* Khái niệm về phân lớp – phân loại văn bản (Document /categorization classification)
* Phân loại tài liệu trong lĩnh vực công nghệ thông tin sử dụng ACM.
* Các phương pháp phân lớp dữ liệu.
* Các thuật toán phân lớp dữ liệu.
* Một số đề tài về phân lớp dữ liệu.

1. Khái niệm về phân lớp – phân loại văn bản (Document /categorization classification)

Phân lớp – phân loại văn bản (document /categorization classification) [[1]](#footnote-2) là việc xác định chuyên mục (categories) cho một văn bản điện tử dựa trên việc phân tích nội dung của văn bản đó. Một văn bản có thể thuộc một hoặc nhiều chuyên mục. Có hai hướng tiếp cận giải quyết bài toán phân lớp văn bản đó là : phân lớp dựa trên hệ chuyên gia và tiếp cận dựa trên máy học.

### Phân lớp văn bản dựa trên cách tiếp cận hệ chuyên gia:

Theo cách tiếp cận này, việc phân lớp văn bản tự động được điều khiển bằng tay bởi các chuyên gia tri thức và hệ chuyên gia có khả năng đưa ra quyết định phân lớp. Hệ chuyên gia bao gồm một tập các luật logic định nghĩa bằng tay, cho mỗi loại, có dạng:

If (DNF formula) then (category).

Công thức DNF (“Disjunctive Normal Form”) là hợp của các mệnh đề liên kết, tài liệu được phân lớp vào category nếu nó thỏa mãn công thức, nghĩa là, nếu nó thỏa mãn ít nhất một mệnh đề trong công thức.

Điều trở ngại của cách tiếp cận này là hạn chế trong quá trình thu nhận tri thức từ tài liệu của các hệ thống chuyên gia. Nghĩa là, các luật phải được định nghĩa bằng tay bởi kỹ sư tri thức với sự giúp đỡ của chuyên gia về lĩnh vực được nêu trong tài liệu. Nếu tập hợp của các loại được cập nhật, thì hai nhà chuyên gia phải can thiệp lại, và nếu phân lớp được chuyển hoàn toàn sang một phạm vi khác, một chuyên gia về lĩnh vực này cần thiết phải can thiệp vào và công việc phải được bắt đầu lại từ tập tài liệu ban đầu.

### Phân lớp văn bản dựa trên cách tiếp cận máy học

Máy học (Machine learning [[2]](#footnote-3)) là một môn khoa học liên quan đến việc thiết kế và phát triển của các thuật toán cho phép các máy tính phát triển các hành vi dựa trên dữ liệu thực nghiệm. Quá trình học của máy sử dụng các dữ liệu được gọi là dữ liệu học để xác định những đặc trưng mà máy cần phải học từ đó xây dựng bộ quyết định cho dữ liệu mới.

### *Theo cách tiếp cận này, quá trình học để xây dựng tự động một bộ phân lớp được thực hiện bằng cách cho máy học các đặc trưng của tập hợp các tài liệu đã được phân bằng tay bởi chuyên gia về lĩnh vực cần phân loại.Từ đó khi cần phân loại một tài liệu hệ thống sẽ dựa vaò bộ phân lớp được học trước đó và các đặc trưng của tài liệu mới để phân loại tài liệu theo đún gnhóm . Trong kỹ thuật máy học, bài toán phân lớp là hoạt động học có giám sát, quá trình học được “giám sát” bởi tri thức của các phân lớp và của các mẫu huấn luyện thuộc chúng.*

### *Học có giám sát [[3]](#footnote-4) là một kĩ thuật của ngành* [*học máy*](http://vi.wikipedia.org/wiki/H%E1%BB%8Dc_m%C3%A1y) *để xây dựng một hàm (function) từ dữ liệu huấn luyện.* [*Dữ liệu huấn luyện*](http://vi.wikipedia.org/wiki/T%E1%BA%ADp_hu%E1%BA%A5n_luy%E1%BB%87n) *bao gồm các cặp gồm đối tượng đầu vào (thường dạng vec-tơ), và đầu ra mong muốn. Đầu ra của một hàm có thể là một giá trị liên tục (gọi là* [*hồi qui*](http://vi.wikipedia.org/wiki/Ph%C3%A2n_t%C3%ADch_h%E1%BB%93i_qui)*), hay có thể là dự đoán một nhãn phân loại cho một đối tượng đầu vào (gọi là* [*phân loại*](http://vi.wikipedia.org/wiki/Ph%C3%A2n_lo%E1%BA%A1i_b%E1%BA%B1ng_th%E1%BB%91ng_k%C3%AA)*). Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho một đối tượng bất kì là đầu vào hợp lệ, sau khi đã xem xét một số ví dụ huấn luyện (nghĩa là, các cặp đầu vào và đầu ra tương ứng). Để đạt được điều này, chương trình học phải tổng quát hóa từ các dữ liệu sẵn có để dự đoán được những tình huống chưa gặp phải theo một cách "hợp lí")*

Một số thuật toán dựa trên tiếp cận máy học được sử dụng phổ biến hiện nay gồm:Cây quyết định, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), thuật toán Naïve Bayes (NB), Neural Network (NNet).

Ngoài ra, còn một số thuật toán được phát triển dựa trên các thuật toán đã nói ở trên ví dụ với SVM ta có các thuật toán cải tiến như Fuzzy Support Vector Machines. Cũng có một số phương pháp người ta kết hợp các thuật toán lại với nhau như Support Vector Machines Nearest Neighbor (SVM-NN) việc kết hợp này sẽ tận dụng những ưu điểm và bổ xung những mặt yếu của các thuật toán mà ta sẽ thấy ở phần trình bày về các thuật toán.

1. <http://en.wikipedia.org/wiki/Document_classification> [↑](#footnote-ref-2)
2. <http://en.wikipedia.org/wiki/Machine_learning> [↑](#footnote-ref-3)
3. <http://vi.wikipedia.org/wiki/H%E1%BB%8Dc_c%C3%B3_gi%C3%A1m_s%C3%A1t> [↑](#footnote-ref-4)