

TÓM TẮT DỮ LIỆU QUAN HỆ SỬ DỤNG PHƯƠNG PHÁP PHÂN CỤM BÁN GIÁM SÁT DỰA TRÊN GIẢI THUẬT DI TRUYỀN

Học viên : Đỗ Thị Hòa
chuyên ngành : Công nghệ phần mềm
Ngành : CNTT
Email : dtkhkhtn@gmail.com

Giáo viên hướng dẫn: Hoàng Xuân Huân
Cơ quan công tác: Khoa CNTT – Đại học Công nghệ -
ĐHQG Hà nội
Email: huanhx@vnu.edu.vn

Từ khóa : Khai phá dữ liệu, thuật toán phân cụm dữ liệu, giải thuật di truyền, cơ sở dữ liệu quan hệ, Kmeans, tóm tắt dữ liệu quan hệ, ứng dụng khai phá dữ liệu

Luận văn tập trung tìm hiểu các phương pháp phân cụm dữ liệu, cơ sở dữ liệu quan hệ, giải thuật di truyền. Qua đó nghiên cứu cách áp dụng để tóm tắt dữ liệu quan hệ trên các tập dữ liệu tài chính, viên gan và đột biến

Nội dung của bản luận văn gồm có phần mở đầu, bốn chương và phần kết luận

toán phân cụm bán giám sát dựa trên giải thuật di truyền. Kết quả thực hiện chương trình là khả quan.

V.KẾT LUẬN

Trình bày tổng hợp các kết quả thực hiện luận văn và phương hướng nghiên cứu tiếp theo về các nội dung của luận văn

TÀI LIỆU THAM KHẢO

- [1] Basu, B., A. Banerjee and R. Mooney, 2002. Semi-supervised clustering by seeding. Proceedings of the 19th International Conference on Machine
- [2] Davies, D.L. and D.W. Bouldin, 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intel., PAMI, 1: 24-227. DOI: 10.1109/TPAMI.1979.4766909
- [3] Kirsten, M. and S. Wrobel, 2000. Extending K-means clustering to first-order representations. Proceeding of the 10th International Conference on Inductive Logic Programming, July 24-27, Springer-Verlag, London, UK., pp: 112-129.
- [4] Rayner, A. and K. Dimitar, 2007. Clustering approach to generalized pattern identification based on multi-instanced objects with DARA. Proceeding of the Communications of the 11th East-European Conference on Advances in Databases and Information Systems, Sept 2007, Technical University of Varna, pp: 1-12
- [5] Rayner Alfred-Journal of Computer Science 6(7):775-784,2010. Summarizing Relational Data Using Semi-Supervised Genetic Algorithm-Based Clustering Techniques
- [6] Salton, G. and M. Michael, 1984. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, USA., ISBN: 0070544840
- [7] Srinivasan, A., S. Muggleton, M.J.E. Sternberg and R.D. King, 1996. Theories for mutagenicity: Study in first-order and feature-based induction. Artif. Intell., 85: 277-299. DOI: 10.1016/0004-3702(95)00122-0
- [8] Witten, I.H. and E. Frank, 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 1st Edn., Morgan Kaufmann, ISBN: 10: 1558605525, pp: 371

I.TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU

Luận văn trình bày khái niệm phân cụm, các phương pháp phân cụm điển hình, xem xét các điểm mạnh, điểm yếu của từng phương pháp này.

II.GIẢI THUẬT DI TRUYỀN

Luận văn trình bày về giải thuật di truyền với các cách biểu diễn dữ liệu, cách xây dựng một giải thuật di truyền và mô tả các phép toán thực hiện trên đó. Tiếp theo là phân tích ứng dụng của giải thuật di truyền trong bài toán phân cụm.

