就活生の情報も同様に数値化を行っている. そして類似度計算にはコサイン類似度 (1) 式を用いている.

$$\cos\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n}(A_i^2)} \cdot \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

本研究は何ら [1] の数値化を部活動に置き換えて使用するため, 類似度計算も彼らの方法を使用することとした.

3　部活動推薦システムの提案手法

部活動のデータは南山大学と名城大学のサイトに記載してある部活動の表を使用する. 図 2 のような各部活動の内容がかかれている表データを取得しデータベースを作成していく. 表データの取得には Web スクレイピングを使用する. Web スクレイピングをするために Octoparse というツール使い, 大学に存在する部活動データを抽出していく. 今回使うデータはサイトに記載してあるためまとめて抽出できないので 1 つずつ抽出をしなければならない. 今回抽出したデータは Excel ファイルに格納する.

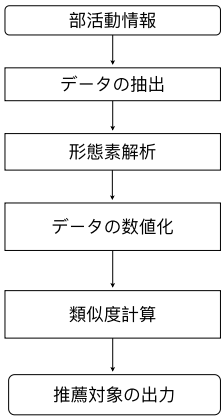

図 1　システムのフローチャート

図 2 にあるように, 部活動の情報は文章で記載されている. このような文字列を処理するには形態素解析を行う. 各項目の内容を分かち書きし, キーワードとなる単語を読み取れるようにする. この形態素解析には様々なツールがあるが, 今回システムを Python3 で実行するので Python3 で実行できる形態素解析に MeCab があるので MeCab を使用する.

活動曜日・時間	火・水・金
部員数	27名
活動場所	名古屋スポーツセンター
所属リーグ	東海リーグ１部
戦績	平成２７年度　愛知県学生アイスホッケー競技会　順位１位　インカレ出場 平成２８年度　愛知県学生アイスホッケー競技会　順位３位 平成２９年度　愛知県学生アイスホッケー競技会　順位４位
クラブ作成Webページ	http://nanzanicehockey.wixsite.com/nanzanicehockey

図 2　南山大学クラブ紹介のサイト

学生のデータは学生に入力させる方式を取り, 入力した内容と部活動の内容を何ら [1] にならって数値化する. 数値化した学生データと各大学に存在する部活動とのコサイン類似度を求めその数値が高いものを出力する.

4　実験結果

入力した学生データと部活動データは数字型データ, 単一選択肢型データ, 複数選択肢型データの 3 種類の数値化によって新たにベクトルを作成する. 数字型データは学生データ側と部活動データ側の数字の偏差の程度をとり, 学生側に 1 を格納し, 部活動側に偏差の程度を格納する. 偏差の程度とは小さいほうが大きい方に占める割合の値である. 今回のデータでは活動日数, 部員数に該当する.

表 2 南山大学の類似度が最も高かった部活

	活動日数	土日の有無	部員数	活動場所	コサイン類似度
アイスホッケー部	3	休日活動なし	27	大学外	1.000
漕艇部	2	土	23	大学外	0.970
フィギュアスケート部	2	土	25	大学外	0.967
準硬式野球部	1	休日活動なし	34	大学外	0.944
航空部	3	土日	14	大学外	0.933

単一選択肢型データでは 2 つの項目に番号をつけ, 選択した番号が一致しているならばお互いに 1, 違っていた場合は学生側に 1, 部活動側に 0 を格納する. 今回のデータでは活動場所が大学内か大学外かというのに該当する.

複数選択肢型データは選択肢すべてに異なる番号をつけ, 選んだ選択肢の類似度出す. そのためにジャッカード指数 (2) 式を用いる. A と B に含まれている共通の番号の個数を数え, その数を A と B の両方に含まれた異なる番号の個数で割った値がジャッカード指数である. これを部活動側に格納し, 学生側は 1 を格納する. 今回のデータでは土日の有無に該当する.

$$Jacc = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

コサイン類似度の結果が高かかったものを表 2 に示す. 入力したデータはアイスホッケー部のデータを入力し, システムの動作を確認する. 表 2 のアイスホッケー部の類似度は 1 を示しており, 完全一致を表しているのでシステム自体はうまく作動していることがわかる. その他の部活動のデータを見るとデータが一致しているもの, または近しいものがどれも 2 項目ずつある. よって本実験の結果は 2 項目ほど一致していれば類似度は高くなることを示した.

類似度が 0.8 台の部活動は一致しているものがなく, 類似度がもっとも低かった部活は逆に一致しているものがあった. しかし類似度が低い原因は数字型データで処理した部員数であった. この部員数はほかのデータに比べて誤差が大きくなることがあるものなので部員数の誤差が類似度に大きく影響することがわかった.

同じことを名城大学のデータを使って行った. 名城大学では部活動データに「部費」の項目があったので「部費」の項目を新たに追加して各部活動との類似度を計算した. 「部費」は数字型データであるが, 学生が希望した月額に対して部活動側がその値より安いのであればそれに越したことはないので, 月額が同じもしくは希望額より安い場合は偏差の程度ではなく両方のデータに 1 を格納する. また名城大学の「部費」の項目は年, 半期, 月といったように部費の額が統一された単位でなかったため, すべて月額で処理している.

類似度の結果は同じように部員数の誤差が大きく関わってきていた. それと項目が少ないことが原因で類似度が似たような値になっている部活動も見られた. 「部費」の項目を 1 つ増やした程度では類似度に差が出ないので, もっと多くの部活動データの項目を増やさなければならないことがわかる.

5 むすび

類似度計算を行ううえで, 項目をもっと増やす必要がある. 項目が少ないと各部活動との類似度に差が出ないので, 似たような値になってしまう. そうすると推薦すべき部活動を定められなくなる. また先行研究 [1] は協調フィルタリングなどの手法も使っており精度は高い. しかしこのようなことをするには部活動側の学生データなど別視点のデータが必要になる.

本研究の精度を高めるには使用するべきデータが不足していることがわかる. そして何より学生が推薦結果をどのように評価するかが重要だが, 現状の結果では入部までの決定を促すことはできないであろう.

参考文献

[1] 何陽, 長谷川忍, “ 就職活動支援システムにおける企業情報推薦機能の開発-企業の採用項目と就活生のアピールのギャップに注目して, ” 情報処理学会, pp. 1-8, 2019.

[2] 湯本 真樹, “ 適性を考慮した条件緩和を用いたラフ集合の決定ルールによる就職先推薦システムの開発, ”電気学会論文誌C（電子・情報・システム部門誌), pp. 100-112, 2020.

[3] 湯本 真樹, “ 利用者による 3 段階評価にもとづくラフ集合による学生向け賃貸物件推薦システムの開発, ” 電気学会論文誌C（電子・情報・システム部門誌）, pp. 441-451, 2018.

[4] 神嶌 敏弘, “ 推薦システムのアルゴリズム 1, ” 人工知能学会, ” pp. 826-837, 2007.

[5] 南山大学「南山大学 学生課 学生生活 クラブ 紹 介 」http://office.nanzan-u.ac.jp/student-services/clubs/c006.html.

[6] 文部科学省「スポーツ実施率：文部科学省」https://www.mext.go.jp/a_menu/sports/jisshi/1294610.htm.