ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN ĐỨC NGỌC

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÂN CỤM NỬA GIÁM SÁT ỨNG DỤNG CHO BÀI TOÁN PHÂN CỤM DỮ LIỆU WEB SERVER LOGS

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN, 2018

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN ĐỨC NGỌC

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÂN CỤM NỬA GIÁM SÁT ỨNG DỤNG CHO BÀI TOÁN PHÂN CỤM DỮ LIỆU WEB SERVER LOGS

Chuyên ngành: Khoa học máy tính
Mã số: 8480101

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: TS. Vũ Việt Vũ

THÁI NGUYÊN, 2018

LỜI CẢM ƠN

Lời đầu tiên, tôi xin được gửi lời cảm ơn sâu sắc tới TS. Vũ Việt Vũ, người đã trực tiếp hướng dẫn tôi thực hiện luận văn. Thầy đã tận tình hướng dẫn, cung cấp tài liệu và định hướng cho tôi trong suốt quá trình nghiên cứu và thực hiện luận văn.

Tôi xin chân thành cảm ơn các thầy cô đã giảng dạy và quản lý đào tạo đã tạo điều kiện cho tôi có một môi trường học tập, nghiên cứu tốt trong suốt 2 năm theo học.

Cuối cùng tôi xin được gửi lời cảm ơn tới gia đình, bạn bè và đồng nghiệp đã giúp đỡ và động viên tôi trong suốt quá trình học tập và hoàn thiện luận văn.

Xin chân thành cảm ơn!

MỤC LỤC

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

Trong vài thập niên gần đây, cùng với sự thay đổi và phát triển không ngừng của ngành công nghệ thông tin nói chung và trong các ngành công nghệ phần cứng, phần mềm, truyền thông và hệ thống các dữ liệu phục vụ trong các lĩnh vực kinh tế - xã hội nói riêng. Việc thu thập thông tin cũng như nhu cầu lưu trữ thông tin càng ngày càng lớn. Bên cạnh đó việc tin học hoá một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu lưu trữ khổng lồ. Hàng triệu Cơ sở dữ liệu đã được sử dụng trong các hoạt động sản xuất, kinh doanh, quản lý..., trong đó có nhiều Cơ sở dữ liệu cực lớn cỡ Gigabyte, thậm chí là Terabyte. Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kĩ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kĩ thuật Khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền Công nghệ thông tin thế giới hiện nay. Một vấn đề được đặt ra là phải làm sao trích chọn được những thông tin có ý nghĩa từ tập dữ liệu lớn để từ đó có thể giải quyết được các yêu cầu của thực tế như trợ giúp ra quyết định, dự đoán,… và Khai phá dữ liệu (Data mining) đã ra đời nhằm giải quyết các yêu cầu đó.

Khai phá dữ liệu được định nghĩa là: Quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong lượng lớn dữ liệu được lưu trữ trong các Cơ sở dữ liệu, kho dữ liệu… . Hiện nay, ngoài thuật ngữ khai phá dữ liệu, người ta còn dùng một số thuật ngữ khác có ý nghĩa tương tự như: Khai phá tri thức từ Cơ sở dữ liệu (knowlegde mining from databases), trích lọc dữ liệu (knowlegde extraction), phân tích dữ liệu/mẫu (data/pattern analysis), khảo cổ dữ liệu (data archaeology), nạo vét dữ liệu (data dredging). Nhiều người coi khai phá dữ liệu và một thuật ngữ thông dụng khác là khám phá tri thức trong Cơ sở dữ

