УДК 004.912

ЗАСТОСУВАННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ ДЛЯ ВИЗНАЧЕННЯ СФЕРИ ДІЯЛЬНОСТІ ПРАЦІВНИКІВ ПРИ ПРОГНОЗУВАННІ КАР'ЄРНОЇ ТРАЄКТОРІЇ

Дворник В. А.

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Україна, Київ

Розглянута задача визначення сфери діяльності працівників на основі їх резюме у текстовому вигляді для прогнозування кар'єрної траєкторії працівника. Приведено постановку задачі. Визначено, до яких моделей зводяться досліджувані проблемні ситуації, та, які методи можуть бути застосовані до розв'язання поставленої задачі. Розглянуто відомі алгоритми кластеризації та їх практичне застосування на прикладі задачі визначення можливих варіантів майбутніх професій. Обґрунтовано вибір методів дослідження: наведено їх детальний опис та висвітлено переваги та недоліки обраних методів.

Ключові слова: кластеризація, класифікація, кар'єрна траєкторія, прогнозування, сфера діяльності.

Дворник В. А. Применение методов кластеризации для определения сферы деятельности работников при прогнозировании карьерной траектории / Национальный технический университет Украины «Киевский политехнический институт имени Игоря Сикорского», Украина, Киев

Рассмотрена задача определения сферы деятельности работников на основе их резюме в текстовом виде для прогнозирования карьерной траектории работника. Приведена постановка задачи. Определено, к каким моделям сводятся

исследуемые проблемные ситуации, и, какие методы могут быть применены к решению поставленной задачи. Рассмотрены известные алгоритмы кластеризации и их практическое применение на примере задачи определения возможных вариантов будущих профессий. Обоснован выбор методов исследования: приведено их подробное описание и освещены преимущества и недостатки выбранных методов.

Ключевые слова: кластеризация, классификация, карьерная траектория, прогнозирования, сфера деятельности.

Dvornyk V. A. Application of clustering methods for determining the field of employees activity in projecting of career trajectory / National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine, Kiev

The problem of determining the area of work of employees on the basis of their resume in text form for forecasting the career trajectory of the employee is considered. The statement of the problem is given. It is determined, to which models are reduced investigated problem situations, and what methods can be applied to the solution of the given problem. Famous algorithms of clustering and their practical application are considered on the example of the problem of determining possible alternatives of future professions. The choice of research methods is justified: methods are described in detail and the advantages and disadvantages of the chosen methods are highlighted.

Keywords: clustering, classification, career trajectory, forecasting, field of activity.

Вступ. У сучасному світі проблема професійного розвитку особистості та побудови кар’єрної траєкторії набуває особливої значущості. Вона обумовлена об’єктивною потребою вивчення питань

мотивації і стимулювання особистості, питань формування професійних навичок та аналізу професійної кар'єри працівника. Сучасні дослідники і теоретики прагнуть розглядати кар'єрне зростання в контексті змін, які зачіпають як і працівника компанії, так і саму компанію, а також навколишнє середовище, в якій існує компанія.

Перевиробництво непотрібних фахівців, підготовка недостатньо кваліфікованих працівників, низька частка зайнятості випускників закладів вищої освіти – все це проблеми, закорінені в неправильному виборі професії [1].

Мета статті. Метою цієї роботи є обґрунтування доцільності визначення сфери діяльності працівників на основі їх резюме у текстовому вигляді для прогнозування кар'єрної траєкторії працівника, обґрунтування доцільності моделей, до яких зводяться досліджувані проблемні ситуації, а також обраних методів, які можуть бути застосовані до розв'язання поставленої задачі.

Для досягнення цілей застосовуються методи кластеризації для автоматичної обробки текстів, написаних природною мовою, якими є резюме у текстовому вигляді, в яких буде міститись інформація про сферу роботи працівника, стаж роботи, набір навичок, вмінь та знань, тематику проектів, на яких він працював, мотиваційні листи, есе та тести з професійної орієнтації з вільними відкритими відповідями. Враховуючи важливість ролі інформаційних систем та технологій у розвитку суспільства, будемо розглядати модель кар'єрного процесу для галузі знань інформаційні технології.

