25-04-2022

Podsumowanie rozprawy

OPRACOWANIE EFEKTYWNYCH METOD EKSPLORACJI CZĘSTYCH WZORCÓW MIĘDZYTRANSAKCYJNYCH

Thanh Ngo Nguyen

Celem niniejszej pracy jest opracowanie skutecznych metod eksploracji częstych wzorców międzytransakcyjnych, eksploracji częstych zamkniętych wzorców międzytransakcyjnych oraz eksploracji częstych maksymalnych wzorców międzytransakcyjnych.

Pytania badawcze

Bazując na wadach istniejących metod i zaletach stosowania zwartych struktur, główne problemy, które należy rozwiązać, aby poprawić czas wydobycia i wykorzystanie pamięci, są następujące:

1. *Jak przycinać rzadkie wzory tak szybko, jak to możliwe*. Podczas procesu kopania obecne metody generują ogromną liczbę nieczęstych i zbędnych kandydatów. Jeśli przytniemy te wzorce w celu zmniejszenia przestrzeni wyszukiwania, czas wymagany do ich wygenerowania i pamięć do ich przechowywania ulegną skróceniu.
2. *Jak szybko określić wsparcie wzorców.* Używanie struktur danych do przechowywania wzorców wiąże się z dużą złożonością czasową i zajmuje ogromne ilości pamięci. Dlatego w oparciu o te struktury ustalenie *obsługi* wzorców wymaga ogromnej ilości pamięci i czasu wykonywania. W tej pracy badamy, jak wykorzystać kompaktowe struktury danych do szybkiego określenia *wsparcia* wzorców podczas procesu eksploracji w celu poprawy czasu eksploracji i wykorzystania pamięci.
3. *Jak ograniczyć skanowanie bazy danych.* Wielokrotne skanowanie dużej bazy danych, szczególnie w kontekście mega-transakcyjnej bazy danych FITPM, przekracza granice transakcji, a przestrzeń wyszukiwania rośnie wykładniczo, ponieważ zmieniają się niektóre parametry eksploracji(*maxSpan* i *minSup)*. Dlatego opracowanie wydajnych metod, które zapobiegają kosztownemu generowanie i testowanie kandydatów oraz powtarzaniu operacji skanowania bazy danych, jest niezbędnym i ważnym zadaniem.

Główne składki

Główne wkłady tej pracy są podzielone na trzy kategorie i zostaną pokrótce przedstawione w następujący sposób:

1. Zaproponowano dwie metody częstszego wydobywania wzorców między transakcjami (FITPM)
   - Pierwsza metoda, w której przyjmujemy strukturę *N-list* i modyfikujemy ją do *ITN-list,* aby była odpowiednia do wyszukiwania FITP w celu przechowywania informacji o kandydatach. Następnie proponowane są dwie strategie przycinania, aby na wczesnym etapie przycinać rzadkie wzorce oraz zmniejszyć przestrzeń wyszukiwania. Bazując na właściwościach *ITN-list* metoda może być wykorzystana do szybkiego określenia *obsługi* ITPs.
   - Druga metoda, w której używamy *diffset* zamiast *tidset*, do przechowywania informacji ITP. Ponadto oferujemy dwie techniki przycinania, aby usunąć rzadkie jednoelementowe wzorce (ang. 1-patterns), aby zmniejszyć przestrzeń wyszukiwania.
   - Zaletą zaproponowanych metod jest to, że wykorzystują one *ITN-list* do przechowywania informacji o kandydatach i szybkiego obliczania częstotliwości, oraz strategie przycinania są wykorzystywane efektywnie. Dlatego proponowane metody DITP-Miner i NL-ITP-Miner przewyższają dotychczasowe metody pod względem czasu pracy i wykorzystania pamięci.
2. Zaproponowano dwie metody częstych zamkniętych wzorców transakcyjnych (FCITPM)
   - W pierwszej metodzie tak jak w powyżej opisanej metodzie używamy również *ITN-list* i *diffset, o których* mowa powyżej. Proponujemy algorytm oparty na *ITN-list do wyszukiwania FCITP*, a także oferujemy dwie strategie przycinania w celu szybkiego określenia zamkniętych właściwości ITP. W oparciu o te dwie strategie proponowana metoda NCITP-Miner znacznie zmniejsza przestrzeń poszukiwań i przyspiesza proces wydobycia.
   - W drugiej metodzie proponujemy algorytm oparty na *diffset*, nazwany FCITP-Miner, do wydobywania FCITP. Algorytm wykorzystuje dwie strategie przycinania na poziomie jednego wzorca, aby zredukować przestrzeń poszukiwań i wydobywać FCITP w sposób przeszukiwania w głąb.

