LỜI CẢM ƠN

Trong suốt khóa học 2005 – 2009 tại trường Đại Học Dân Lập Hải Phòng với sự giúp đỡ của quý thầy cô và giáo viên hướng dẫn về mọi mặt, từ nhiều phía nhất là trong thời gian thực hiện đề tài, nên đề tài của em đã được hoàn thành đúng thời gian quy định.

Em xin gửi lời cảm ơn chân thành nhất tới thầy giáo hướng dẫn Th.s Nguyễn Trịnh Đông đã tận tình hướng dẫn, giúp đỡ, tạo điều kiện để em hoàn thành khóa luận này.

Em xin gửi lời cảm ơn chân thành tới Bộ môn Công Nghệ Thông Tin cùng toàn thể các thầy cô trong khoa cũng như toàn thể các thầy cô trong trường đã giảng dạy những kiến thức chuyên môn làm cơ sở để em thực hiện tốt cuốn luận văn tốt nghiệp này và đã tạo điều kiện thuận lợi để em hoàn thành khóa học.

Em xin chân thành cảm ơn !

Hải Phòng, ngày 28 tháng 6 năm 2009

Sinh Viên

Vũ Thị Thắm

MỤC LỤC

GIỚI THIỆU

Thuật toán SOM là một biểu tượng của lớp mạng neural học không giám sát. Trong đó, sơ khai đầu tiên của SOM được phát minh bởi giáo sư Teuvo Kohonen tại trung tâm nghiên cứu của mạng Neural- Network (1981-1982). Ông đã ứng dụng SOM vào rất nhiều những chương trình phiên bản một cách nhanh chóng và hiệu quả.

Trọng tâm của SOM là đưa và hiển thị dữ liệu hoặc cụm dữ liệu một cách rõ ràng lên mảng một hoặc hai chiều. Nếu các biến trong bản ghi dữ liệu là các vector thì các biến đó sẽ được mô tả như một dữ liệu thống kê, được sử dụng độc lập các mức xám hoặc các mã màu nền riêng. Dùng SOM khai phá để tìm ra được mối quan hệ hữu ích, phụ thuộc lẫn nhau giữa các biến và cấu trúc của dữ liệu.

Lĩnh vực khai phá dữ liệu văn bản cho đến nay đã đạt mục tiêu chính: đó là chứng minh được bằng lý thuyết và thực nghiệm rằng bản đồ văn bản tự tổ chức là một công cụ trọng tâm có nhiều triển vọng, và việc xây dựng những bản đồ như vậy là hoàn toàn tự động. Tuy nhiên, mọi thành quả chỉ mới là ở giai đoạn sơ khai, còn tồn đọng rất nhiều vấn đề không thể giải quyết một cách bao quát được, đặc biệt quan trọng là vấn đề chọn lựa đặc trưng cho nội dung văn bản trong quá trình xây dựng bản đồ, cũng như việc đánh giá chất lượng bản đồ kết quả. Đó là những điều rất đáng phải suy nghĩ

Tính cấp thiết của đề tài nằm ở những mối quan tâm đó - những gì còn chưa đầy đủ và không thể bao quát được của mô hình đã có - khi ứng dụng vào của Tiếng Việt. Trong giai đoạn tiền xử lý, bao hàm trọng tâm là phương pháp chọn lựa đặc trưng cho văn bản, thật ra còn quyết định chất lượng bản đồ nhiều hơn là các yếu tố khác. Sự triển khai lĩnh vực khai phá dữ liệu văn bản trong các ngôn ngữ đặc thù thì dường như là những đề tài vô tận.

Đề tài nghiên cứu mọi khía cạnh tổng quát của mô hình khai phá dữ liệu văn bản với thuật toán bản đồ tự tổ chức, sau đó triển khai với một ngữ liệu văn bản Tiếng Việt

Nội dung cụ thể của đề tài bao gồm việc trình bày tổng quan về các lĩnh vực nghiên cứu có liên quan, thu thập, tổ chức ngữ liệu văn bản và tiền xử lý; xây dựng mới và nghiên cứu các thuật toán chọn lựa đặc trưng: xác định ngữ đoạn, xác định cụm từ, xác định các từ vựng theo chỉ số hữu ích từ vị của Rosengren, xác định các từ khóa theo quan điểm Guiraud; nghiên cứu các phương pháp mã hóa văn bản dựa trên từ vựng, cụm từ, ngữ đoạn;nghiên cứu thuật toán bản đồ tự tổ chức (Self Organizing Map), thuật toán chiếu ngẫu nhiên; đánh giá bản đồ văn bản theo những phương pháp khác nhau.

Ngoài ra, đề tài còn triển khai hai vấn đề quan trọng, đó là cơ sở của việc khám phá và quản lý tri thức trên bản đồ: gom nhóm trên bản đồ và gán nhãn trên bản đồ. Ứng dụng ngữ đoạn trong việc gán nhãn các đơn vị bản đồ và các vùng văn bản. Những vấn đề này đã được một số tác giả nước ngoài nghiên cứu bước đầu.

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.TIẾNG VIỆT

1.1. Giới thiệu đặc trưng của ngữ pháp tiếng Việt

Khi đi sâu tìm hiểu về tiếng Việt, ta có thể thấy rằng có khá nhiều khác biệt so với các ngôn ngữ khác như tiếng Anh, tiếng Pháp, … về tất cả các khía cạnh: âm tiết, từ, câu và các quy tắc liên kết các thành phần đó lại với nhau. Những khác biệt đó cho ta cơ sở để xây dựng và cải tiến cho chương trình kiểm lỗi chính tả đối với tiếng Việt.

Đặc trưng nổi bật của tiếng Việt đó là thuộc dòng Nam Á và là loại hình ngôn ngữ đơn lập, không biến hình. Trong tiếng Việt thì quan hệ giữa các từ được biểu thị không phải bằng các phụ tố chứa trong bản thân từ mà bằng những phương tiện nằm ngoài từ như trật tự từ, hư từ. Chính đặc điểm này bao quát ngữ pháp tiếng Việt cả về ngữ âm, ngữ pháp và ngữ nghĩa.

Trong tiếng Việt, có các đơn vị chính cấu tạo nên đó là:

- Tiếng
- Từ
- Câu

Mỗi đơn vị đó lại có những đặc trưng nổi bật riêng biệt mà ta sẽ tìm hiểu sau đây:

1.1.1. Tiếng

Về giá trị ngữ âm thì tiếng chính là âm tiết. Khi nói thì cứ phát âm ra một hơi thì thành một âm tiết. Về mặt cấu tạo thì tiếng gồm có phụ âm đầu, vần, phụ âm cuối và dấu thanh.

Bảng 2.1.1: Bảng các thành phần âm tiết

Phụ âm đầu	b c d đ g h k l m n q r s t v x ch gh gi kh ng nh ph qu th tr ngh
Nguyên âm	a â ă e ê i o ô ơ u ư y ai ao au ây eo êu ia iu iê oa oi oe oă oo ôi ơi ua uy ui uâ uô uê uơ ưa ưi ươ ưu yê iêu oai oao oay oeo uôi uây uyê ươi ươu uya uyu uêu yêu
Phụ âm cuối	c p t m n ch ng nh
Dấu thanh	huyền, hỏi, ngã, sắc, nặng

Về mặt giá trị ngữ nghĩa tiếng là đơn vị nhỏ nhất có thể có nghĩa. Về mặt giá trị ngữ pháp, tiếng là đơn vị ngữ pháp để cấu tạo nên từ tiếng Việt.

1.1.2. Từ

Từ chính là đơn vị cấu tạo nên câu trong tiếng Việt. Từ trong tiếng Việt có đặc trưng nổi bật là đa âm tiết, cụ thể là một từ có thể có một hoặc nhiều âm tiết khác biệt so với tiếng Anh, mỗi từ chính là một âm tiết.

Từ tiếng Việt có một số đặc trưng đã được thống nhất. Thứ nhất, về mặt hình thức, từ là một khối thống nhất về cấu tạo (về chính tả, về ngữ âm, ...). Thứ hai, về mặt nội dung, từ có nghĩa hoàn chỉnh. Và thứ ba, về khả năng của từ thì nó có khả năng hoạt động tự do và độc lập về ngữ pháp. Từ có hai dạng cấu tạo chủ yếu là từ đơn và từ ghép.

- ***Từ đơn*** có cấu tạo là chỉ có một tiếng (âm tiết) duy nhất và nó thuần nhất về cấu tạo.
- ***Từ ghép*** thì có hai dạng cấu tạo là láy và ghép. Trong đó:
  - ***Láy***: Đó là sự sắp đặt các tiếng kế cận nhau sao cho có quan hệ phối hợp ngữ âm và sự phối hợp này tạo nên nghĩa của từ láy. (ví dụ: long lanh, lờ mờ, ...)
  - ***Ghép***: Đó là sự sắp đặt các tiếng kế cận nhau sao cho có quan hệ ngữ nghĩa. Sự phối hợp này tạo nên nghĩa của từ ghép.

Về mặt phân loại, từ có 8 dạng chính:

- ***Danh từ***: Là những từ chỉ sự vật hay sự việc hoặc thực thể có thuộc tính.

Có các tiểu loại là danh từ chung và danh từ riêng. Trong đó:

- ***Danh từ riêng*** là danh từ chỉ tên riêng của người, vật, địa điểm
- ***Danh từ chung*** là các danh từ chỉ đơn vị, sự vật, khái niệm trừu tượng.

- ***Động từ***: đó là các thực từ chỉ trạng thái vận động của người, vật, hay sự việc. Nó gồm có 2 dạng phân loại là dạng độc lập và dạng không độc lập.
  - ***Dạng độc lập*** là dạng động từ mà bản thân nó đã mang nghĩa.
  - Ví dụ: cắt, giặt, ...
  - ***Dạng không độc lập*** là dạng động từ trống nghĩa, biểu thị tình thái vận động, và tự bản thân nó không mang nghĩa trọn vẹn.

  Ví dụ: nên, cần, dám, ...
- ***Tính từ:*** Là những từ thể hiện đặc trưng tính chất của sự vật, sự việc.
- ***Đại từ***: Là lớp từ có tính chất trung gian giữa thực từ và hư từ. Có các dạng sau:

- Đại từ nhân xưng
- Đại từ chỉ định
- Đại từ thay thế.

- ***Phụ từ***: Là các hư từ, có chức năng dẫn suất, sở biểu hình thái.
- ***Trạng từ***: Là các từ chỉ nơi chốn, trạng thái.
- ***Trợ từ***: Là những từ có chức năng gia tăng một sắc thái ý nghĩa, có các dạng sau:
  - Trợ từ tình thái
  - Trợ từ nhấn mạnh
- ***Cảm từ***: là những từ biểu thị tình cảm, cảm xúc.
- ***Số từ***: Là những từ biểu hiện ý nghĩa về số lượng. Gồm có các dạng:

- Số từ xác định
- Số từ không xác định.

1.1.3. Câu

Trong các ngôn ngữ nói chung và tiếng Việt nói riêng, câu là đơn vị ở bậc cao hơn cả. Hai đặc điểm nổi bật của câu là nó có nghĩa hoàn chỉnh và có cấu tạo rất phong phú và đa dạng.