Постановка задачі. У задачі визначення сфери діяльності працівників при прогнозуванні кар'єрної траєкторії в якості вхідної інформації розглядатимуться резюме у текстовому вигляді, в яких буде міститись уся інформація про професійну кар'єру працівника, а

також мотиваційні листи, есе та тести з професійної орієнтації з вільними відкритими відповідями.

На виході отримаємо набір професійних сферу роботи працівника, до яких він прив'язаний зараз, тобто, отримаємо класифікацію робітника до набору певних професій.

При подальшому прогнозуванні кар'єрної траєкторії на виході отримаємо список рекомендованих професій для подальшого розвитку кар'єри та побудови кар'єрної траєкторії.

Для фільтрації, рубрикації і кластеризації вхідних даних, автоматичного анотування документів, пошуку схожих документів і дублікатів, автоматизованої оцінки якості вільних розгорнутих відповідей пропонується залучити методи кластеризації текстових даних.

Опис методу розв'язання задачі. Розглянемо методи кластеризації. Кластеризація – це поділ множини вхідних векторів на групи (кластери) за ступенем «схожості» один на одного. Для того, щоб можна було порівнювати два об'єкти, потрібно мати критерій, на підставі котрого і буде відбуватися порівняння. Зазвичай, як правило, таким критерієм є відстань між об'єктами [2].

Алгоритми кластеризації поділяються на дві основні категорії [3]:

1. Чіткі та нечіткі алгоритми
2. Ієрархічні і плоскі

Чіткі алгоритми ставлять у відповідність кожному об'єкту з вибірки номер визначеного кластера, тобто кожен об'єкт належить лише одному кластеру. Нечіткі алгоритми кожному об'єкту ставлять у відповідність набір дійсних значень (чисел), що показують ступінь відношення об'єкта до усіх кластерів. Тобто, кожен об'єкт відноситься до кожного кластеру з певною ймовірністю.

Наступна категорія – це ієрархічні алгоритми, котрі будують не одне розбиття вибірки на непересічні кластери, а систему вкладених розбиттів. Таким чином, на виході отримується дерево кластерів, коренем якого є вся вибірка, а листками – найдрібніші кластери. Плоскі алгоритми будують одне розбиття об'єктів на кластери.

На рисунку 1 показано більш детальну класифікацію алгоритмів кластеризації [4].