3. Zaproponowano trzy metody częstszego, maksymalnego wydobywania wzorców między transakcjami (FMITPM)
   - Zaproponowano trzy metody wydobywania FMITP używające *ITN-list*, *diffset*, *tidset*, nazwane odpowiednio nMITP-Miner, dMITP-Miner i tMITP-Miner,
   - Zaproponowano strategię przycinania zastosowaną w algorytmie nMITP-Miner w celu znacznego zmniejszenia przeszukiwanej przestrzeni. W większości przypadków nMITP-Miner działa szybciej niż algorytmy dMITP-Miner i tMITP-Miner pod względem czasu wykonywania i użycia pamięci.
4. Przeprowadzone eksperymenty służą weryfikacji proponowanych metod. Wyniki dla zbiorów danych z repozytorium FIMD (http://fimi.uantwerpen.be/data/) pokazują, że wszystkie proponowane metody są bardziej wydajne niż istniejące metody.

25-04-2022

Summary of the dissertation

DEVELOPING EFFICIENT METHODS FOR MINING FREQUENT INTER-TRANSACTION PATTERNS

Thanh Ngo Nguyen

The aim of this thesis is to develop efficient methods for mining frequent inter-transaction patterns, mining frequent closed inter-transaction patterns, and mining frequent maximal inter-transaction patterns.

Research questions

Based on the disadvantages of existing methods and the advantages of using the compact structures, the main problems that need to be solved to improve mining time and memory usage are as follows:

1. *How to prune infrequent patterns as soon as possible*. During the mining process, the current methods generate a huge number of infrequent and redundant candidates. If we prune these patterns to reduce the search space, the time required to generate them and the memory to store them will be reduced.
2. *How to quickly determine the support of patterns.* Using data structures to store patterns has high time complexities and takes up a huge amount of memory usage. Therefore, based on these structures, determining the *support* of patterns requires an enormous amount of memory usage and runtime. In this thesis, we study how to use the compact data structures to quickly determine the *support* of patterns during the mining process to improve the mining time and memory usage.
3. *How to reduce database scans*. Multiple scans of a large database, especially in the context of a mega-transaction database of FITPM, breaks the transaction boundaries, and the search space increases exponentially, as some of the mining parameters, *maxSpan* and *minSup*, change. Therefore, developing efficient methods that prevent the costly candidate-generation-and-test and repeated database scan operations is a necessary and significant task.

The main contributions

The main contributions of this thesis are classified into three categories and briefly presented as follows:

1. Proposed two methods for frequent inter-transaction pattern mining (FITPM)
   - We adopt the *N-list* structure and modify it to *ITN-list* to be suitable for mining FITPs to store candidates' information. Then, two pruning strategies are proposed to prune infrequent patterns early on to reduce the search space. Based on the *ITN*-list's properties, it can be used to quickly determine the *support* of ITPs.
   - We use *diffsets*, instead of *tidsets*, to store the information of ITPs. In addition, we offer two pruning techniques to remove infrequent 1-patterns to cut down the search space.
   - The advantages of these methods are that they take advantage of the *ITN-List* to store candidate information and quickly calculate frequency, and the pruning strategies are used efficiently. Therefore, the proposed methods, DITP-Miner and NL-ITP-Miner, outperform the current methods in terms of runtime and memory usage.
2. Proposed two methods for frequent closed inter-transaction pattern mining (FCITPM)
   - We also use *ITN-list* and *diffset,* mentioned above.
   - We propose an *ITN-list* based algorithm for mining FCITPs, and also offer two pruning strategies to qucikly determine the closed properties of ITPs. Based on these two strategies, the proposed method, NCITP-Miner, significantly reduces the search space and speeds up the mining process.
   - We propose a diffset-based algorithm, named FCITP-Miner, for mining FCITPs. The algorithm uses the two pruning strategies at the 1-pattern level to reduce the search space and mine FCITPs in the depth-first search manner.
3. Proposed three methods for frequent maximal inter-transaction pattern mining (FMITPM)
   - We propose three methods for mining FMITPs, named nMITP-Miner, dMITP-Miner, and tMITP-Miner, using *ITN-list*, *diffset*, and *tidset*, respectively.
   - We propose a pruning strategy applied on the nMITP-Miner algorithm to reduce the search space significantly. In most cases, the nMITP-Miner runs faster than the dMITP-Miner and tMITP-Miner algorithms in terms of runtime and memory usage.

4. Experiments are used to verify the proposed methods. Results for datasets from the FIMD Repository (http://fimi.uantwerpen.be/data/) show that all proposed methods are more efficient than existing methods.