1.2 Khó khăn trong việc nhận dạng từ Tiếng Việt

- Một phần của tiếng Việt Nam giống với tiếng Trung Quốc hoặc tiếng Nhật, nên rất khó định nghĩa một cách chính xác, gây lên sự khác nhau giữa các từ điển, vì vậy góp phần làm cho việc nhận ra các ranh giới của từ khó hơn.

- Phần lớn vốn từ Tiếng Việt có từ tiếng Trung Quốc, các đơn vị này ghép lại với nhau tạo thành đơn vị từ Tiếng Việt. Ví dụ: "công nhân","thương nhân" và "nhân" (là một từ của trung Quốc)

- Có một lớp từ đặc biệt trong Tiếng Việt, đó là từ láy. Thông thường từ láy có hai âm tiết, trong đó có 1 hoặc thậm chí không có âm tiết nào có nghĩa, âm tiết còn lại chỉ là một biến đổi âm của âm tiết kia. Kiểu này rất thông dụng đặc biết là tính từ, trong thực tế hầu hết các tính từ đều là dạng từ láy.

2. NHỮNG PHƯƠNG PHÁP PHÂN TÍCH, KHAI PHÁ DỮ LIỆU

Những phương pháp thường dùng trong phân tích, khai phá dữ liệu đối với các tập dữ liệu nhiều chiều là phương pháp xử lý dữ liệu đầu vào được biểu diễn dưới dạng vector mà không cần có bất kỳ giả thiết nào về sự phân bố dữ liệu. Điều này cũng giả định rằng không có thêm thông tin nào bên ngoài nào khác được dùng. Vấn đề được giải quyết dựa trên cấu trúc thật sự của dữ liệu chứ không phải bằng các giả thuyết có trước về cấu trúc lớp. Mặc dù quá trình phân tích diễn ra theo chế độ không kiểm soát nhưng các nhãn lớp có thể được dùng sau đó để giúp cho việc diễn dịch ý nghĩa của kết quả chứ không ảnh hưởng đến cấu trúc được tìm thấy.

Những vector trong tập dữ liệu đầu vào sẽ được ký hiệu là x_k, k =1,….N, $x_k \in R^n$.

Trong thống kê, các thành phần của vector thường được gọi là các quan sát (observation) ghi nhận trên các biến số. Trong nhận dạng mẫu, người ta thường gọi các thành phần của vector là các đặc trưng.

Các phương pháp được giới thiệu sau đây có điểm chung là đều làm sáng tỏ những cấu trúc bên trong của tập dữ liệu cho trước. Trong các ứng dụng thực tiễn, việc lựa chọn và tiền xử lý dữ liệu thực ra còn có tầm quan trọng nhiều hơn việc lựa chọn phương pháp phân tích dữ liệu. Các vấn đề sau đây giữ vai trò then chốt trong việc áp dụng một phương pháp vào trong các tập dữ liệu nhiều chiều: những loại cấu trúc nào có thể được rút ra từ tập dữ liệu, làm thế nào để mô tả các cấu trúc, và làm thế nào để thu giảm số chiều của dữ liệu cũng như giảm số lượng dữ liệu

2.1 Hiển thị trực quan dữ liệu đa chiều

Một số phương pháp đồ họa được đưa ra để hiển thị trực quan dữ liệu nhiều chiều bằng cách để tạo cho mỗi chiều chi phối một số khía cạnh nào đó của hiển thị, và sau đó tích hợp các kết quả vào trong một hình ảnh. Các phương pháp này có thể dùng để hiển thị trực quan cho bất cứ loại vector dữ liệu nhiều chiều nào, hoặc là bản thân dữ liệu hoặc là các vector mang ý nghĩa mô tả nào đó về tập dữ liệu

Hạn chế của việc áp dụng những phương pháp này trong khai thác dữ liệu là chúng không thu giảm số lượng dữ liệu

2.2 Các phương pháp gom nhóm dữ liệu

Mục đích của phương pháp gom nhóm là thu giảm số lượng dữ liệu bằng cách phân loại hoặc nhóm những mục dữ liệu tương tự lại với nhau. Cách gom nhóm như vậy phản ánh quá trình con người xử lý thông tin, và một trong những lý do để sử dụng các thuật giải gom nhóm là chúng được cung cấp các công cụ tự động trợ giúp cho việc gom nhóm hoặc phân loại. Các phương pháp này dùng để giảm thiểu hóa tối đa yếu tố con người trong quá trình xử lý.

Các phương pháp gom nhóm có thể chia thành hai loại: gom nhóm phân cấp và gom nhóm phân hoạch

- Gom nhóm phân cấp thực hiện việc trộn các nhóm nhỏ thành các nhóm lớn hoặc phân tách các nhóm lớn thành các nhóm nhỏ hơn. Các phương pháp gom nhóm loại này khác biệt nhau ở nguyên tắc thực hiện việc trộn hoặc tách nhóm. Kết quả cuối cùng của thuật giải là một dạng cây biểu diễn các nhóm.
- Gom nhóm phân hoạch nhắm đến phân rã trực tiếp tập dữ liệu thành một tập các nhóm rời nhau. Hàm tiêu chuẩn nhấn mạnh đến cấu trúc cục bộ hoặc

cấu trúc toàn cục dữ liệu. Thông thường, tiêu chuẩn toàn cục yêu cầu tối thiểu hóa một số độ đo về sự khác biệt giữa các nhóm.

Một số phương pháp gom nhóm phân hoạch phổ biến là K- trung bình. **Trong gom nhóm K- trung bình**, hàm tiêu chuẩn là khoảng cách bình phương trung bình của các mục dữ liệu x_k đến trung tâm nhóm gần nhất

$$E_k = \sum_k \| x_k - m_{c(k)} \|^2 \quad (1)$$

Trong đó, $c(x_k)$ là chỉ số của trung tâm nhóm gần x_k nhất. Một thuật giải có thể có để tối thiểu hóa hàm giá thành bắt đầu bằng cách khởi tạo một tập K trung tâm nhóm, ký hiệu là m_i , i=1,….K. Vị trí của m_i được điều chỉnh trong quá trình lặp: ngay lần đầu tiên gán các mẫu dữ liệu vào các nhóm gần nhất, và tính toán lại các trung tâm nhóm cho lần lặp tiếp theo. Vòng lặp kết thúc khi E không thay đổi nữa. Trong một thuật giải lặp, các nhóm chọn ngẫu nhiên sẽ được đánh giá lần lượt, và trung tâm điểm gần nhất được cập nhật.

Phương trình trên cũng dùng trong phương pháp lượng tử hóa vector. Trong lượng tử hóa vector, mục đích tối thiểu hóa lỗi lượng tử hóa bình phương trung bình, là khoảng cách giữa mẫu x và biểu diễn $m_{c(x)}$ của nó. Thuật giải để tối thiểu hóa phương trình trên là tổng quát hóa thuật giải tối thiểu hóa lỗi lượng tử hóa trung bình trên không gian một chiều

Một vấn đề đối với các phương pháp gom nhóm tỏ ra thích hợp với một số kiểu nhóm nào đó, và các thuật giải sẽ gán dữ liệu vào trong các nhóm kiểu như vậy ngay cả khi trong dữ liệu không thực sự có các nhóm như vậy. Tuy nhiên, mục đích không phải là tập dữ liệu mà phải rút ra được cấu trúc các nhóm dữ liệu trong tập dữ liệu. Điều then chốt là phân tích xem tập dữ liệu có bộc lộ một khuynh hướng gom nhóm dữ liệu hay không. Các kết quả phân tích nhóm sau đó cũng cần được kiểm tra tính đúng đắn

Một vấn đề tiềm tàng khác là việc chọn số lượng nhóm: các loại nhóm khác nhau có thể xuất hiện khi K thay đổi. Sự khởi tạo các nhóm sẽ có tính quyết định. Một số nhóm có thể trống nếu trung tâm của chúng được khởi tạo khác xa với sự phân bố dữ liệu .

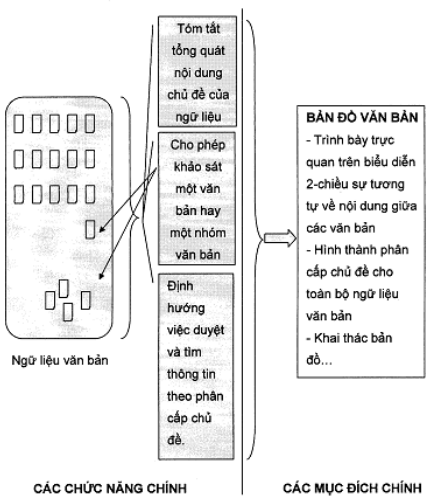
2. 3 Các phương pháp chiếu

Gom nhóm làm giảm số lượng dữ liệu bằng cách nhóm chúng lại với nhau. Một phương pháp khác cũng được dùng để giảm số chiều của dữ liệu. Các phương pháp đó được gọi là các phương pháp chiếu. Mục đích của phép chiếu là biểu diễn các mục dữ liệu đầu vào trong một không gian ít chiều hơn, theo cách thức sao cho một số tính chất nào đó của cấu trúc tập dữ liệu được giữ lại nguyên vẹn đến mức có thể.

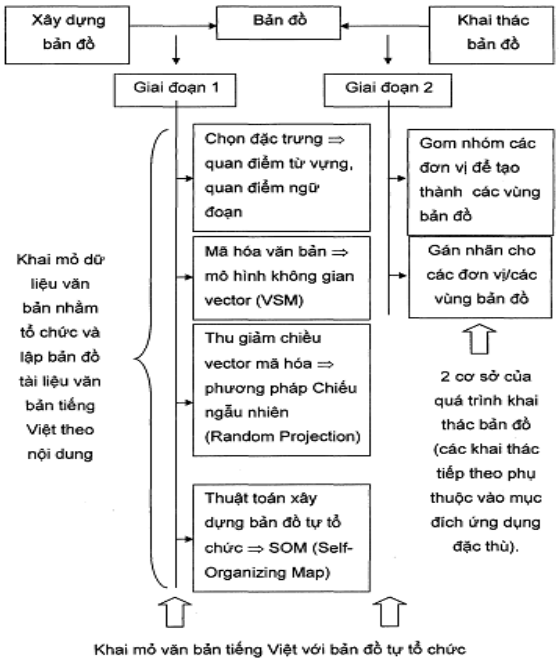
Tính chất nhiều chiều của những tập dữ liệu lớn có thể thu giảm bằng các mạng neuron. Các mạng neuron này chấp nhận những dữ liệu đầu vào được biểu diễn bởi một số lượng nhỏ các biến số, thay vì dùng nhiều chiều cho mỗi mục dữ liệu. Các neuron tìm cách tái cấu trúc những dữ liệu đầu vào đến mức có thể, và sự biểu diễn các mục dữ liệu đã cấu trúc lên mạng neuron được xem như là sự biểu diễn giảm chiều của dữ liệu.

3. KHAI PHÁ DỮ LIỆU VĂN BẢN TIẾNG VIỆT.

3.1.Những chức năng chính của một hệ thống khai phá dữ liệu văn bản.



