

MỘT THUẬT TOÁN TÌM TẬP THƯỜNG XUYỀN TRÊN CƠ SỞ DỮ LIỆU GIAO TÁC CÓ TRỌNG SỐ

A NEW METHOD TO MINE FREQUENT ITEMSETS ON TRANSACTION WEIGHTED DATA

Tác giả: Nguyễn Hữu Trọng; Lê Đức An; Trần Xuân Việt; Nguyễn Anh Hào

Tóm tắt bằng tiếng Việt:

Tìm tất cả các tập mục dữ liệu thường xuyên là công việc cơ bản trong khai phá dữ liệu. Trong hơn 20 năm từ ngày bởi Agrawal và các cộng sự đưa ra thuật toán AIS, các nhà nghiên cứu đã đề xuất nhiều thuật toán cải tiến về tốc độ để phát hiện nhanh những tập mục dữ liệu thường xuyên. Những thuật toán này, một hướng tập trung vào cải tiến cách duyệt qua tập dữ liệu và cách tính độ hỗ trợ của từng tập mục dữ liệu ứng viên, một hướng khác, tập trung vào việc rút gọn kích thước và phân nhỏ cơ sở dữ liệu được xử lý. Trong báo cáo này, chúng tôi đề xuất thuật toán CABOWD (clustering algorithm based on weighted data), nhằm rút ngắn thời gian tìm các tập thường xuyên theo hướng phân nhỏ và rút gọn kích thước dữ liệu.

Từ khóa: Cơ sở dữ liệu giao tác; luật kết hợp; tập thường xuyên; phân cụm; trọng số

Tóm tắt bằng tiếng Anh:

Finding all of the frequent itemsets is a basic work in data mining. For more than 20 years from the date of Agrawal and colleagues provide AIS algorithm, the researchers have proposed many improved algorithms for detecting fast pace of the frequent itemsets. These algorithms, an improved focus on navigating the data sets and how to calculate the support of each candidate item set, a different direction, focus on the reduced size and split of data into several clusters. In this paper, we propose the clustering algorithm based on weighted data (CABOWD) in order to shorten time by way of splitting data and reducing data sizes.

Key words: database transactions; association rules; frequent practice; clustering; weight