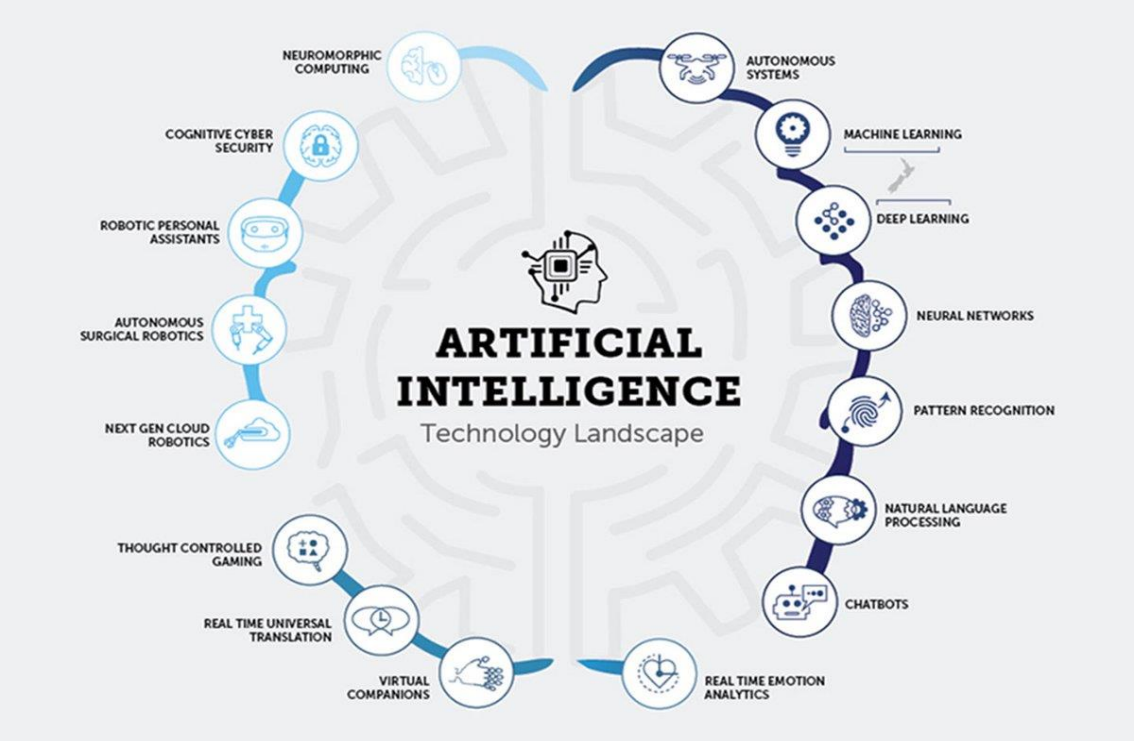
liệu(Knowlegde Discovery in Databases – KDD) là như nhau. Tuy nhiên trên thực tế, khai phá dữ liệu chỉ là một bước thiết yếu trong quá trình Khám phá tri thức trong Cơ sở dữ liệu. Ngay từ những ngày đầu khi xuất hiện, Data mining đã trở thành một trong những xu hướng nghiên cứu phổ biến trong lĩnh vực học máy tính và công nghệ tri thức. Nhiều thành tựu nghiên cứu của Data mining đã được áp dụng trong thực tế. Data mining có nhiều hướng quan trọng và một trong các hướng đó là phân cụm dữ liệu (Data Clustering). Phân cụm dữ liệu là quá trình tìm kiếm để phân ra các cụm dữ liệu, các mẫu dữ liệu từ tập Cơ sở dữ liệu lớn. Phân cụm dữ liệu là một phương pháp học không giám sát.

Trong những năm trở lại đây, do phương pháp phân cụm dữ liệu không giám sát còn một số hạn chế vì vậy dựa trên học không giám sát và học có giám sát đã ra đời một phương pháp phân cụm dữ liệu mới đó là phương pháp phân cụm dữ liệu nửa giám sát. Phương pháp phân cụm nửa giám sát không phải là một phương pháp phân cụm hoàn thiện nhưng nó đã phần nào khắc phục được những hạn chế và phát huy ưu điểm của phương pháp phân cụm không giám sát.

Chương 1. TỔNG QUAN

1.1. Khái niệm về học máy và bài toán phân cụm dữ liệu

Học máy (Machine Learning) là một nhánh nghiên cứu của Trí tuệ nhân tạo nhằm xây dựng các thuật toán thực hiện trên hệ thống máy tính có thể học được qua các dữ liệu mẫu thống kê có sẵn. Trí tuệ nhân tạo (artificial intelligence) gồm rất nhiều lĩnh vực nghiên cứu [1]. Hình 1.1 minh họa các hướng nghiên cứu trong lĩnh vực trí tuệ nhân tạo. Chúng ta có thể kể đến học máy, học sâu, nhận dạng đối tượng, các hệ thống tự động, xử lý ngôn ngữ tự nhiên, trợ lý ảo,… Trí tuệ nhân tạo là một trong ba trụ cột của cuộc cách mạng công nghiệp 4.0 cùng với dữ liệu lớn (Big Data) và Internet vận vật. (IoT).



Hình 1.1. Các hướng nghiên cứu của Trí tuệ nhân tạo [1]

Trên thực tế có 4 dạng học cơ bản bao gồm:

- Học có giám sát: Máy tính được học một số mẫu gồm đầu vào (Input) và đầu ra (Output) tương ứng trước. Sau khi học xong các mẫu này, máy tính