Các chức năng và mục đích chính của hệ thống khai phá dữ liệu văn bản



Nội dung và phạm vi của đề tài

3.2.Nhu cầu thông tin và những vấn đề liên quan đến văn bản.

Mục tiêu của hệ thống khai phá dữ liệu văn bản là để trợ giúp cho việc người dùng đáp ứng nhu cầu thông tin của họ. Trong một số trường hợp có thể xác định rõ ràng một câu hỏi nào đó cần được trả lời hay một văn bản nào đó cần được tìm kiếm. Ngược lại, trong những trường hợp khác, người ta lại muốn có một cái nhìn tổng quát về một chủ đề nào đó. Đôi khi nhu cầu chỉ đơn thuần là tìm vài thứ quan tâm, hay đạt được một sự hiểu biết chung chung, hay để tìm ra những thông tin mới lạ nào đó ngoài mong đợi. Hơn nữa nhu cầu có thể được người dùng hiểu một cách không rõ ràng, và trong nhiều trường hợp thì khó diễn đạt bằng ngôn ngữ tự nhiên

Những công việc chính liên quan đến các nhu cầu thông tin khác nhau có thể được xem như các chức năng tìm kiếm, khảo duyệt, và hiển thị trực quan mà một hệ thống khai phá dữ liệu văn bản có thể cung cấp.

Tìm kiếm thông tin: trong tiếp cận tìm kiếm, người dùng đặc tả một yêu cầu thông tin bằng các từ dưới dạng truy vấn và yêu cầu hệ thống xác định những văn bản thích hợp với truy vấn. Những cơ chế tìm kiếm trên Internet là ví dụ quen thuộc về những công cụ đặc biệt cho công việc này .

Mô hình tìm kiếm là một dạng rất khiêm tốn của Khai phá dữ liệu văn bản, cho rằng người dùng đã biết khá rõ về những gì cần phải tìm thấy, và bắt buộc họ cũng phải khéo léo trong việc diễn đạt nhu cầu thông tin. Tuy nhiên, nhu cầu có thể là mơ hồ, hay lĩnh vực chưa biết, hoặc đặc biệt khó khăn trong việc sử dụng thuật ngữ để biểu đạt truy vấn.

Khảo duyệt thông tin: trong khi duyệt thông tin, người dùng tự định hướng trong việc chọn lựa văn bản, ví dụ thông qua những liên kết giữa các văn bản như trong WWW, hay thông qua vài cấu trúc phân cấp như thu mục nội dung của một cuốn sách, hay những cấu trúc chủ đề của website.

Cách thức duyệt thông tin cho phép nhu cầu thông tin là mờ hơn hay không biết, bắt nguồn từ việc không yêu cầu có sự mô tả nhu cầu rõ ràng. Thay vì vậy, nhu cầu được truyền đạt ngầm qua những chọn lựa được thực hiện lúc duyệt.

Trong cả hai hướng tiếp cận tìm kiếm và duyệt thông tin, giả sử khi nhu cầu thông tin là rất mơ hồ, hay chung chung, thì việc cung cấp truy cập đến hầu hết những văn bản thích ứng vẫn không thể được đáp ứng. Trong những trường hợp như thế thông tin dạng tổng quát có thể là thích hợp và hữu dụng hơn.

***Hiển thị trực quan*:** có những nhu cầu thông tin đòi hỏi phải đạt đến kết quả là sự đánh giá và chuyển đạt được tính chất tương tự, cũng như sự khác biệt, sự chồng lấn và những mối quan hệ khác giữa các thành phần trong tập dữ liệu.

Những công cụ hữu ích nhất cho việc Khai phá dữ liệu văn bản trong tương lai sẽ xoay quanh các khía cạnh đã đề cập ở trên, cung cấp sự đa dạng về ý nghĩa trong việc khám phá những ngữ liệu văn bản lớn bằng cách cho phép sự đan xen giữa các chức năng: hiển thị trực quan, khảo duyệt, và tìm kiếm.

3.3.Khai phá dữ liệu văn bản với bản đồ biểu diễn trực quan

Việc nghiên cứu những phương pháp phân tích, khảo sát và trình bày những trực quan dữ liệu đã được phổ biến, cung cấp những phương tiện có khả năng minh họa các thuộc tính và mối quan hệ giữa những tập hợp dữ liệu phức tạp .

Thông tin có thể được chuyển tải một cách trực quan bằng cách kết hợp những điểm, đường nét, ký hiệu, từ vựng, màu sắc, và độ bóng trên một bản đồ. Đặc biệt, dùng bản đồ có thể giúp tạo được cảm nhận đối với những tập dữ liệu lớn phức tạp và không thể quản lý được bằng những cách khác. Sự xấp xỉ về mặt không gian được dùng để chuyển đạt tính tương tự của các văn bản, và thông tin tổng quát sẽ được diễn giải tự động bởi người lĩnh hội thông qua thể hiện đồ họa.

HỆ THỐNG KHAI MỎ VĂN BẢN VỚI BẢN ĐỒ TỰ TỔ CHỨC

CHƯƠNG 2: BẢN ĐỒ TỰ TỔ CHỨC – SOM

Bản đồ tự tổ chức SOM (Self- Organizing Map), (Kohonen, 1990, 1995, 1996) là một thuật toán mạng neuron đã được dùng rộng rãi trong nhiều ứng dụng, đặc biệt trong các vấn đề về phân tích dữ liệu.

- Bản đồ tự tổ chức (SOM) là mạng nơ ron hai tầng, sử dụng phương pháp học không chuyên gia.

Một số vấn đề có thể áp dụng SOM bao gồm:

.Gom cụm

. Phân nhóm

. Trực quan dữ liệu

. Phân tích các nhân tố ẩn

2.1 Nội dung thuật toán

Học cạnh tranh là một tiến trình thích nghi, trong đó các neuron của mạng neuron trở nên thích nghi với những loại đầu vào khác nhau, đó là những tập hợp mẫu trong một miền đặc biệt nào đó của không gian đầu vào.

Sự cạnh tranh giữa các neuron diễn ra như sau: Khi xuất hiện một đầu vào *x*, neuron nào có thể biểu diễn tốt nhất cho *x* sẽ được tuyển chọn.

Nếu tồn tại một trật tự học giữa các neuron, nghĩa là các neuron được đặt trên một bản đồ tổ chức, thuật toán học cạnh tranh có thể được tổng quát hóa: không chỉ có neuron chiến thắng mà còn có các lân cận của nó trên bản đồ được phép học, các neuron lân cận sẽ thích ứng để biểu diễn những đầu vào tương tự nhau, và những biểu diễn đó trở nên có trật tự trên bản đồ. Đây là bản chất của thuật toán SOM

Các neuron biểu diễn dữ liệu đầu vào bằng những vector tham chiếu m_i, trong đó các thành phần của nó tương ứng với các trọng số. Một vector tham chiếu được kết hợp cho mỗi neuron - một đơn vị - của bản đồ. Đơn vị, chỉ mục *c*, có vector tham chiếu gần nhất với đầu vào *x* chính là neuron chiến thắng trong tiến trình cạnh tranh:

$$c=c(x) = \text{argmin}\{\| x_i - m_i\|^2\} \qquad (5)$$

Thông thường khoảng cách Euclide được dùng mặc dù những khoảng cách khác có thể tốt hơn .

Đơn vị chiến thắng và các đơn vị lân cận tự động điều chỉnh vector tham chiếu của chúng theo mỗi đầu vào hiện thời để trở nên thích ứng với việc biểu diễn. Số lượng các đơn vị học được triển khai bởi một *lân cận h của nhân*, đây là một hàm giảm theo thời gian, xác định khoảng cách lân cận tính từ đơn vị chiến

thắng. Vị trí của các đơn vị *i* và *j* trên bản đồ được ký hiệu bởi các vector hai chiều r_i và r_j

thì $h_{ij}=(||r_i - r_j||;t)$, trong đó *t* ký hiệu thời gian.

Trong tiến trình học, ở thời điểm *t* các vector tham chiếu được thay đổi lặp đi lặp lại tương ứng với qui tắc thích nghi sau đây, trong đó *x(t)* là đầu vào ở thời điểm *t* và *c*=c (x*(t)*) là chỉ số của đơn vị chiến thắng:

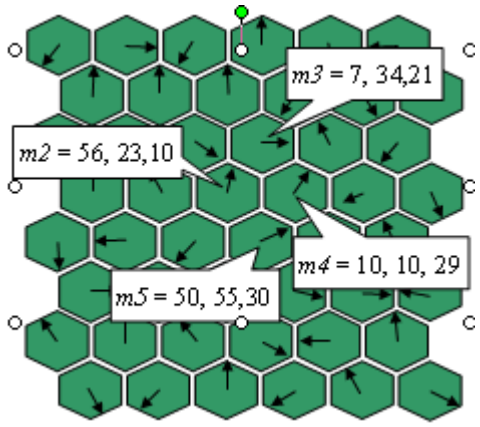
$$m_i\ (t+1)= m_i\,(t)+ h_{ci}(t)\ [x(t)-m_i(t)] \qquad (6)$$

Trong ứng dụng , lân cận của nhân phải có độ rộng rất lớn vào thời điểm bắt đầu tiến trình học để đảm bảo trật tự toàn cục của bản đồ.

Tiến trình học cạnh tranh lựa chọn đơn vị chiến thắng theo phương trình (5) và thay đổi thích nghi trọng số theo phương trình (6).

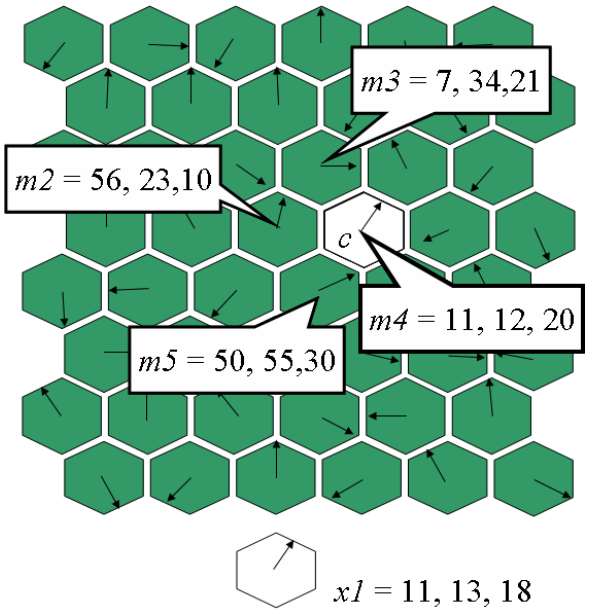
Áp dụng thuật toán SOM (Khởi tạo ngẫu nhiên)

Bản đồ được khởi tạo ngẫu nhiên và mỗi nơ ron được gán với một vecto tham chiếu, ký hiệu là *m*. Các vector được minh họa bằng các mũi tên



