Journal of educational equipment: Applied research, Volume 1, Issue 324 (November 2024)
ISSN 1859 - 0810

Tìm hiểu một số thuật toán khai phá tập mục lợi ích cao và ứng dụng

Nguyễn Nam Phương*, Vũ Anh Đức**

**ThS. Trường Cao Đẳng Yên Bái*

Received: 03/10/2024; Accepted: 14/10/2024; Published: 30/10/2024

Abstract: *High-utility set mining is an important problem in data mining, which considers the benefits of items (such as profits and interest rates) discovered from transactional databases that support the business of units. It is a method that uses techniques and algorithms in the field of data science to discover valuable patterns, rules or important information from available large data sets. This paper presents some high-utility set mining algorithms and its applications.*

Keywords: *Data Mining*

1. Đặt vấn đề

Chúng ta đang sống trong trong thời đại của công nghệ thông tin. Ngoài việc phát triển của Internet, các kỹ thuật tiên tiến về lưu trữ dữ liệu những khối dữ liệu khổng lồ phát sinh từ các doanh nghiệp, các tổ chức khoa học và chính phủ cũng phát triển mạnh mẽ. Vấn đề làm sao khai thác được thông tin có giá trị trong kho dữ liệu khổng lồ đó. Đây là phương pháp sử dụng các kỹ thuật và thuật toán trong lĩnh vực khoa học dữ liệu để khám phá ra các mẫu, quy luật hoặc thông tin quan trọng có giá trị từ bộ dữ liệu lớn có sẵn. Bài báo này trình bầy một số thuật toán khai thác tập lợi ích cao và ứng dụng của nó.

2. Nội dung nghiên cứu

2.1. Khái niệm khai phá dữ liệu tập lợi ích cao

Khai phá dữ liệu (KTDL) là một quá trình trích xuất tri thức từ lượng lớn dữ liệu; là tiến trình trích lọc, sản sinh những tri thức hoặc các mẫu tiềm ẩn, chưa biết nhưng hữu ích từ các cơ sở dữ liệu lớn; là tiến trình khái quát các sự kiện rời rạc trong dữ liệu thành các tri thức mang tính khái quát, tính quy luật hỗ trợ tích cực cho các tiến trình ra quyết định. Khai phá dữ liệu có một số ứng dụng sau:

- KPDL được sử dụng để phân tích dữ liệu, hỗ trợ ra quyết định.

- Trong sinh học: nó dùng để tìm kiếm, so sánh các hệ gen và thông tin di truyền, tìm mối liên hệ giữa các hệ gen và chẩn đoán một số bệnh di truyền.

- Trong y học: KPDL giúp tìm ra mối liên hệ giữa các triệu chứng, chẩn đoán bệnh.

- Tài chính và thị trường chứng khoán: KPDL dùng để phân tích tình hình tài chính, phân tích đầu tư, phân tích cổ phiếu.

- Khai thác dữ liệu web.

- Trong thông tin kỹ thuật: KPDL dùng để phân tích các sai hỏng, điều khiển và lập lịch trình.

Trong đó khai phá tập Tập lợi ích cao (TLIC) là một vấn đề quan trọng trong khai thác dữ liệu, xem xét các lợi ích của các mục (chẳng hạn như lợi nhuận và lãi suất) được khám phá từ cơ sở dữ liệu (CSDL) giao dịch hỗ trợ cho việc kinh doanh của các đơn vị.

Hiện nay, trong lĩnh vực kinh doanh việc tính toán doanh số và tối ưu hóa lợi nhuận bán hàng là công việc cực kỳ quan trọng, nó ảnh hưởng trực tiếp đến doanh thu và chiến lược bán hàng của các công ty, siêu thị hay các đơn vị bán lẻ. Đặc biệt, với số lượng hàng hóa lớn, giá cả khác nhau, nên việc tính toán lợi nhuận tối ưu bán hàng càng quan trọng. Với số lượng giao tác mỗi giờ có thể lên đến hàng chục nghìn giao tác, việc tính toán xem mặt hàng nào đem lại doanh số cao, mặt hàng nào kinh doanh không hiệu quả dù bán với số lượng lớn càng trở nên khó khăn do dữ liệu quá lớn, liên tục. Kho dữ liệu điển hình trong những doanh nghiệp cho phép người dùng hỏi và trả lời những câu hỏi như "Doanh số bán ra là bao nhiêu tính theo khu vực, theo nhân viên bán hàng". Trong khi đó, KTDL cho phép người ra quyết định kinh doanh hỏi và trả lời cho những câu hỏi như là "Ai là khách hàng chính yếu của công ty đối với một mặt hàng cụ thể?" hoặc "Dòng sản phẩm nào sẽ bán trong khu vực này và ai sẽ mua chúng, dựa vào việc bán những sản phẩm tương tự ở ở khu vực đó?".