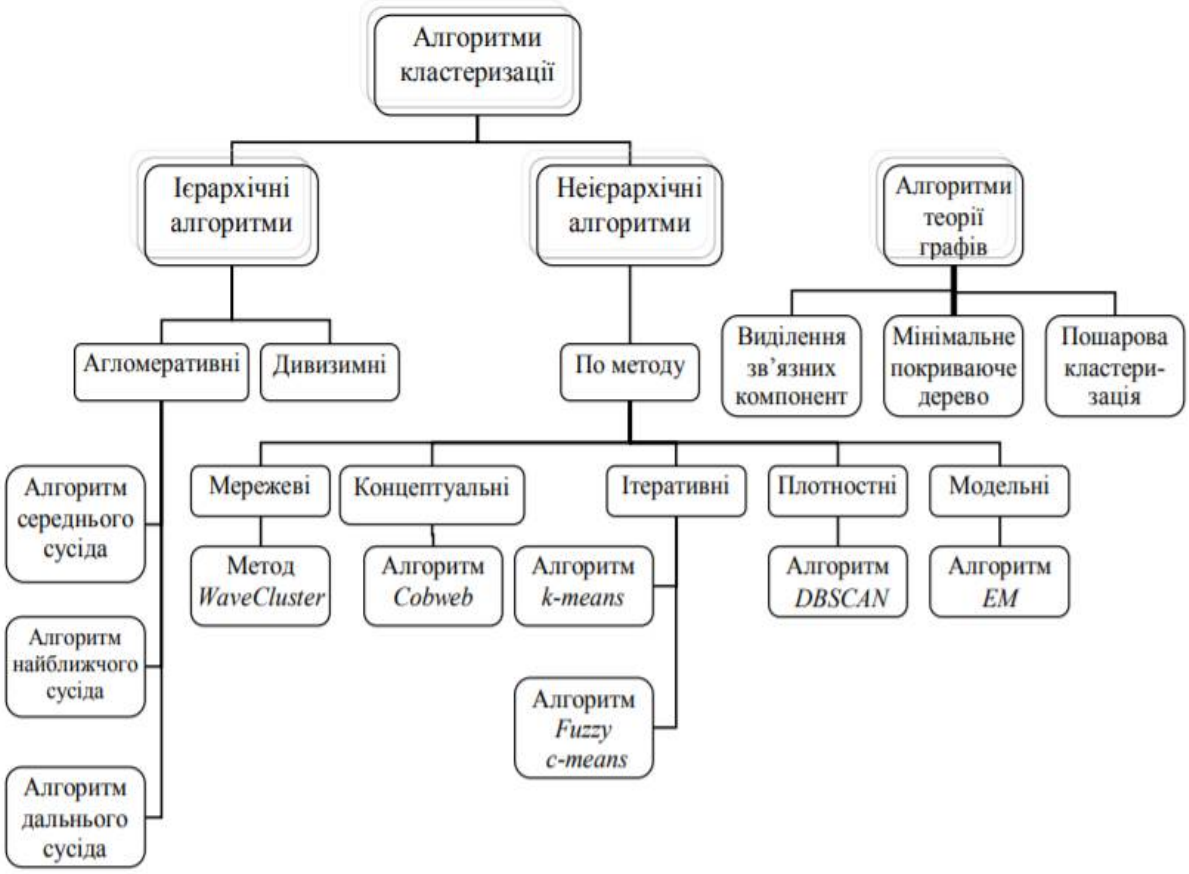

Рисунок 1. Класифікація алгоритмів кластеризації

Метод k-середніх (k-means) – це спеціальний алгоритм кластеризації, котрий на вході має масив даних, який ми хочемо згрупувати у кластери, а точніше – в k кластерів.

Вхідними даними в методі k-середніх є тільки матриця. Як правило, формується вона так, щоб кожен рядок представляв окремий приклад (зразок), а кожен стовпець – окрему ознаку або,

користуючись термінами з статистики, фактор. Зазвичай говорять, що є N прикладів і D ознак, так що утворюється матриця розмірності $N \times D$.

В алгоритмі методу k-середніх є два основних етапи. Спочатку вибирається k різних центрів кластерів – як правило, це просто випадкові точки в наборі даних. Потім переходять до основного циклу, який також складається з двох етапів. Перший – це вибір, до якого з кластерів належить кожна точка з X. Для цього береться кожен приклад і вибирається кластер, чий центр ближче всього. Спочатку вибираються центри випадковим чином. Другий етап – заново обчислити кожен центр кластера, ґрунтуючись на безлічі точок, які до нього приписані. Для цього беруться всі відповідні приклади і обчислюється їх середнє значення, звідси і назва методу – «метод k-середніх». Все це робиться до тих пір, поки не припиниться зміна в розподілі точок по кластерам або в координатах центрів кластерів [5].

Проте, на жаль, алгоритм k-середніх не справляється із задачею, коли об’єкт не належить жодному кластеру або належить до різних кластерів у однаковій мірі.

З цією проблемою k-means чудово справляється алгоритм c-середніх (c-means). Замість точної відповіді на запитання до якого кластеру відноситься об’єкт, алгоритм визначає ймовірність належності об’єкту до того чи іншого кластеру. Таким чином, твердження вигляду «об’єкт В належить до кластеру 1 з імовірністю 85%, до кластеру 2 – 15%» вірне і набагато зручніше.

Алгоритм c-середніх (c-means) – це модифікація методу k-means. Далі наведено кроки роботи алгоритму [4]:

1. Вибір початкового нечіткого розбиття n об'єктів на k кластерів шляхом вибору матриці належності U розміром $n \times k$.

2. Визначення значення критерію нечіткої похибки із застосуванням матриці

$$E^2(X,U) = \sum_{i=1}^{N}\sum_{k=1}^{K} U_{ik} \left\| x_i^{(k)} - c_k \right\|^2,$$

де c_k – це «центр мас» нечіткого кластера k.

3. Перегрупування (перестановка) об’єктів із метою зменшення значення критерію нечіткої помилки.

4. Перехід до пункту 2 до тих пір, поки зміни матриці U не стануть незначними.

