Projekt z przedmiotu MED

Informacje kontaktowe:

dr inż. Dominik Ryżko, pok. 318, email: dominik.ryzko@pw.edu.pl

Konsultacje projektowe odbywają się w środy w godzinach 13-15. Preferowany sposób konsultacji: zdalne, z wykorzystaniem aplikacji MsTeams. Aby umówić się na spotkanie przez MsTeams w godzinach konsultacji projektowych proszę o wcześniejszą informację przez czat lub o przesłanie emaila w tej sprawie.

Cel projektu

Celem projektu jest zbadanie własności zaimplementowanego przez siebie algorytmu: czas wykonania, uzyskiwane wyniki dla różnych wartości parametrów algorytmu oraz kilku zbiorów wejściowych. Zależnie od tematu projekt może być realizowany przez jedną lub dwie osoby.

Liczność grup

Przy każdym z tematów znajduje się informacja o liczności grupy. W przypadku projektów jednoosobowych istnieje możliwość rozszerzenia zakresu zadania i powiększenia grupy do 2 osób. W takiej sytuacji należy ustalić z prowadzącym zakres zadania projektowego.

Ważne daty

17.04.2024– termin wybrania tematu projektu.

- Proszę przesłać emailem swoje preferencje projektowe, w temacie wiadomości proszę dodać **[projekt MED]**.
- Proszę uwzględnić w wiadomości 3 preferowane tematy w kolejności preferencji. Projekty zostaną przydzielone zgodnie z kolejnością zgłoszeń i preferencjami.
- Dla projektów realizowanych przez zespoły proszę o adresowanie wiadomości do wszystkich członków zespołu.
- Osoby, które nie prześlą preferencji w wyznaczonym terminie zostaną przypisane do tematu projektu arbitralnie wybranego przez prowadzącego.

08.05.2024 – prezentacja kluczowych elementów rozwiązania (część 1). Kluczowe elementy rozwiązania należy przedstawić osobiście w terminie konsultacji projektowych. Przedstawienie obejmuje:

- przedstawienie idei rozwiązania w formie prezentacji (np. w formacie .ppt);
- omówienie planowanego rozwiązania.

12.06.2024 – ostateczny termin oddania całego projektu (część 2) w terminie konsultacji projektowych. Oddanie projektu obejmuje:

1. prezentację pokazującą główne zagadnienia związane z realizowanym projektem – należy przygotować prezentację (np. w formacie .ppt);
2. pokaz działania oprogramowania;
3. rozmowę dotyczącą testów, uzyskanych wyników i wniosków.

Zasady zaliczenia projektu

Maksymalna liczba punktów z części 1 wynosi 15, a z części 2 wynosi 30 (sumaryczna maksymalna liczba punktów do uzyskania z obu części wynosi 45). Ostateczna ocena zależy od sumarycznej liczby punktów, z uwzględnieniem przedziałów podanych poniżej:

- 0 – 20: (ndst) **2.0**
- 21 – 25: (dst) **3.0**
- 26 – 30: (+dst) **3.5**
- 31 – 35: (db) **4.0**
- 36 – 40: (+db) **4.5**
- 41 – 45: (bdb) **5.0**

Dokumentacja

Dokumentacja powinna zawierać następujące rozdziały:

1. Wprowadzenie i definicja problemu
2. Charakterystyka proponowanego algorytmu/rozwiązania z odniesieniem do literatury
3. Opis implementacji
4. Instrukcja użytkownika (jak uruchomić/korzystać z implementacji)
5. Charakterystyka wykorzystywanych zbiorów danych
6. Wyniki eksperymentów pokazujących właściwości proponowanego rozwiązania
7. Wnioski
8. Bibliografia

Eksperymenty należy przeprowadzić z wykorzystaniem publicznie dostępnych zbiorów danych, np.:

- https://archive.ics.uci.edu/ml/datasets.php,
- http://fimi.uantwerpen.be/data/,
- https://github.com/deric/clustering-benchmark,
- http://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php

Implementacja algorytmu/rozwiązania powinna zostać przygotowana w jednym z popularnych języków programowania. Obydwie części dokumentacji należy przesłać mailem w terminach przedstawionych powyżej.

Proponowane tematy

1. ***Temat własny – wymaga uzgodnienia***

2. ***Reguły asocjacyjne***

Sugerowana liczba osób w grupie: 1.