2.2. Bài toán khai phá luật kết hợp

Cho cơ sở dữ liệu giao tác DB, ngưỡng độ hỗ trợ tối thiểu minsup và ngưỡng độ tin cậy tối thiểu minconf.

Yêu cầu: Tìm tất cả các luật kết hợp $X \rightarrow Y$ trên cơ sở dữ liệu DB sao cho $sup(X \rightarrow Y) \geq minsup$ và $conf(X \rightarrow Y) \geq minconf$.

Bài toán khai phá luật kết hợp này được goi là bài toán cơ bản hay bài toán nhị phân, vì ở đây, giá trị của mục dữ liệu trong cơ sở dữ liệu là 0 hoặc 1 (xuất hiện hay không xuất hiện).

2.3. Một số hướng mở rộng của bài toán khai phá tập mục phổ biến

Một hướng mở rộng bài toán có nhiều ứng dụng là quan tâm đến cấu trúc dữ liệu và mức độ quan trọng khác nhau của các mục dữ liệu, các thuộc tính trong cơ sở dữ liệu. Một số mô hình mở rộng bài toán đã được nghiên cứu là:

- Quan tâm đến mức độ quan trọng khác nhau của các mục dữ liệu bằng cách gán cho mỗi mục một giá trị được gọi là trọng số.

- Quan tâm đến các kiểu thuộc tính khác nhau trong cơ sở dữ liệu như nhị phân, đa phân, định lượng.

2.4. Một số thuật toán hiệu quả khai phá tập mục lợi ích cao

Bài toán: Vấn đề khai thác tập phổ biến cũng như khai thác itemset có ích đều gặp phải là số lượng các item quá nhiều trong cơ sở dữ liệu. Giả sử có 105 items, thì hơn 109 2-itemsets ứng viên có thể được sinh ra, và số lượng khổng lồ k-itemsets được sinh ra khi k lớn.

Vấn đề cần giải quyết là các thuật toán tìm itemset có ích cần phải thu gọn không gian xử lý. Thuật toán đầu tiên đề cập đến phương pháp khai thác itemset có ích do Hong Yao và Howard Hamilton đề xuất có hiệu suất rất thấp. Họ dùng hai công thức xác định chặn trên của itemset.

a) Thuật toán Two-Phase

Định nghĩa 1(lợi ích giao dịch): Lợi ích của một giao dịch Tq, ký kiệu TU(Tq) là tổng giá trị lợi ích của tất cả các item trong giao dịch Tq, tức là:

$$TU(T_q) = \sum_{S \in T_q} u(S, T_q)$$

TU(T1) = 4.5 + 1.38 = 58, TU(T2) = 16, TU(T3) = 68, TU(T4) = 20, TU(T5) = 11, TU(T6) = 106, TU(T7) = 18, TU(T8) = 144, TU(T9) = 110, TU(T10) = 34.

Định nghĩa 2 (Lợi ích trọng số giao dịch): Lợi ích trọng số giao dịch của một tập item X, ký hiệu là TWU(X), là tổng giá trị lợi ích giao dịch của tất cả các giao dịch có chứa X:

$$TWU(X) = \sum_{X \subseteq T_q \in T} TU(T_q)$$

b) Thuật toán Two-Phase

Quét CSDL theo chiều ngang, đầu tiên tính TWU của tất cả tập 1-itemset, nếu itemset nào có TWU lớn hơn hoặc bằng ngưỡng minutil thì được chọn và có khả năng là tập mục có lợi ích cao, những itemset nào có TWU nhỏ hơn ngưỡng minutil thì bị loại và chắc chắc các itemset đó là tập mục không có lợi ích cao.