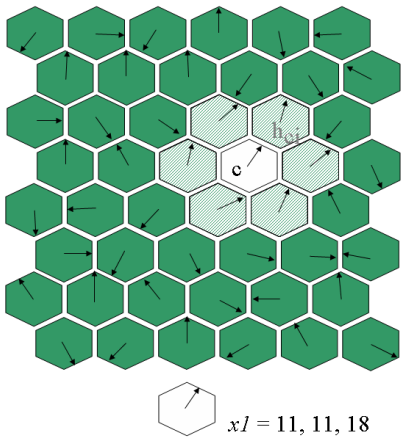
Bước 1: Định vị vector khớp nhất

Mỗi đơn vị dữ liệu đầu vào, được biểu diễn bởi vector *x*, được so sánh với vector tham chiếu $m_{1,2...n}$ của mạng.Vector khớp nhất, vector *c*, được xem như nơron chiến thắng



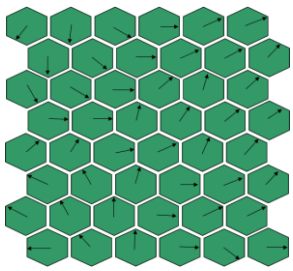
Bước 2: Pha huấn luyện

Các nơ ron trong vùng lân cận h_{ci} của nơ ron chiến thắng c, hướng đến, hay học cái gì đó từ vector dữ liệu đầu vào *x*. Mức độ học hỏi ít nhiều của các nơ ron này phụ thuộc vào yếu tố tốc độ học α



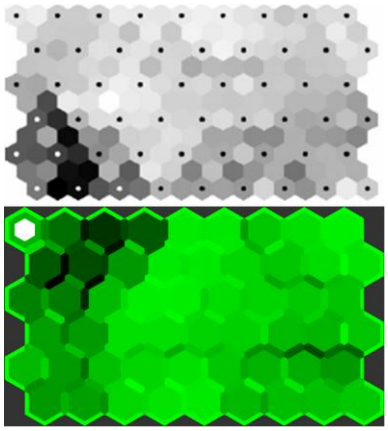
Huấn luyện mạng:

Bước 1 & 2 được lặp lại cho toàn bộ các vector dữ liệu đầu vào, với một số lần cho trước hoặc cho đến khi một chỉ tiêu dừng nào đó được thỏa. Mạng được huấn luyện sẽ biểu diễn một số nhóm các vector. Các nhóm này chuyển tiếp nhau một cách uyển chuyển



Trực quan hóa bản đồ SOM

Phương pháp U_matrix thường được dùng để trực quan hóa SOM_s. Phương pháp U_matrix biểu diễn các khoảng cách nhỏ với các màu sáng, các khoảng cách lớn với các màu tối, tạo nên một bức tranh với các điểm lồi lõm. Cũng có thể biểu diễn các văn bản đồ U_matrix ở dạng màu.



2.2 Những tính chất đặc biệt.

Trình bày có trật tự: một sự trình bày có trật tự các mục dữ liệu giúp cho dễ hiểu về cấu trúc của tập dữ liệu. Ngoài ra, với cùng một sự trình bày có thể dùng để chuyển tải nhiều loại thông tin khác nhau.

Hiển thị trực quan các nhóm: bản đồ được trình bày một cách có trật tự sẽ dùng để minh họa mật độ gom nhóm trong những vùng khác nhau của không gian dữ liệu. Mật độ các vector tham chiếu trên bản đồ được tổ chức sẽ phản ánh mật độ của các mẫu vào. Trong những vùng được gom nhóm, các vector tham chiếu sẽ gần với nhau, và trong những khoảng không gian trống giữa các nhóm chúng sẽ thưa nhau hơn. Cấu trúc nhóm trong tập dữ liệu có thể thấy được qua việc trình bày khoảng cách giữa những vector tham chiếu của các đơn vị lân cận .

Sự trình bày các nhóm có thể được tổ chức như sau: khoảng cách giữa mỗi cặp vector tham chiếu được tính toán và được tỉ lệ sao cho chúng nằm trong một khoảng giá trị tối thiểu và tối đa nào đó. Khi trình bày bản đồ, mỗi giá trị tỉ lệ khoảng cách sẽ xác định mức xám hoặc màu sắc của điểm trung tâm của các đơn vị bản đồ tương ứng. Giá trị mức xám của những điểm tương ứng với các đơn vị bản đồ được đặt bằng trung bình của một số giá trị khoảng cách gần nhất. Sau khi những giá trị này đã được xác lập, chúng có thể dùng để trình bày bản đồ.

Không đầy đủ dữ liệu: một vấn đề thường xuyên gặp khi áp dụng các phương pháp thống kê là sự thiếu dữ liệu, chẳng hạn như một số thành phần của vector dữ liệu không phải luôn được định nghĩa đối với mọi mục tiêu dữ liệu.

Trong trường hợp của SOM, vấn đề này được xử lý như sau: khi chọn một đơn vị chiến thắng theo phương trình (5) , vector đầu vào *x* có thể so sánh với vector tham chiếu m_i chỉ bằng các thành phần vector hữu hiệu trong *x*. Lưu ý là không có thành phần nào của vector tham chiếu bị thiếu. Nếu chỉ có một tỉ lệ nhỏ thành phần của vector dữ liệu bị thiếu thì kết quả của việc so sánh có thể tương đối chính xác. Khi các vector tham chiếu được điều chỉnh thích nghi theo phương trình (6), chỉ có các thành phần hiện hữu trong *x* bị thay đổi.

Phương pháp trên đã được chứng minh rằng vẫn cho kết quả tốt hơn là việc loại bỏ hẳn những mục dữ liệu do chúng chỉ thiếu một ít thành phần vector dữ liệu. Tuy nhiên, đối với những mục dữ liệu mà đa số các thành phần của vector dữ liệu bị thiếu thì nhất định phải loại bỏ chúng.

Dữ liệu rơi rải: Là những dữ liệu khác biệt nhiều với những dữ liệu khác. Trong trình diễn bản đồ, mỗi dữ liệu rơi rải chỉ ảnh hưởng lên một đơn vị bản đồ và những đơn vị lân cận của nó trong khi phần còn lại của bản đồ vẫn có thể dùng để khám phá những dữ liệu rơi rải có thể bị loại bỏ ra khỏi tập dữ liệu.

2.3 Đặc điểm toán học.

Hàm chi phí: Trong trường hợp tập dữ liệu rời rạc và lân cận của nhân cố định, hàm chi phí:

$$E=\sum_k \sum_i h_{ci} \| x_k - m_i \|^2 \qquad (7)$$

Trong đó chỉ số *c* phụ thuộc vào x_k và các vector tham chiếu m_i (phương trình 5)

Quy tắc học của SOM, phương trình (6), tương ứng với một bước giảm gradient trong khi tối thiểu hóa mẫu

$$E_i= \sum_i h_{ci} \| x_k - m_i \|^2 \qquad (8)$$

Nhận được bằng cách chọn ngẫu nhiên một mẫu *x(t)* ở bước lặp *t*

Liên hệ với gom nhóm K-trung bình: hàm chi phí của SOM, phương trình (7), khá giống với phương trình (1) của thuật toán K-trung bình. Điểm khác biệt là trong SOM, mỗi đầu vào được tính khoảng cách đến tất cả các vector tham chiếu (7), thay vì chỉ tính khoảng cách từ mỗi đầu vào đến vector tham chiếu gần nó nhất (1). Các hàm của SOM được xem là giống với thuật toán gom nhóm qui ước nếu lân cận của nhân là 0.

Mặc dù thuật toán gom nhóm K-trung bình và SOM liên hệ mật thiết với nhau nhưng những phương cách tốt nhất để dùng chúng trong khai phá dữ liệu lại khác nhau. Trong thuật toán gom nhóm K-trung bình, cần phải xác định con số K

nhóm ứng với số lượng có trong tập dữ liệu. Đối với SOM, số lượng các vector tham chiếu có thể chọn lớn hơn bất kể số lượng nhóm.

Liên hệ đến với các đường cong chính yếu: Thuật toán SOM tạo ra một biểu diễn cho tập dữ liệu đầu vào dựa trên sự phân bố của dữ liệu. Biểu diễn của tập dữ liệu do vậy cũng được tổ chức. Các đường cong chính yếu có thể cung cấp một nhìn nhận về đặc trưng toán học của tổ chức.

Mỗi điểm trên đường cong là trung bình của tất cả những điểm chiếu vào nó. Đường cong được hình thành trên những kỳ vọng có điều kiện của dữ liệu. Trong SOM, mỗi vector tham chiếu biểu diễn cho các kỳ vọng có điều kiện, cục bộ của các mục dữ liệu.

Các đường cong chính yếu cũng có một đặc tính khác có thể dùng để giải thích cho thuật toán SOM. Tính chất của một đường cong trong việc biểu diễn một sự phân bố dữ liệu là có thể đánh giá bằng khoảng cách (bình phương) trung bình của các điểm dữ liệu trên đường cong, giống như tính chất của thuật toán K-trung bình được đánh giá bằng khoảng cách (bình phương) trung bình của các điểm dữ liệu đến nhóm gần nhất.

Phân rã hàm chi phí: Hàm chi phí của SOM, phương trình (7), có thể được phân rã thành hai thành phần như sau:

$$E=\sum_k \| x_k - n_c \|^2 + \sum_i \sum_j h_{ij} N_j \| n_i - m_j \|^2 \qquad (9)$$

Trong đó , N_j ký hiệu số lượng các mục dữ liệu gần với vector tham chiếu m_i nhất, và

$$n_i = 1/N_i \sum_{x_k \in V_i} x_k$$

Với V_k là vùng Vonoroi tương ứng với vector tham chiếu m_i

Thành phần thứ nhất trong phương trình (9) tương ứng với hàm chi phí của thuật toán K-trung bình, đó là khoảng cách trung bình từ các điểm dữ liệu đến tâm nhóm gần nhất. Ở đây, các nhóm không được định nghĩa bằng các tâm nhóm mà bằng vector tham chiếu m_i .Thành phần thứ nhất cho biết sự biểu diễn chính xác của bản đồ đối với sự phân bố của dữ liệu.

Thành phần thứ hai có thể diễn dịch như là trật tự của các vector tham chiếu. Khi đánh giá thành phần thứ hai cần lưu ý rằng n_i và m_i rất gần nhau, vì n_i là tâm điểm của nhóm được định nghĩa bởi m_i. Để tối thiểu hóa thành phần thứ hai, các đơn vị gần nhau trên bản đồ phải có vector tham chiếu tương tự nhau.

2.4 Topology và qui luật học.

Thuật toán SOM định nghĩa một phép chiếu phi tuyến từ không gian đặc trưng nhiều chiều R^n vào một bảng 2- chiều chứa M neuron. Các vector đầu vào n- chiều trong không gian gốc được ký hiệu là $x \in R^n$, và mỗi neuron được liên kết với một vector tham chiếu n- chiều w_i.

Thuật toán học cạnh tranh tuyển chọn của SOM dựa trên việc tìm kiếm neuron thích hợp nhất cho mỗi vector đầu vào, bằng cách tính toán khoảng cách hoặc tính điểm giữa mỗi vector đầu vào với tất cả những vector tham chiếu để tìm ra neuron *chiến thắng (winner)*. Sự điều chỉnh vector tham chiếu sẽ xảy ra không chỉ đối với neuron chiến thắng mà còn đối với một số neuron lân cận của nó. Do vậy, những neuron lân cận của neuron chiến thắng cũng được học cùng với một vector đầu vào. Việc học cục bộ này được lặp đi lặp lại nhiều lần sẽ dẫn đến một trật tự toàn cục. Trật tự toàn cục này bảo đảm sao cho những vector gần nhau trong không gian đặc trưng n- chiều ban đầu sẽ xuất hiện trong những neuron lân cận trên bảng 2- chiều.