1. Implementacja algorytmu do odkrywania reguł asocjacyjnych (Apriori, Eclat …) wraz z implementacją wyliczanie miar: współczynnika podniesienia (lift) oraz 3 wybranych: conviction, cosine, gini, Jaacard/ coherence, certainty factor, improvement,

mutualInformation, odd ratio. Porównanie zachowania się wybranych miar w odniesieniu do wartości współczynnika podniesienia.

- Bayardo, R. , R. Agrawal, and D. Gunopulos (2000). Constraint-based rule mining in large, dense databases.
- Berzal, Fernando, Ignacio Blanco, Daniel Sanchez and Maria-Amparo Vila (2002). Measuring the accuracy and interest of association rules: A new framework.
- Brin, Sergey, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur (1997). Dynamic itemset counting and implication rules for market basket data.
- Tan, Pang-Ning, Vipin Kumar, and Jaideep Srivastava (2004). Selecting the right objective measure for association analysis

2. Implementacja algorytmu do odkrywania uogólnionych reguł asocjacyjnych opisanego w: Srikant, Ramakrishnan, and Rakesh Agrawal. "Mining generalized association rules." (1995)
3. Tworzenie nieredundantnych reguł asocjacyjnych na podstawie zamkniętych częstych zbiorów elementów: Mohammed J. Zaki. 2000. Generating non-redundant association rules. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '00). ACM, New York, NY, USA, 34-43

3. ***Wzorce sekwencyjne***

Sugerowana liczba osób w grupie: 1.

Implementacja jednego z algorytmów do odkrywanie częstych wzorców sekwencyjnych: GSP (uogólniony), SPADE, PrefixSpan, ERMiner.

2.1. R. Srikant, R. Agrawal , “Mining sequential patterns: Generalizations and performance improvements,” In Proceedings of International Conference on Extending Database Technology,
2.2. Zaki, M. J. ,“SPADE: An efficient algorithm for mining frequent sequences”, Machine learning, vol.42.no.1-2,pp.31-60,2001,
2.3. Pei, J., Han, J., Mortazavi-Asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M., “Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach”, IEEE Trans. Knowledge and Data Engineering,
2.4. Founier, P., Zida, S., Guenieche, T. and Tseng V., “ERMiner: Sequential Rule Mining using Equivalence Classes”, Advanced in intelligent data Analysis

4. Grupowanie

Implementacja jednego z algorytmów przedstawionych poniżej.

Sugerowana liczba osób w grupie: 1.

1. Guha S., Rastogi R., Shim K., ROCK: A robust clustering algorithm for categorical attributes, Proceedings of the International Conference on Data Engineering, Sydney 1999, pp. 512–521.
2. Karypis G., Han E., Kumar V., CHAMELEON: A hierarchical clustering algorithm using dynamic modeling, IEEE Computer: Special Issue on Data Analysis and Mining, vol. 32, no. 8, 1999, pp. 68–75.

3. Ng R. T. , Han J., Efficient and effective clustering methods for spatial data mining, Proc. 20th Int. Conf. on Very Large Data Bases, Morgan Kaufmann, Santiago 1994, pp. 144–155.
4. Estivill–Castro V., Lee I.: AMOEBA: hierarchical clustering based on spatial proximity using Delaunay diagram, In Proceedings of the 9th International Symposium on Spatial Data Handling, 2000
5. Estivill–Castro V., Lee I.: AUTOCLUST: automatic clustering via boundary extraction for mining massive point–data sets, Proceedings of the 5th International Conference on Geocomputation, 2000
6. Zhang T., Ramakrishnan R., Linvy M., BIRCH: an efficient data clustering method for very large databases, Proc. ACM SIGMOD Int. Conf. on Management of Data, ACM Press, 1996, pp. 103–114.
7. Kryszkiewicz M., Lasek P. (2010) A Neighborhood-Based Clustering by Means of the Triangle Inequality. In: Fyfe C., Tino P., Charles D., Garcia-Osorio C., Yin H. (eds) Intelligent Data Engineering and Automated Learning – IDEAL 2010. IDEAL 2010

5. Klasyfikacja

Sugerowana liczba osób w grupie: 2.

5.1. Implementacja i porównanie algorytmów klasyfikacji: Naïve Bayesian Classifier, Lazy Classification with Contrast Patterns, SPRINT.

- Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
- John C. Shafer, Rakesh Agrawal, and Manish Mehta. 1996. SPRINT: A Scalable Parallel Classifier for Data Mining. In Proceedings of the 22th International Conference on Very Large Data Bases (VLDB '96)