Tính trọng số lợi ích cao của các ứng viên, ta có: TWU(ABCD) = 144. Và duyệt CSDL, những giao dịch có trọng số lợi ích cao lớn hơn hoặc bằng 150 sẽ được chọn. Kết quả không có itemset.

c) Thuật toán FHM

Input: D: a transaction database, minutil: a user-specified threshold

Output: the set of high-utility itemsets

1. Scan D to calculate the TWU of single items;
2. I* ← each item I such that twu(i) ≥ minutil;
3. Let be the total order of TWU ascending values on I*;
4. Scan D to built the utility-list of each item i ∈ I* and built the EUCS structure
5. Search (∅, I*, minutil, EUCS);

2.5. Chương trình thực nghiệm ứng dụng

a) Bài toán: Phát hiện nhóm mặt hàng mang lại lợi nhuận cao trên tập dữ liệu bán hàng của siêu thị Yên Bái.

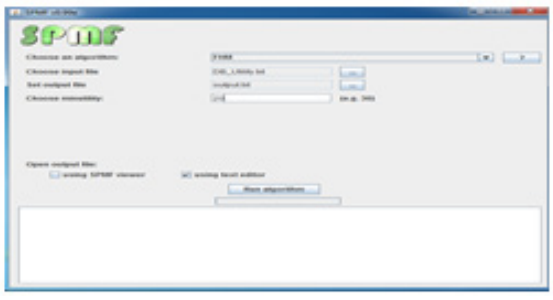
Khai thác tập mục lợi ích cao có nhiều ứng dụng chẳng hạn như phát hiện các nhóm mặt hàng trong các giao dịch của một cửa hàng tạo ra nhiều lợi nhuận nhất. Một cơ sở dữ liệu chứa thông tin tiện ích là một cơ sở dữ liệu, nơi các mục có thể có số lượng và đơn giá. Yêu cầu của bài toán là phát hiện các nhóm mặt hàng trong các giao dịch bán hàng của một siêu thị Yên Bái

b) Kết quả thực nghiệm

- Xây dựng chương trình

Chương trình được cài đặt bằng ngôn ngữ Java, hệ điều hành Windows 10, CPU i5- 4790 2.6GHz, RAM 4GB.

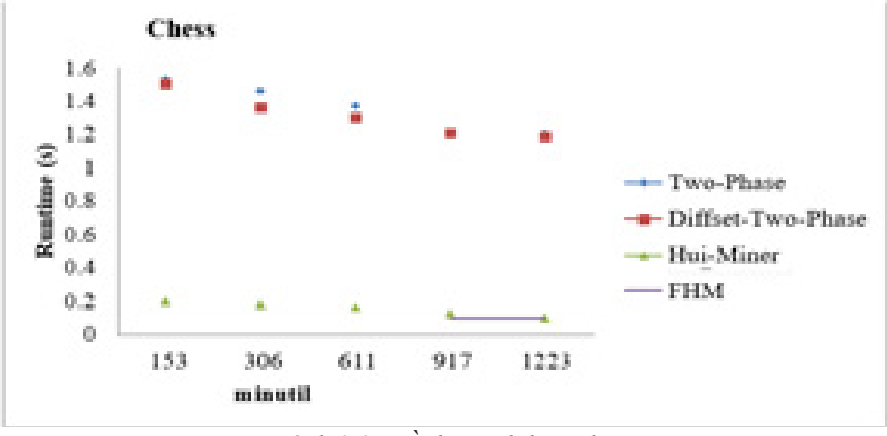
Giao diện chương trình (hình 2.1)