Mỗi lần lặp trong tiến trình học SOM sẽ gồm những bước sau:

1. Chọn ngẫu nhiên một vector đầu vào, liên kết nó với tất cả vector tham chiếu.
2. Chọn neuron chiến thắng, nghĩa là neuron có vector tham chiếu gần (giống) nhất với vector đầu vào theo tiêu chuẩn đánh giá được định nghĩa trước.
3. Hiệu chỉnh các vector tham chiếu của neuron chiến thắng *j* và của một số neuron lân cận với nó. Các neuron lân cận được chọn lựa dựa trên một hàm đánh giá nào đó.
4. Mô tả chi tiết hơn về tiến trình học cạnh tranh tuyển chọn, không kiểm soát của SOM như sau: Vector đầu vào được so sánh với tất cả các vector tham chiếu w_i i=1,....,M trong bảng 2 – chiều chứa M neuron, bằng cách tính khoảng cách $d(x,w_i)$, để tìm ra neuron chiến thắng. Neuron chiến thắng *j* chính là neuron có khoảng cách tối thiểu giữa các vector tham chiếu với vector đầu vào:
    1. $||x - w_i|| = \min || x - w_k||$, k=1,...,M
5. Quy luật học cạnh tranh tuyển chọn (qui luật Kohonen) được dùng để hiểu chỉnh các vector tham chiếu:
    a. $w_k\ (t+1) = w_k(t) + h_j\ (N_j(t),t)\ (x - w_k\ (t)\ )$,i=1,...,M
6. Mức độ hiệu chỉnh phụ thuộc vào mức độ giống nhau giữa vector đầu vào và vector tham chiếu của neuron, biểu diễn bởi $(x - w_k(t))$ và một hệ số tính bởi hàm $h_j(N_j(t),t)$ có ý nghĩa như là tỷ lệ học.

    1. $\Delta w_k\ (t+1) = h_j\ (N_j(t),t)\ (x - w_k\ (t)\ )$

Tỷ lệ học, còn được gọi *là lân cân của nhân* (neighborhood kernel), là hàm phụ thuộc vào hai thông số: thời gian và không gian lân cận của neuron chiến thắng $N_j(t)$. Không gian lân cận này là một hàm số biến thiên theo thời gian, định nghĩa một tập hợp các neuron chiến thắng. Các neuron trong không gian lân cận được điều chỉnh trọng số theo cùng một qui tắc học nhưng với mức độ khác nhau tùy theo vị trí khoảng cách của chúng đối với neuron chiến thắng.

2.5 Lân cận của nhân.

Thông thường lân cận của nhân được định nghĩa dựa trên đánh giá khoảng cách:

$$h_j(N_j(t),t) = h_j(||r_j - r_i||,t)$$

Trong đó, $0 \leq h_j(N_j(t),t) \leq 1, r_j, r_i \in R^2$ là vector vị trí tương đối của neuron chiến thắng j đối với neuron của i. Đối với lân cận của neuron chiến thắng $r_i \in N_j(t)$, hàm số $h_j(||r_j - r_i||,t)$ trả về giá trị khác 0 cho phép hiệu chỉnh vector tham chiếu. Khoảng cách càng xa thì $h_j(||r_j - r_i||,t)$ giảm dần đến 0. Hàm này giữ vai trò then chốt để tạo nên một trật tự toàn cục từ những thay đổi cục bộ. Sự hội tụ của tiến trình học đòi hỏi hàm $h_j(||r_j - r_i||,t)$ giảm dần đến 0 khi $t \to \infty$

Lân cận của nhân $h_j(N_j(t),t) = h_j(||r_j - r_i||,t)$ thường được quan niệm theo hai cách:

- Tập hợp các neuron xung quanh vị trí hình học của neuron chiến thắng.
- Hàm Gauss xung quanh neuron chiến thắng.

Tập hợp các neuron xung quanh vị trí hình học của neuron chiến thắng phải thu nhỏ dần theo diễn tiến của tiến trình học. Định nghĩa $N_j(t) = N_j(r(t),t)$ là tập hợp các neuron chiến thắng và các neuron lân cận nó trong khoảng bán kính $r(t)$, tính từ neuron chiến thắng đi các hướng.

Sự hội tụ của tiến trình học đòi hỏi bán kính $r(t)$ phải giảm dần trong quá trình học:

$$r(t_1) \geq r(t_2) \geq r(t_3) \geq \ldots$$

trong đó , $(t_1 < t_2 < t_3 < ..)$ là thứ tự các bước lặp. Đầu tiên bán kính rất rộng, sau đó hẹp dần về 0.

Khi hàm $N_j(r(t),t)$ cố định $h_j(N_j(t),t)$ được định nghĩa như sau:

$$h_j(N_j(t),t) = h_j(||r_j - r_i||) = \eta(t)$$

trong đó $\eta(t)$ là tỷ lệ học. Trong tiến trình học, cả bán kính $r(t)$ và $\eta(t)$ giảm đơn điệu theo thời gian.

Có thể chọn $\eta(t)$ như sau:

$$\eta(t)= \eta_{max}(t)(1-t/T)$$

Trong đó T là số bước lặp của tiến trình học.

Một hàm khác dùng để định nghĩa lân cận của nhân là hàm Gauss:

$$h_j(N_j(t),t)= h_j(||r_j - r_i||,t) = \eta(t).\exp((||r_j - r_i||^2) / (2\delta^2(t))$$

trong đó, r_j là vị trí của neuron chiến thắng j và r_i là vị trí của neuron thứ i. $\delta^2(t)$ là bán kính nhân, là lân cận $N_j(t)$ xung quanh neuron chiến thắng j. $\delta^2(t)$ cũng là hàm giảm đơn điệu theo thời gian.

Sau tiến trình học, một bảng 2- chiều hình thành nên một bản đồ, trong đó mỗi neuron i mã hóa cho một hàm mật độ xác xuất $p(x)$ của dữ liệu đầu vào.

Kohonen (1989) cũng đã đề xuất một cách tính theo tích điểm thay vì khoảng cách:

Neuron chiến thắng j: $w_j x= max(w_k, x)$, k=1,….M

Qui tắc học như sau:

$$w_i(t+1) = (w_i(t) + \eta(t)x).(||w_i(t) + \eta(t)x||), i \in N_j(t)$$

với $N_j(t)$ là tập hợp các neuron lân cận của neuron chiến thắng j

và $0 \leq N_j(t) \leq \infty$ là hàm số giảm dần theo tiến trình học.

2.6 Lỗi lượng tử hóa trung bình.

Nếu quan điểm mạng SOM là một dạng mạng lượng tử hóa vector thì có thể định nghĩa lỗi lượng tử hóa trung bình (average quantization error) cho một vector đầu vào như sau:

$d_{SOM}(x,w_j) = \min(x, w_k), k=1,...,M$

Trong đó j là chỉ số của neuron chiến thắng. Khoảng cách có thể được định nghĩa như là bình phương khoảng cách *Euclide* $||x-w_i||^2$. Đối với L vector đầu vào, lỗi lượng tử hóa trung bình được định nghĩa như sau:

$$J_{dist} = \frac{1}{L}\sum_{l=1}^{L}\min_{k=1,...,M}\|x_l - w_k\|^2$$

Chương 3: ỨNG DỤNG SOM TRONG KHAI PHÁ DỮ LIỆU VĂN BẢN TIẾNG VIỆT

1. BIỂU DIỄN VĂN BẢN TIẾNG VIỆT.

Vấn đề lớn nhất đối với dữ liệu văn bản, cũng như đối với bất kỳ kiểu dữ liệu nào khác, đó là việc tìm kiếm một sự biểu diễn thích hợp, hay một mô hình, cho những dữ liệu đang tồn tại, với những tài nguyên hiện hữu trong một thời gian hữu hạn. Cho nên, hiệu năng của mô hình yêu cầu cả chất lượng lẫn tốc độ.

1 .1 Mô hình biểu diễn văn bản.

Hiện nay hầu hết những nghiên cứu trong lĩnh vực Khai phá dữ liệu văn bản đều xem như văn bản nhưng được đặc trưng bởi một tập hợp từ vựng. Cách tiếp cận này, thường đươc gọi là mã hóa kiểu ”gói từ” (bag of word), bỏ qua trật tự của từ và những thông tin về cấu trúc câu, nhưng ghi nhận lại số lần mỗi từ xuất hiện .

Mã hóa như vậy thực ra đã làm đơn giản hóa những thông tin phong phú được thể hiện trong văn bản, cách làm này đơn thuần chỉ là sự thống kê từ vựng hơn là sự mô tả trung thực nội dung. Việc phát triển những mô hình tốt hơn nhưng vẫn khả thi về tính toán và cho phép đánh giá được dữ liệu trên thực tế vẫn còn là một vấn đề thách thức.

Mặc dù độ phức tạp chỉ dừng lại ở cấp độ từ vựng của ngôn ngữ nhưng việc mã hóa trên từ vựng vẫn tạm được xem là có khả năng cung cấp một lượng thông tin ít nhiều thích đáng về những mối kết hợp giữa từ vựng và văn bản, có thể trong chừng mực nào đó đủ cho việc gom nhóm theo chủ đề cũng như việc tìm kiếm thông tin từ những ngữ liệu lớn.

1.2 Mô hình không gian vector (Vector Space Model- VSM).

Mô hình này biểu diễn văn bản như những điểm (hay những vector) trong không gian Euclide t-chiều, mỗi chiều tương ứng với một từ trong vốn từ vựng. Thành phần thứ *i,* và d_i của vector văn bản cho biết tần số lần mà từ vị có chỉ mục *i* xuất hiện trong văn bản. Hơn nữa, mỗi từ có thể có một trọng số tương ứng để mô tả sự quan trọng của nó. Sự tương tự giữa hai văn bản được định nghĩa hoặc là khoảng cách giữa các điểm, hoặc là góc giữa những vector (không quan tâm chiều dài của văn bản).

Bất chấp tính đơn giản của nó, mô hình không gian vector và những biến thể của nó cho đến nay vẫn là cách thông thường nhất để biểu diễn văn bản trong khai phá dữ liệu văn bản. Một lý giải cho điều này là những tính toán vector được

thực hiện rất nhanh, cũng như đã có nhiều thuật toán hiệu quả để tối ưu việc lựa chọn mô hình, thu giảm chiều, và hiển thị trực quan trong không gian vector. Ngoài ra, mô hình không gian vector và những biến thể của nó vẫn còn được đánh giá cao, chẳng hạn như trong lĩnh vực truy tìm thông tin.

Một số vấn đề với mô hình không gian vector là số chiều lớn: kích thước vốn từ của một ngữ liệu văn bản thường là từ vài chục ngàn cho đến vài trăm ngàn từ. Hơn nữa, trong mô hình VSM các từ được xem là độc lập với nhau.