Застосовування алгоритму с-середніх може бути недоцільним, якщо число кластерів заздалегідь невідоме або є необхідність віднесення кожного об’єкту до певного кластеру однозначно.

Для кластеризації також можуть бути застосовані такі алгоритми [3]:

1. Алгоритми, котрі засновані на теорії графів
2. Алгоритм виділення зв’язних компонент
3. Алгоритм мінімального покриваючого дерева
4. Алгоритм пошарової кластеризації

Обґрунтування вибору методів дослідження. Розглянуті алгоритми кластеризації достатньо прості, легко реалізуються та показують достатньо високу якість роботи. Незважаючи на це, алгоритми мають і свої недоліки. Наведемо далі переваги та недоліки методу k-середніх [2]:

Переваги:

- простота алгоритму;
- не вимагає обчислення і зберігання матриці відстаней;
- можливість паралелізації;
- лінійна просторова й тимчасова складність.

Недоліки:

- необхідно наперед знати кількість кластерів;

- алгоритм дуже чутливий до вибору початкових центрів кластерів;
- не може впоратись з завданням, коли об’єкт належить до різних кластерів у рівних степенях або не належить ніякому.

Наведемо далі переваги та недоліки методу c-середніх [2]:

Переваги:

- можливість визначення ступеня приналежності елемента до кластеру;
- нечіткість при віднесення об’єкта до кластеру дозволяє включати об’єкти, які знаходяться на границі, в кластери.

Недоліки:

- число кластерів повинно бути відоме заздалегідь;
- комплікативність роботи з об’єктами;
- шукає кластери сферичної форми;
- обчислювальна складність.

Висновки. В роботі розглянуто застосування методів кластеризації текстів природної мови для задачі визначення сфери діяльності працівників на основі їх резюме у текстовому вигляді в процесі прогнозування кар’єрної траєкторії працівника. Описана змістовна постановка задачі, а також методи її розв’язання з їх детальним покроковим описом. Обґрунтовано вибір методів дослідження: наведено їх детальний опис та висвітлено переваги та недоліки обраних методів.

Література:

1. Митюгина, С.В. (2006). Модели професиональной карьеры личности. <https://www.dissercat.com/content/modeli-professionalnoi-karery-lichnosti>(2019, ноябрь, 27)

2. Кластеризация: алгоритмы k-means и c-means. <https://habr.com/ru/post/67078/> (2019, ноябрь, 27)

3. Обзор алгоритмов кластеризации данных. <https://habr.com/ru/post/101338/> (2019, ноябрь, 27)

4. Волосюк, Ю.В. (2014). Аналіз алгоритмів кластеризації для задач інтелектуального аналізу даних. <https://www.mnau.edu.ua/files/faculty/off/kaf-ist/volosyuk/9.pdf> (2019, листопад, 27)

5. Кластеризация методом k-средних. <https://craftappmobile.com/кластеризация-методом-k-средних/> (2019, ноябрь, 27)

References:

1. Mitjugina, S.V. (2006). Modeli profesional'noj kar'ery lichnosti [Models of professional personality careers]. Retrieved from https://cyberleninka.ru/article/v/proektirovanie-individualnoy-obrazovatelnoy-traektorii-i-marshruta-studenta-vuza-buduschego-bakalavra [in Russian]. (2019, November, 27)

2. Klasterizacija: algoritmy k-means i c-means [Clustering: k-means and c-means algorithms]. Retrieved from https://habr.com/ru/post/67078/ [in Russian]. (2019, November, 27)

3. Obzor algoritmov klasterizacii dannyh [Overview of Data Clustering Algorithms]. Retrieved from https://habr.com/ru/post/101338/ [in Russian]. (2019, November, 27)

4. Volosiuk, Yu.V. (2014). Analiz alhorytmiv klasteryzatsii dlia zadach intelektualnoho analizu danykh [Analysis of clustering algorithms for the tasks of intellectual analysis]. Retrieved from https://www.mnau.edu.ua/files/faculty/off/kaf-ist/volosyuk/9.pdf [in Ukrainian]. (2019, November, 27)

5. Klasterizacija metodom k-srednih [K-means clustering]. Retrieved from https://craftappmobile.com/кластеризация-методом-k-средних/ [in Russian]. (2019, November, 27)