Nghiên cứu các kỹ thuật phân cụm dữ liệu và ứng dụng

Nguyễn Thị Huế

Trường Đại học Công nghệ
Luận văn ThS. ngành: Hệ thống thông tin; Mã số: 60 48 05
Người hướng dẫn: GS.TS. Vũ Đức Thi
Năm bảo vệ: 2011

Abstract. Tổng quan về khai phá tri thức, khai phá dữ liệu; qui trình khai phá tri thức, khai phá dữ liệu … Trình bày tổng quan về phân cụm dữ liệu, một số phương pháp phân cụm dữ liệu dữ liệu phổ biến như phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dựa trên lưới; trình bày một số giải thuật điển hình của mỗi phương pháp phân cụm; … Ứng dụng, triển khai bài toán với giải thuật DBSCAN.

Keywords. Hệ thống thông tin; Dữ liệu; Công nghệ thông tin; Phân cụm dữ liệu

Content

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin trong các lĩnh vực của đời sống, kinh tế, xã hội trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được các cơ quan thu thập và lưu trữ ngày một tích lũy nhiều lên. Hơn nữa, các công nghệ lưu trữ và phục hồi dữ liệu phát triển một cách nhanh chóng vì thế cơ sở dữ liệu ở các cơ quan, doanh nghiệp, đơn vị ngày càng nhiều thông tin tiềm ẩn phong phú và đa dạng. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật khai phá tri thức và khai phá dữ liệu (*KDD - Knowledge Discovery and Data Mining*). Khai phá tri thức trong cơ sở dữ liệu có thể được coi như quá trình tìm tri thức có ích, cần thiết, tiềm ẩn và chưa được biết trước trong cơ sở dữ liệu lớn (*discovery of interesting, implicit, and previously unknown knowledge from large databases*)[5]

Kỹ thuật khai phá tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau ở các nước trên thế giới, tại Việt Nam kỹ thuật này tương đối còn mới mẻ tuy nhiên cũng đang được nghiên cứu và dần đưa vào ứng dụng trong những năm gần đây. Những vấn đề được quan tâm là phân lớp nhận dạng mẫu, luật kết hợp, phân cụm dữ liệu, phần tử dị biệt,…

Phân cụm cơ sở dữ liệu là một trong những phương pháp quan trọng trong quá trình tìm kiếm tri thức. Phân cụm là phương pháp học từ quan sát (*learning from obversation*) hay còn gọi là học không thầy (*unupervised learning or automatic classfication*) trong trí tuệ nhân tạo. Phân cụm đặc biệt hiệu quả khi ta không biết về thông tin của các cụm, hoặc khi ta

quan tâm tới những thuộc tính của cụm mà chưa biết hoặc biết rất ít về những thông tin đó. Phân cụm được coi như một công cụ độc lập để xem xét phân bố dữ liệu, làm bước tiền xử lý cho các thuật toán khác. Việc phân cụm dữ liệu có rất nhiều ứng dụng như trong tiếp thị, sử dụng đất, bảo hiểm, hoạch định thành phố … Hiện nay, phân cụm dữ liệu là một hướng được nghiên cứu rất nhiều trong Tin học. Chính vì lý do đó mà em chọn đề tài "*Nghiên cứu các kỹ thuật phân cụm dữ liệu và Ứng dụng*" là hướng nghiên cứu chính cho luận văn của mình.

Nội dung chính của luận văn được trình bày trong 3 chương:

Chương 1: Tổng quan về khai phá tri thức và khai phá dữ liệu. Trong chương này trình bày tổng quan về khai phá tri thức, khai phá dữ liệu; qui trình khai phá tri thức, khai phá dữ liệu; …

Chương 2: Phân cụm và các kỹ thuật phân cụm. Trong chương này trình bày tổng quan về phân cụm dữ liệu, một số phương pháp phân cụm dữ liệu dữ liệu phổ biến như phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dựa trên lưới; trình bày một số giải thuật điển hình của mỗi phương pháp phân cụm; …

Chương 3: Ứng dụng, triển khai bài toán với giải thuật DBSCAN

Phần kết luận trình bày tóm tắt về các nội dung thực hiện trong luận văn, đồng thời đưa ra các vấn đề nghiên cứu tiếp cho tương lai. Phần phụ lục trình bày một số modul chương trình cài đặt bằng thuật toán DBSCAN.

References

Tiếng Việt

[1]. Vũ Đức Thi, *Cơ sở dữ liệu – Kiến thức và thực hành*, Nhà xuất bản Thống kê, 1997

[2]. Nguyễn Xuân My, Hồ Sĩ Đàm, Trần Đỗ Hùng, Lê Sĩ Quang, *Một số vấn đề về chọn lọc trong môn tin học*, Nhà xuất bản Giáo dục, 2002, Trang 73- 108

[3]. Phan Đình Diệu, *Tri thức là gì*? Đại học Quốc gia Hà Nội.

Tiếng Anh

[4]. Han J. and Kamber M., *Data Mining: Concepts and Techniques*, Morgan Kaufman, Academic Press. 2001.

[5]. Burosch G., Demetrovics J., Katona G. O. H. (1987), *The poset of closures as a model of changing databases*, Oder 4, pp. 127-142.

[6]. J.R, QUINLAN, *Machine Learning 1*, 81-106, 1986, © 1986 Kluwer Academic Publishers, Boston - Manufactured in The Netherlands.

[7]. H. Huang, X. Wu, and R.Relue (2002), *Association analysis with one scan of databases.* In IEEE International Conference on Data Mining, pages 629-836, Decenber.

[8]. Utgoff P.E, *Article: Incremental induction of Decision Trees*, Univerity of Massacuhsetts, 1989.

[9].Tutorial: *Decision Tree*: ID3, Monhash University, 2003, http://www.cs.bham.ac.uk/resources/courses/ai-intro/docs/dt/

[10]. Agrawal, Rakesh, Johannes Gehrke, Dimitrios Gunopulos and Prahhakar Raghavan. *Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications*. Proceedings of the 1998 ACM-SIGMOD International Conference on Management of Data, Seattle, Washington, June 1998.

[11]. Ester, Martin, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu.(1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland, Oregon, August 1996.

[12]. Guha, Sudipto, Rajeev Rastogi and Kyuseok Shim. (1998). *CURE: An Efficient Clustering Algorithm for Large Databases. Proceedings of ACM SIGMOD-International Conference on Management of Data*. New York, NY 1998. pp 73—84. (CURE)

[13]. Hinneburg, Alexander and Daniel A Keim. (1998). *An Efficient Approach to Clustering in Large Multimedia Databases with Noise*. Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, (KDD98). New York, August 1998. pp. 58—65.

[14]. Pavel Berkhin, *Survey of Clustering Data Mining Techniques*. Accrue Software, Inc., San Jose.

[15] Ng, Raymond T. and Jiawei Han. *Efficient and Effective Clustering Methods for Spatial Data Mining*. Proceedings of the 20th Very Large Databases-Conference (VLDB 94), Santiago, Chile. pp 144-155. (CLARAN)

[16] Wang, Wei, Jiong Yang, and Richard Muntz, *STING: A Statistical Information Grid Approach to Spatial Data Mining*. Proceedings of the 23rd Very Large Databases Conference (VLDB 1997), Athens, Greece, 1997.

[17] Zhang, Tian, Raghu Ramakrishnan, and Miron Chinh hay Livny. (1996). *BIRCH: An Efficient Data Clustering Method for Very Large Databases*. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Canada, pp. 103-114, 1996.