Nhiều nỗ lực đã được tiến hành để có thể biểu diễn văn bản với số chiều ít hơn, thích hợp theo cách tiếp cận trực tiếp dữ liệu. Các phương pháp này thường bắt đầu với mô hình không gian vector chuẩn. Một trong những phương pháp này là chiếu ngẫu nhiên (Random Projection) sẽ được khảo sát chi tiết ở các phần sau.

1.3.Trọng số từ vựng.

Trong khi xem xét ngữ nghĩa của một văn bản người ta cảm thấy rằng dường như là một số từ thể hiện ngữ nghĩa nhiều hơn là những từ khác. Hơn nữa, có sự phân biệt cơ bản giữa những từ ngữ chức năng và những từ ngữ mang nội dung, trong đó có một số từ ngữ mang nội dung dường như thể hiện nhiều về các chủ đề hơn những từ khác.

Bất kể phương pháp nào được dùng để giảm chiều hay để suy ra những chiều tiềm ẩn, việc gán trọng số cho từ vựng chỉ cần đòi hỏi miễn sao nguyên tắc gán trọng số có thể diễn giải được tốt về tầm quan trọng của từ vựng đối với việc biểu diễn văn bản. Trọng số có thể dựa trên mô hình phân bố từ, chẳng hạn như sự phân bố Poisson, hay sự đánh giá thông tin về các chủ đề thông qua entropy.

Một sơ đồ trọng số được dùng thông dụng là *tf * idf* với *tf* là tần suất của một từ vựng trong văn bản, và *idf* là nghịch đảo của số lượng văn bản mà từ vựng đó xuất hiện. Sơ đồ này dựa trên khái niệm rằng những từ vựng xuất hiện thường xuyên trong văn bản thì thường ít quan trọng đáng kể về ngữ nghĩa, và những từ hiếm xuất hiện có thể chứa đựng nhiều ngữ nghĩa hơn.

Ví dụ trọng số W_{ij} của một từ w_i xuất hiện trong văn bản d_j có thể được tính toán như sau:

$$W_{ij}= (1+\log tf_{i,j}).\log \frac{N}{dfi}$$

với tf_{ij} là tần xuất của thuật ngữ i trong văn bản j, và df_i là số lần xuất hiện văn bản, nghĩa là số lượng văn bản mà thuật ngữ i xuất hiện trong đó. Sơ đồ này gán trọng số cực đại cho những từ chỉ xuất hiện trong văn bản duy nhất.

Vì trọng số của từ vựng trong mô hình không gian vector ảnh hưởng trực tiếp đến khoảng cách giữa các văn bản, do vậy các kết quả cụ thể phụ thuộc chủ yếu vào phương pháp gán trọng số.

Những sơ đồ trọng số toàn cục nói trên chỉ nhằm mô tả tầm quan trọng của một từ bất kể ngữ cảnh riêng của nó, chẳng hạn như những từ lân cận hay vị trí của từ cấu trúc văn bản. Thông tin về cấu trúc của văn bản cũng chưa được tận dụng, ví dụ như nhấn mạnh lên những từ tiêu đề hay những từ xuất hiện đầu văn bản.

1.4 Phương pháp chiếu ngẫu nhiên.

Đối với nhiều phương pháp và ứng dụng, vấn đề trọng tâm trong việc biểu diễn văn bản là định nghĩa khoảng cách giữa những văn bản. Một không gian dữ liệu có số chiều lớn sẽ được chiếu lên một không gian có số chiều ít hơn, sao cho những khoảng cách gốc được duy trì một cách gần đúng. Kết quả là những vector cơ sở trực giao trong không gian gốc được thay thế bởi những vector có xác suất trực giao gần đúng.

Thuận lợi của phép chiếu ngẫu nhiên là sự tính toán cực nhanh, phép chiếu ngẫu nhiên có độ phức tạp tính toán là $\Theta(Nl)+\Theta(n)$, với N là số lượng văn bản, l là số lượng trung bình những từ khác nhau trong mỗi văn bản, và n là số chiều gốc của không gian đầu vào. Hơn nữa, phương pháp trên có thể áp dụng được cho mọi biểu diễn vector có số chiều lớn, và với mọi thuật toán dựa trên khoảng cách vector

Những phương pháp thu giảm số lượng chiều tựu chung có thể để đến hai nhóm: nhóm các phương pháp dựa trên việc đúc kết các đặc trưng của dữ liệu và nhóm các phương pháp tỉ xích đa chiều (multidimensional scaling method). Những phương pháp chọn lựa đặc trưng có thể thích ứng cao với tính chất tự nhiên của mỗi loại dữ liệu, và vì vậy chúng không thể thích hợp một cách tổng quát cho mọi dữ liệu. Mặt khác, những phương pháp tỉ xích đa chiều cũng có độ phức tạp tính toán lớn, và nếu số chiều của những vector dữ liệu gốc lớn thì cũng không thể áp dụng được, cho việc giảm chiều.

Một phương pháp giảm chiều mới sẽ tỏ ra cần thiết trong những trường hợp mà các phương pháp giảm chiều hiện có quá tốn kém, hoặc không thể áp dụng được. Chiếu ngẫu nhiên là một phương pháp khả thi về mặt tính toán cho việc giảm chiều dữ liệu, bảo đảm sao cho tính chất tương tự giữa những vector dữ liệu được bảo toàn gần đúng.

(Ritter & Kononen) đã tổ chức các từ vựng dựa trên những thông tin về ngữ cảnh mà chúng có khuynh hướng xuất hiện trong đó. Số chiều của các biểu

diễn ngữ cảnh được giảm nhờ thay thế mỗi chiều của không gian gốc bằng một chiều ngẫu nhiên trong một không gian có số chiều ít hơn.

Phép chiếu ngẫu nhiên có thể giảm số chiều dữ liệu theo cách đảm bảo toàn cấu trúc của tập dữ liệu gốc trong mức độ hữu dụng. Mục đích chính là giải thích bằng cả chứng minh phân tích và thực nghiệm xem tại sao phương pháp này làm việc tốt trong những không gian có số chiều lớn.

1.4.1 Nội dung.

Trong phương pháp chiếu ngẫu nhiên (tuyến tính), vector dữ liệu gốc, ký hiệu $n \in R^N$, được nhận với ma trận ngẫu nhiên R

$$x = Rn \quad (1)$$

Phép chiếu ánh xạ cho các kết quả là một vector giảm chiều $n \in R^d$. Ma trận R gồm những giá trị ngẫu nhiên.

Một điều cần xem xét là những gì đã xảy ra đối với mỗi chiều của không gian gốc R^N trong phép chiếu. Nếu cột thứ i^{th} của R ký hiệu là r_i, việc ánh xạ ngẫu nhiên (1) có thể được biểu diễn như sau:

$$x = \sum_i n_i r_i \quad (2)$$

Thành phần thứ i^{th} của n được kí hiệu n_i .Trong vector gốc n, các thành phần n_i là những trọng số của những vector đơn vị trực giao. Trong (2), mỗi chiều i của không gian dữ liệu gốc đã được thay thế bởi một chiều ngẫu nhiên không trực giao r_i trong không gian giảm chiều.

1.4.2 Đặc điểm.

Ích lợi của phương pháp này chiếu ngẫu nhiên trong việc gom nhóm về cơ bản phụ thuộc vào việc nó ảnh hưởng ra sao đến những tính chất tương tự giữa các vector dữ liệu.

Sự biến đổi đối với các tính chất tương tự: Cosine của góc giữa hai vector thường được dùng để đo lường sự tương tự của chúng. Các kết quả sẽ hạn chế cho những vector có chiều dài đơn vị. Trong trường hợp đó cosine có thể được tính toán như tính của những vector.

Tích của hai vector x và y, đạt được bằng phép chiếu ngẫu nhiên các vector m và n tương ứng, có thể được biểu diễn (1) như sau:

$$x^T y = n^T R^T R_m \quad (3)$$

Ma trận $R^T R$ có thể được phân tích như sau:

$$R^T R = I + \varepsilon \quad (4)$$

Với $\varepsilon_{ij} = R_i^T R_j$

Cho i ≠ j và $\varepsilon_{ij} = 0$ cho tất cả giá trị *i*. Những thành phần trên đường chéo $R^T R$ đã được thu gom thành ma trận đồng nhất *i* trong (4). Chúng luôn bằng đơn vị vì những vector r_i đã được chuẩn hóa. Những đơn vị không nằm trên đường chéo bị thu gom thành ma trận ε. Nếu tất cả những mục trong ε đều bằng 0, nghĩa là những vector r_i và r_j là trực giao, ma trận $R^T R$ sẽ bằng *i* và sự tương tự giữa các văn bản sẽ được bảo toàn một cách chính xác trong phép chiếu ngẫu nhiên, trong thực tế những phần tử trong ε sẽ rất nhỏ nhưng không bằng 0.

Những đặc điểm thống kê của ε: cho phép phân tích những đặc tính thống kê của các phần tử ε, nếu chúng ta cố định sự phân bổ những tử trong ma trận chiếu ngẫu nhiên R, nghĩa là sự phân bố của những thành phần của các vector cột r_i. Giả sử những thành phần được chọn ban đầu là độc lập, phân bố chuẩn và đồng nhất (với kỳ vọng 0), và chiều dài của tất cả r_i được chuẩn hóa. Kết quả của thủ tục này là chiều dài của r_i sẽ được phân bổ đồng nhất

$$E[\varepsilon_{ij}] \qquad (6)$$

Với mọi *i* và *j*, E biểu diễn kỳ vọng trên tất cả những chọn lựa ngẫu nhiên cho các thành phần của R.

Trong thực tế chúng ta luôn luôn dùng một thể hiện đặc biệt của ma trân R ,và vì vậy chúng ta cần biết nhiều hơn sự phân bố ε_{ij} để kết luận về ích lợi của phương pháp ánh xạ ngẫu nhiên. Đã chứng minh được rằng nếu số chiều d của không gian được giảm chiều lớn ε_{ij} xấp xỉ phân bố chuẩn. Sự khác biệt, được biểu diễn bởi σ_ε^2 có thể xấp xỉ bằng:

$$\sigma_\varepsilon^2 \sim 1/d \qquad (7)$$

Những đặc tính thống kê đối với các tính chất tương tự: Cần phải đánh giá xem những tính chất tương tự của các vector trong không gian gốc bị biến đổi như thế nào trong phép chiếu ngẫu nhiên.

Cho hai vector n và m trong không gian dữ liệu gốc, có thể suy ra sự phân bổ tính chất tương tự của các vector x và y nhận được một cách tương ứng bằng phép chiếu ngẫu nhiên của n và m.

Sử dụng (3),(4),(5) tích giữa các vector được chiếu có thể biểu diễn như

$$x^T y = n^T m + \sum_{k \neq l} \varepsilon_{kl} n_k m_l \qquad (8)$$

Ký hiệu $\delta = \sum_{k \neq l} \varepsilon_{kl} n_k m_l$. Kỳ vọng của δ là 0 khi kỳ vọng của mỗi thành phần trong tổng là (8) là 0.

