Khai phá mẫu phổ biến, luật kết hợp và thước đo tương quan

Vũ Mỹ Hạnh

Trường Đại học Công nghệ
Luận văn Thạc sĩ ngành: Hệ thống thông tin; Mã số: 60 48 05
Người hướng dẫn: TS. Nguyễn Công Điều
Năm bảo vệ: 2011

Abstract: Tổng quan về khai phá mẫu phổ biến, luật kết hợp và các thước đo tương quan. Một số phương pháp cơ bản và mở rộng trong khai phá luật kết hợp. Giới thiệu công cụ khai phá dữ liệu Weka và mô phỏng.

Keywords: Khai phá dữ liệu; Công nghệ thông tin; Mẫu phổ biến; Luật kết hợp

Content

Trong những năm gần đây, cùng với sự phát triển vượt bậc của khoa học và công nghệ, khả năng thu thập và lưu trữ dữ liệu được nâng cao đáng kể, điều này đồng nghĩa với việc một lượng lớn dữ liệu được lưu trữ trên các thiết bị nhớ tăng lên không ngừng. Cũng chính vì thế mà từ đây con người gặp phải một bất cập lớn trong việc phân tích một kho dữ liệu khổng lồ để rút ra được các quyết định hữu ích, ứng dụng trong hầu hết các lĩnh vực khoa học, kinh tế, xã hội.

Khai phá dữ liệu – Data mining là một lĩnh vực khoa học liên ngành, nhằm tự động hóa quá trình khai phá thông tin, tri thức hữu ích tiềm ẩn trong cơ sở dữ liệu của các tổ chức, doanh nghiệp,...Đây là lĩnh vực khoa học tiềm năng, mang lại nhiều lợi ích thiết thực, đồng thời thể hiện ưu thế vượt trội hơn hẳn so với các công cụ phân tích dữ liệu truyền thống.

Khai phá Mẫu phổ biến – Frequent pattern mining (hay còn gọi là "*Mẫu thường xuyên* ") đóng vai trò thiết yếu trong khai phá luật kết hợp, mối tương quan, và mối quan hệ thú vị khác nhau trong dữ liệu. Hơn nữa, nó giúp phân lớp, phân cụm dữ liệu, và hỗ trợ khá tốt các nhiệm vụ khai phá dữ liệu. Do vậy, khai phá mẫu phổ biến đã trở thành nhiệm vụ khai phá dữ liệu quan trọng và là một chủ đề cần khai phá và tìm kiếm dữ liệu [1].

Khai phá luật kết hợp - Accessociation rule mining là một kỹ thuật quan trọng của khai phá dữ liệu lần đầu tiên được Rakesh Agrawal, Tomas Imielinski, Arun Swami đề xuất năm 1993. Những nghiên cứu về luật kết hợp gần đây tập trung vào việc xây dựng các thuật toán khai phá luật kết hợp theo hai hướng là cải tiến đưa thuật toán mới và cải tiến hiệu quả của thuật toán cũ.

Trong luận văn này tập trung trình bầy những khái niệm tổng quan về khai phá dữ liệu, mẫu phổ biến, luật kết hợp. Từ đó tìm hiểu các phương pháp khai phá tập mục phổ biến sinh ứng viên đối với khai phá khai phá luật kết hợp nhị phân. Đồng thời, dựa trên phân tích về những bất cập của phương pháp này, luận văn xem xét một số phương phương pháp cải tiến, khai phá tập mục không cần sinh ứng viên, cùng với những phân tích đánh giá chi tiết về ưu điểm và nhược điểm của phương pháp này. Bên cạnh đó, luận văn cũng đề cập đến một hướng tiếp cận khác trong việc khai phá luật kết hợp đó là khai phá luật kết hợp định lượng. Đây là một trong những hướng phát triển để hoàn thiện những khía cạnh còn thiếu sót của

khai phá luật kết hợp nhị phân. Hơn nữa, các thước đo tương quan cũng được trình bầy để giúp đánh giá một luật được đưa ra có thực sự mạnh và đáng quan tâm hay không. Cuối cùng, tìm hiểu về công cụ Weka và sử dụng công cụ này để mô phỏng các phương pháp khai phá đã đề cập.

Luận văn bao gồm ba chương: Chương 1: Tổng quan về khai phá mẫu phổ biến, luật kết hợp và các thước đo tương quan. Chương 2: Một số phương pháp cơ bản và mở rộng trong khai phá luật kết hợp. Chương 3: Giới thiệu công cụ khai phá dữ liệu Weka và mô phỏng.

References

Tiếng Việt

[1]. Hà Quang Thụy (Chủ biên), Phan Xuân Hiếu, Đoàn Sơn, Nguyễn Trí Thành, Nguyễn Thu Trang, Nguyễn Cẩm Tú (2010), *Giáo trình Khai phá dữ liệu Web*, NXB Giáo dục Việt Nam.

Tiếng Anh

[2]. Jiawei Han (2006), *Data mining – Concept and Techniques* – 2nd edition.

[3]. Fayyad, Piatetsky-Shapiro, Smyth (1996). *From Data Mining to Knowledge Discovery: An Discovery and Data Mining*, AAAI Press/ The MIT Press, Menlo Park, CA, 1-34.

[4]. J.Han, J.Pei, Y.Yin, and R.Mao (2004), *Mining Frequent Patterns without Candidate Generation: A Frequent-pattern Tree Approach. Data Mining and Knowledge Discovery*.

[5]. Y.Aumann, and Y.Lindell (1999), *A statistical theory for quantitative association rules*. Proc. Of the 5th KDD.

[6]. R.Srikant, and R.Agrawal (1996), *Mining Quantitative Associatin Rules in Large Rational Tables*.

[7]. Rakesh Agrawal and Ramakrishnan Srikant (September 1994). *Fast Algorithms for Mining Association Rules*. In Pro. Of the 20th Int'l Conference on Very Large Databases, Santiago, Chile.

[8]. R. Agrawal, T.Imielinski, and A.N.Swami (1993), *Mining association rules between sets of items in large databases.* In International Conference on 1993 ACM SIGMOD Inter national Conference on Management of Data, Washington, D.C.

[9]. B. Goethals (2003), *Survey on Frequent Parttern Mining*. Technical Report, Helsinki, Institute for Information Technology.