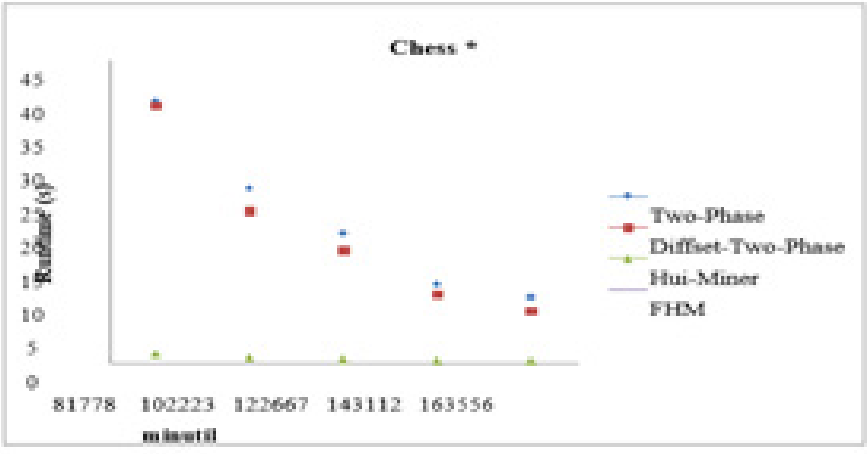
Hình 2.1: Giao diện chương trình

- Kết quả

Với bản thực nghiệm trên, chúng ta có thể xem sự khác biệt kết quả thực nghiệm thông qua đồ thị. So sánh giữa nhóm (Two-phase, Diffset-Two-Phase) và (Hui-Miner, FHM): Hình 2.2 và 2.3



Hình 2.2. Đồ thị minh họa chess



*Hình 2.3. Đồ thị minh họa cho dữ liệu chess**

Kết quả thực nghiệm cho thấy các thuật toán Two-Phase có kết quả chậm hơn nhiều so với các thuật toán sử dụng utility-list. Trong họ thuật toán Two-Phase thì thuật toán Diffset-Two-Phase vẫn là tốt nhất. Và trong họ thuật toán có dung utility-list thì thuật toán FHM đúng là tốt hơn thuật toán HUI-Miner trong hầu hết mọi trường hợp.

3. Kết luận

Việc nghiên cứu là lâu dài và khó khăn, trong khuôn khổ thời gian thực hiện luận văn, bao gồm quá trình tìm hiểu và đọc tài liệu, chạy các thực nghiệm cũng như kiểm tra tính đúng đắn của các thuật toán đã được cài đặt trong công cụ SPMF. Khu vực nghiên cứu của luận văn chỉ mới nằm trong các giới hạn sau:

- Tác giả tìm hiểu các thuật toán khai thác tập có ích cao trên dữ liệu tĩnh (dữ liệu không có biến động). Trong thực tế dữ liệu tĩnh chỉ phù hợp trong phân tích, rút kinh nghiệm trong một quãng thời gian nào đó và không có tính tương tác kịp thời. Dữ liệu động phù hợp trong các bài toán cần có tính tương tác cao. Với mỗi loại dữ liệu có tập các thuật toán phù hợp để khai thác.

- Dữ liệu nghiên cứu được lấy từ nguồn dữ liệu nghiên cứu chuẩn (chưa thử nghiệm trên dữ liệu thực). Những dữ liệu trong luận văn chỉ là dữ liệu giả lập, việc này dễ hơn rất nhiều so với thao tác trên dữ liệu thực tế vì dữ liệu thực tế còn phải qua bước tiền xử lý dữ liệu.

- Việc đánh giá chỉ mới đánh giá dựa trên tốc độ xử lý dữ liệu của các thuật toán (chưa đánh giá tính có ích thực sự so với ý kiến thực của khách hàng). Trong thực tế cần đánh giá độ trùng khớp kết quả thu được với tri thức của khách hàng, nếu độ trùng khớp cao thì kết quả của hệ thống mới có giá trị sử dụng, và các thông tin bất thường mới thực sự thú vị với mong muốn khai thác dữ liệu.

Tài liệu tham khảo

[1] Philippe Fournier-Viger, Cheng-Wei Wu, Souleymane Zida, Vincent S.Tseng (2014) *FHM: Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning*. Proc. 21st International Symposium on Methodologies for Intelligent Systems (ISMIS 2014), Springer, LNAI, pp. 83-92.

[2] Ying Liu, Wei-keng Liao, and Alok Choudhary: *A two-phase algorithm for fast discovery of high utility itemsets*. In: Proc. PAKDD 2005, pp. 689-695 (2005)