Phương sai của δ, ký hiệu là σ_δ^2 có thể biểu diễn như sau

$$\sigma_\delta^2 = [1+(\sum_k n_k m_k)^2 - 2\sum_k n_k^2 m_k^2]\,\sigma_\delta^2 \qquad (9)$$

Khi chiều dài của các vector dữ liệu gốc *n* và *m* cố định là đơn vị, tích của chúng lớn nhất là 1, và theo phương trình (7)

$$\sigma_\delta^2 \le 2_\delta^2 \approx 2 / d \qquad (10)$$

1.4.3 Chiếu ngẫu nhiên và SOM.

Thuật toán xây dựng một ánh xạ từ không gian đầu vào lên trên một bản đồ 2- chiều. Mỗi vị trí bản đồ được gọi là một đơn vị bản đồ, chứa vector tham chiếu, những vector tham chiếu của các đơn vị bản đồ lân cận cùng học dần dần để có thể biểu diễn những vector đầu vào tương tự nhau. Phép chiếu trở nên có trật tự. Kết quả, bản đồ là một sự biểu diễn tóm tắt, trực quan cho tập dữ liệu.

Thuật toán SOM bao gồm hai bước áp dụng lặp đi, lặp lại. Trước hết đơn vị chiến thắng, đơn vị có vector tham chiếu đối với đầu vào hiện tại được chọn gần nhất, và sau đó những vector tham chiếu của những đơn vị lân cận với đơn vị chiến thắng trên bản đồ được cập nhật.

Vì phép chiếu ngẫu nhiên là tuyến tính, những lân cận hẹp trong không gian gốc sẽ được ánh xạ lên trên những lân cận hẹp trong không gian ít chiều hơn. Trong SOM, những vector tham chiếu của các đơn vị lân cận nói chung là gần nhau và vì vậy những lân cận nhỏ trong không gian gốc hầu hết sẽ được ánh xạ lên trên một đơn vị bản đồ đơn lẻ hay lên trên một tập hợp những đơn vị bản đồ lân cận. Vì thế bản đồ tự tổ chức SOM sẽ không qua nhạy cảm với những sai lệch về tính tương tự gây ra bởi phép chiếu ngẫu nhiên.

Trước khi xem xét các hiệu quả từ phép chiếu ngẫu nhiên cho những dữ liệu đầu vào trên việc học của SOM, cần phải xem xét khái niệm về không gian trống của toán tử chiếu R. Các dòng hình thành một tập hợp các vector ngẫu nhiên trong không gian gốc. Không gian trống của R là không gian con của không gian gốc đã chiếu thành vector zero.

Mỗi vector đầu vào n hiện có trong không gian dữ liệu gốc có thể được phân tích thành tổng của hai thành phần trực giao riêng biệt $n^\wedge$ và $n^\sim = n - n^\wedge$, với $n^\sim$ thuộc về không gian trống của R, và $n^\wedge$ là phần bù của nó. Khi vector đầu vào *n* được chiếu với toán tử ngẫu nhiên, kết quả chỉ phản ánh những phần của n trực giao với không gian trống

$$Rn = Rn^\wedge \qquad (11)$$

Vì vậy, kết quả phép chiếu loại bỏ những thành phần của *n* hiện có trong không gian trống của R

Khi vector *Rn(t)* là đầu vào cho SOM, ở bước thời gian t, những vector tham chiếu m_i được cập nhật theo nguyên tắc sau:

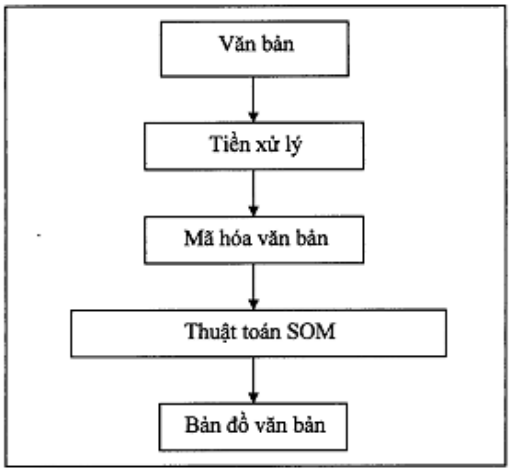
$$M_i(t+1) = m_i(t) + h_{ci}(t)\,[Rn - m_i(t)] \quad (12)$$

Trong đó, h_{ci} là lân cận của nhân, là hàm khoảng cách giữa những đơn vị *i* và *c* trên bản đồ. Ở đây, *c* chỉ là mục của đơn vị có vector tham chiếu gần nhất với *Rn(t)* .

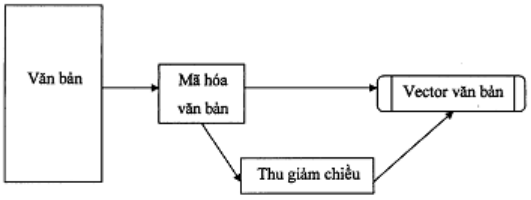
2. BẢN ĐỒ VĂN BẢN TIẾNG VIỆT.

2.1 Mô hình tổng quát.

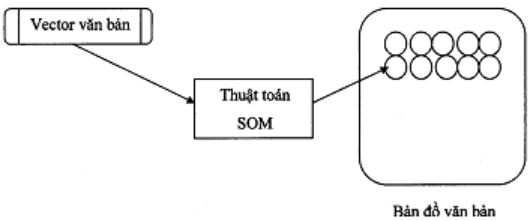
Mô hình tổng quát được xây dựng dựa trên phương pháp WEBSOM. Trong mô hình này, thuật toán SOM được dùng để chiếu những văn bản, được biểu diễn trong không gian ban đầu có số chiều rất lớn, lên trên một bản đồ 2-chiều. Kết quả là những vị trí gần nhau trên bản đồ sẽ chứa đựng những văn bản tương tự nhau. Sau đó, bản đồ có thể được khai thác để trình bày thông tin về ngữ liệu văn bản một cách trực quan, hoặc khảo sát sự gom nhóm, hoặc dùng cho việc tìm kiếm trên các văn bản.



MÔ HÌNH TỔNG QUÁT HÓA CÁC BƯỚC XÂY DỰNG BẢN ĐỒ VĂN BẢN



Văn bản được mã hóa thành vector
Có thể thu giảm số chiều của vector để tính toán



Vector văn bản được chiếu vào bản đồ bằng thuật toán SOM

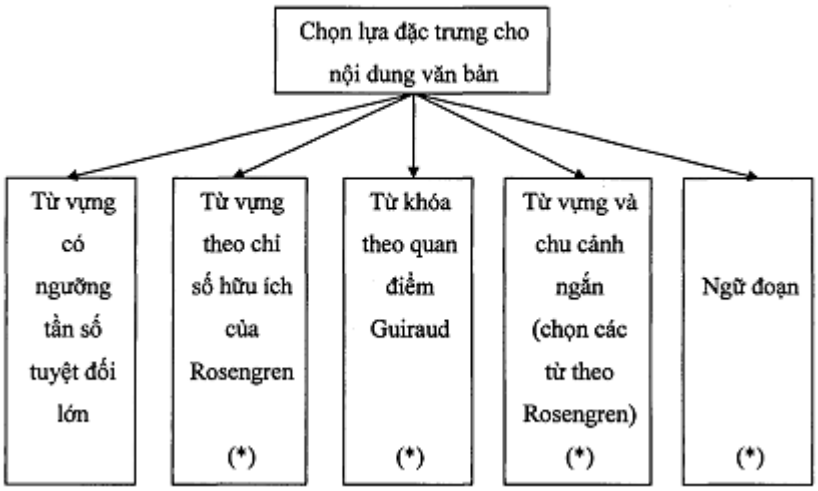
2.2 Tiền xử lý.

Trích tách các đặc trưng là bước quan trọng nhất trong phân tích khám phá dữ liệu cũng như Khai phá dữ liệu văn bản. Tất cả các phương pháp học không kiểm soát đều tìm kiếm một số cấu trúc nào đó trong tập dữ liệu, và các cấu trúc căn bản cũng được xác định bởi các đặc trưng được chọn để biểu diễn các mục dữ liệu. Tính hữu ích của những phương pháp tiền xử lý khác nhau tùy thuộc vào mục đích ứng dụng.

Các thực nghiệm đã công bố trong lĩnh vực Khai phá dữ liệu văn bản hầu như cho đến nay đều sử dụng những phương pháp tiền xử lý khá đơn giản trong việc loại bỏ dữ liệu dư thừa và chọn lựa đặc trưng. Trong các thực nghiệm như vậy, những tiêu đề văn bản, những chữ số, công chức, và tất cả những ký hiệu phi ngôn ngữ đều bị loại bỏ. Văn bản được xem là đặc trưng bởi tập hợp các từ vựng có tần số tuyệt đối lớn, những từ ít xuất hiện bị loại bỏ theo một tần số ngưỡng nào đó (các tác giả đã chọn tần số ngưỡng là 50 cho hầu hết các thực nghiệm, một số ít trường hợp chọn tần số ngưỡng là 10 và 5).

Đề tài tập trung chú ý đến các phương pháp chọn lựa đặc trưng bởi vì đây là yếu tố nền tảng quyết định sự thành công của môt hệ thống khai phá dữ liệu văn bản. Điều này đã được hầu hết các tác giả nhận định, như đã trình bày ở phần 2, những công việc trong giai đoạn tiền xử lý thật ra còn quan trọng và quyết định hơn cả việc chọn lựa các phương pháp phân tích. Đây là một lý lẽ tất yếu, bởi vì các phương pháp, các mô hình hiện nay đều đã có những bề dày lý thuyết ổn định và được triển khai rất nhiều trong thực nghiệm.

Phương pháp chọn lựa đặc trưng dựa trên cơ sở những từ vựng có tần số tuyệt đối lớn có lẽ chỉ thuyết phục và chứng tỏ được mức độ hiệu quả của chúng khi được so sánh và đối chiếu với các phương pháp khác.



(*): Những phương pháp lần đầu tiên được nghiên cứu và thử nghiệm trong đề tài

LÝ DO TRIỂN KHAI CÁC PHƯƠNG PHÁP MỚI:

1. Sự khác biệt cơ bản về loại hình ngôn ngữ đơn lập của tiếng Việt so với những ngôn ngữ biến hình đã được nghiên cứu trong lĩnh vực này, như tiếng Anh và tiếng Phần lan. Cụ thể là quan điểm về đơn vị từ vựng.
2. Phương pháp chọn lựa từ vựng đặc trưng dựa trên tần số ngưỡng có thể không phải là cách thức hiệu quả nhất

NHỮNG PHƯƠNG PHÁP CHỌN LỰA ĐẶC TRƯNG

2.2 .1 Chọn lựa đặc trưng: phương pháp đánh giá độ hữu ích từ vị.

Rosengren định nghĩa tần số hiệu chỉnh KF của một dạng thức W trên *n* khối ngữ liệu K_i i=1,2,…,n, bằng các công thức:

$$KF=(\sum_{i=1}^{n} \sqrt{difi} \ )^2$$

Với d_i là trọng số của K_i trong toàn mẫu, f_i là tần số của W trên K_i. Tần số hiệu chỉnh Rosengren còn được gọi là chỉ số hữu ích của từ vị.