III.TÓM TẮT DỮ LIỆU QUAN HỆ SỬ DỤNG PHƯƠNG PHÁP PHÂN CỤM BÁN GIÁM SÁT DỰA TRÊN GIẢI THUẬT DI TRUYỀN

Chương này đi sâu phân tích khái niệm, cấu trúc quan hệ các bảng trong cơ sở dữ liệu, cách chuyển đổi dữ liệu. Thông qua đó luận văn trình bày thuật toán phân cụm bán giám sát dựa trên giải thuật di truyền để tóm tắt dữ liệu.

IV.KẾT QUẢ THỬ NGHIỆM THUẬT TOÁN

Chương này trình bày các kết quả thực nghiệm về phương pháp tóm tắt dữ liệu quan hệ sử dụng thuật

SUMMARIZING RELATIONAL DATA USING SEMI-SUPERVISED GENETIC ALGORITHM – BASED CLUSTERING TECHNIQUES

Student name: Do Thi Hoa
minor, major, Department: Department of Information
Systems, Information Technology
email: dtkhkhtn@gmail.com

Professor name: Prof . PhD Hoang Xuan Huan
Affiliation: University Of Engineering and Technology,
VietNam National University, Ha Noi
Email: huanhx@vnu.edu.vn

Keywords : Data mining, data clustering algorithms, genetic algorithms, relational databases, Kmeans, summarizing relational data, data mining applications

The thesis concentrates in studying the methods of data clustering, relational data, genetic algorithm. Therefore applying to study relational data summary of data on the financial, mutagenesis and hepatitis databases

The structure of it is following:

I.OVERVIEW OF DATA CLUSTERING

The thesis concentrates definition of clustering, some typical clustering methods, and find out the advantages and disadvantages of those methods.

II.GENETIC ALGORITHMS

The thesis concentrates genetic algorithms with various data's representation, how to build a genetic algorithm and describe the operations performed on it. Then analysing of applications of genetic algorithms in clustering problems.

III.SUMMARIZING RELATIONAL DATA USING SEMI-SUPERVISED GENETIC ALGORITHM-BASED CLUSTERING TECHNIQUES

This chapter analyzes in depth concepts, relational tables in the database, the data conversion. So that thesis presents the summarizing relational data using semi-supervised genetic algorithm-based clustering techniques.

IV.RESULTS OF EXPERIMENT

This chapter presents the results of experiment about summarizing relational data using semi-supervised genetic algorithm-based clustering techniques. Results of experiment is satisfactory.

V.CONCLUSION

Review about results of the thesis and methods for further research on the contents of the thesis

REFERENCES

- [1] Basu, B., A. Banerjee and R. Mooney, 2002. Semi-supervised clustering by seeding. Proceedings of the 19th International Conference on Machine
- [2] Davies, D.L. and D.W. Bouldin, 1979. A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intel., PAMI, 1: 24-227. DOI: 10.1109/TPAMI.1979.4766909
- [3] Kirsten, M. and S. Wrobel, 2000. Extending K-means clustering to first-order representations. Proceeding of the 10th International Conference on Inductive Logic Programming, July 24-27, Springer-Verlag, London, UK., pp: 112-129.
- [4] Rayner, A. and K. Dimitar, 2007. Clustering approach to generalized pattern identification based on multi-instanced objects with DARA. Proceeding of the Communications of the 11th East-European Conference on Advances in Databases and Information Systems, Sept 2007, Technical University of Varna, pp: 1-12
- [5] Rayner Alfred-Journal of Computer Science 6(7):775-784,2010. Summarizing Relational Data Using Semi-Supervised Genetic Algorithm-Based Clustering Techniques
- [6] Salton, G. and M. Michael, 1984. Introduction to Modern Information Retrieval. McGraw-Hill, Inc., New York, USA., ISBN: 0070544840
- [7] Srinivasan, A., S. Muggleton, M.J.E. Sternberg and R.D. King, 1996. Theories for mutagenicity: Study in first-order and feature-based induction. Artif. Intell., 85: 277-299. DOI: 10.1016/0004-3702(95)00122-0
- [8] Witten, I.H. and E. Frank, 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. 1st Edn., Morgan Kaufmann, ISBN: 10: 1558605525, pp: 371