2.2.2 Chọn lựa đặc trưng: phương pháp xác định từ khóa theo quan điểm Guiraud.

Phân hoạch vốn từ vựng dựa trên giả thuyết và phân bố Laplace-Gausse của từ vị:

Từ vựng của khối ngữ liệu K_i so với K_0 có thể được phân hoạch qua đại dương Z,

$$Z = \frac{X - E(X)}{\sigma}$$

E(X) là kỳ vọng của biến ngẫu nhiên X, σ = σ (X) là độ chêch lệch chuẩn của X, Z được gọi là độ lệch thu gọn.

- Nếu Z > 2.58, Guiraud gọi là W là một từ khóa của K_i.
- Nếu Z > 1.96, Muller và Camlong gọi là W là một từ chủ đề của K_i.

Ngoài ra, các từ khóa theo tiêu chuẩn Guiraud cũng là những từ ngữ để theo tiêu chuẩn Muller và Camlong.

2.2.3 Chọn lựa đặc trưng: phương pháp xác định cụm từ trong chu cảnh ngắn.

Chu cảnh ngắn: của một từ là khái niệm dùng để chỉ những từ xuất hiện xung quanh từ đó, được hiểu là một từ đứng trước và một từ đứng sau nó. Đề tài sử dụng 2,757 từ vựng có chỉ số KF của Rosengren cao nhất để làm nòng cốt cho các kết cấu 3- từ. Sau khi xác định tất cả những kết cấu từ có thể, loại bỏ những kết cấu từ có tần số xuất hiện ít hơn 50 lần trong toàn bộ ngữ liệu văn bản. Kết quả giữ lại 5,090 kết cấu từ.

2.2.4 Chọn lựa đặc trưng: phương pháp sử dụng ngữ đoạn.

Câu và ngữ đoạn: Theo tiêu chuẩn Ngữ pháp chức năng, câu không được cấu tạo bằng những đơn vị ngôn ngữ: những từ, những hình vị, những âm vị. Câu được cấu tạo bằng những đơn chức năng gọi là *ngữ đoạn.*

Một ngữ đoạn không được định nghĩa bằng thuộc tính nội tại của nó (vì nó không có những thuộc tính nội tại nhất định, không có cương vị ngôn ngữ học nhất định), mà bằng chức năng cú pháp của nó, và một ngữ đoạn cũng được cấu tạo bằng những ngữ đoạn ở bậc thấp hơn, chứ không phải bằng những đơn vị ngôn ngữ.

Chọn lựa ngữ đoạn đặc trưng: Đề tài sử dụng phương pháp phân tích ngữ đoạn (phần 5) để xây dựng một vốn ngữ đoạn, bao gồm những dạng trung tâm ngữ đoạn đặc trưng cho toàn bộ các văn bản trong ngữ liệu.

2.3 Mã hóa văn bản.

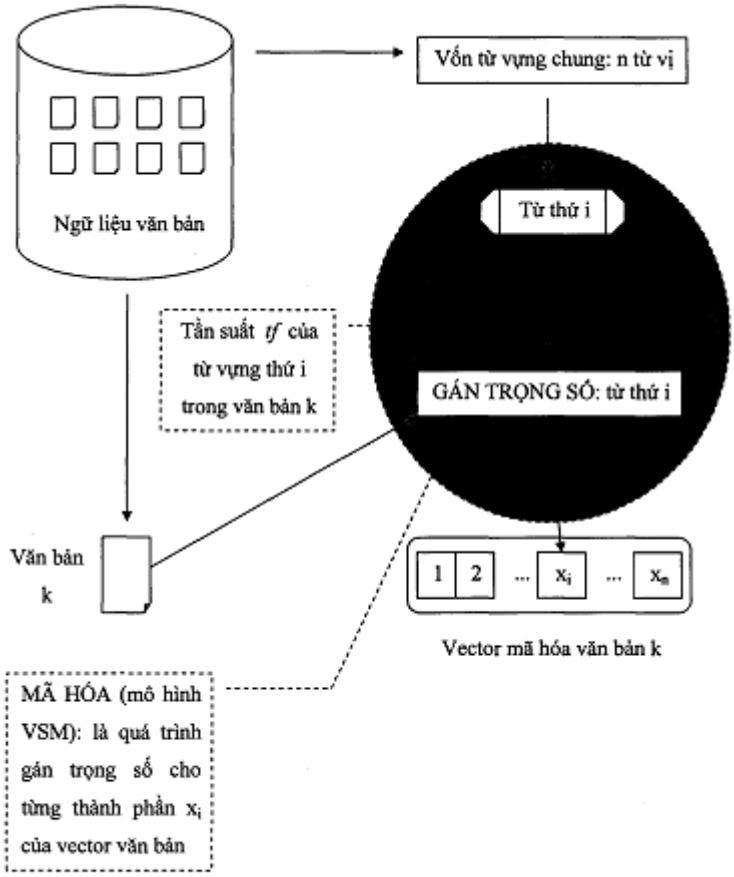
Trọng số: có nhiều phương pháp gán trọng số khác nhau được sử dụng. Thông thường, có thể áp dụng một trong các phương pháp sau đây:

- Dùng tần xuất *tf* của từ vựng.
- Dùng $tf \times idf$, trong đó *idf* là nghịch đảo số văn bản mà từ vựng xuất hiện trong đó

- Dung entropy Shannon trong trường hợp đã có các nhóm giả định trước.

Đề tài sử dụng tần suất xuất hiện của từ vựng trong văn bản để đánh giá trọng số. Khoảng cách Euclide được dùng để tính khoảng cách giữa hai văn bản.

Giảm chiều: mặc dù giai đoạn tiền xử lý đã giảm bớt vốn từ vựng chung ban đầu nhưng đối với những ngữ liệu lớn thì số lượng từ vựng đặc trưng còn lại vẫn rất cao. Các thực nghiệm của đề tài sử dụng phương pháp chiếu ngẫu nhiên để giảm chiều vector văn bản. Số chiều sau khi rút gọn để mã hóa cho một vector văn bản trong thực nghiệm là 100.



MÃ HÓA VĂN BẢN

2.4 Xây dựng bản đồ.

Đề tài cài đặt lại thuật toán SOM và sử dụng trong mô hình xây dựng bản đồ văn bản.

2.4.1 Xác định những thông số quan trọng cho thuật toán SOM.

- Bản đồ gồm 4000 neuron , kích thước 20 $\times$ 20. Trung bình mỗi đơn vị bản đồ có 13.3125 văn bản tập trung, điều này phù hợp với kinh nghiệm cho rằng số lượng văn bản trung bình trên một bản đồ nên khoảng từ 10-15 văn bản.
- Bản đồ được xây dựng chữ T=100,000 bước lặp trong thuật toán SOM.
- Lân cận của neuron chiến thắng được xác định theo những vị trí hình học vuông xung quanh neuron đó $h_j(N_j(t),t)= \eta(t)$
- Hàm tỉ lệ học $\eta(t)= \eta_{max}(t)(1-t/T)$, với η_{max} cho trước băng 50% kích thước bản đồ.

2.4.2 Cài đặt thuật toán SOM.

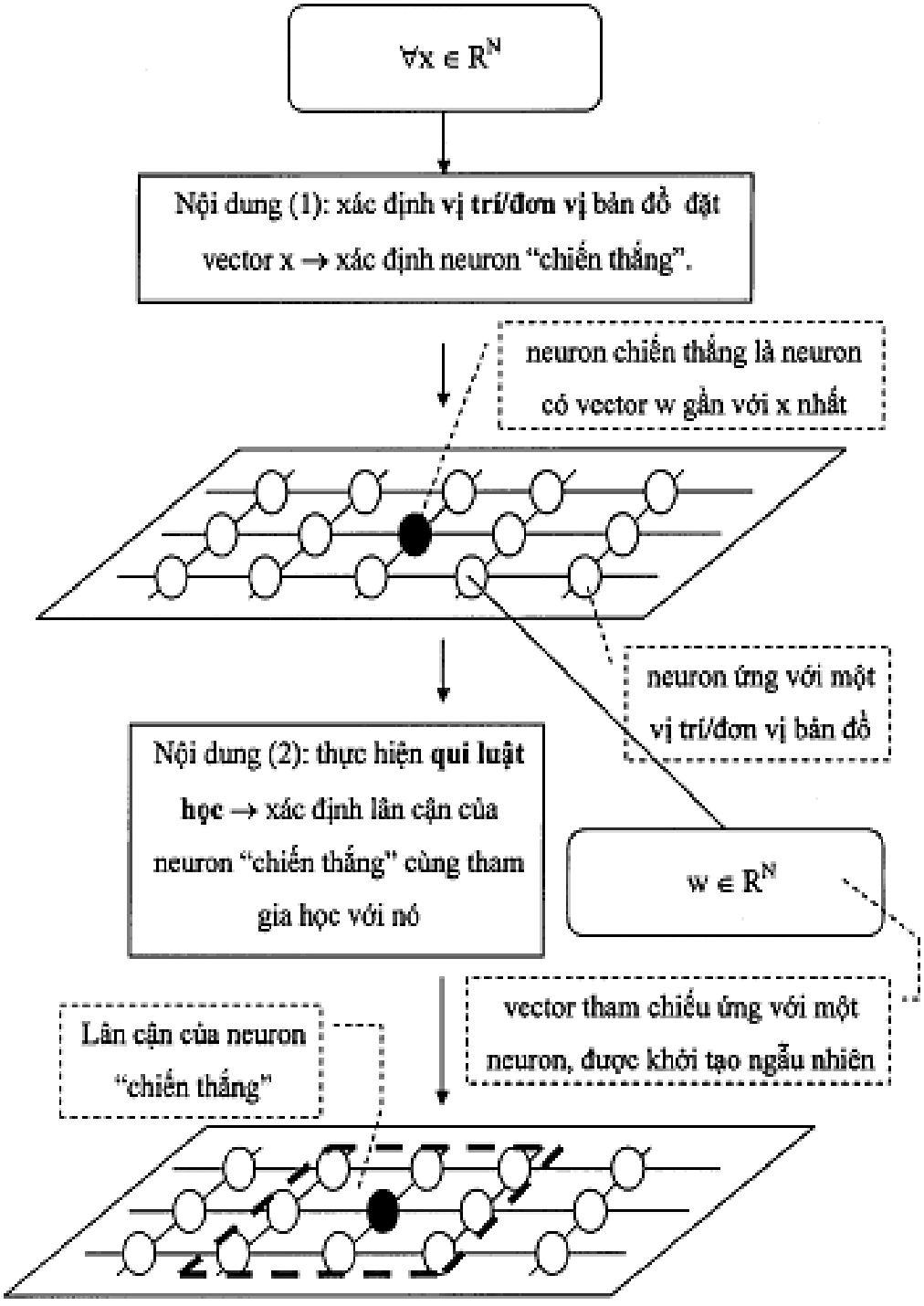
Đầu vào:

- Mạng 2- chiều gồm M neuron.
-Tập hợp dữ liệu gồm L vector đầu vào n-chiều.
- Số bước học T.
- Hàm lân cận của nhân $h_j(N_j(t),t)$.
- Hàm tỉ lệ học $\eta(t)= \eta_{max}(t)(1-t/T)$, với η_{max} cho trước.

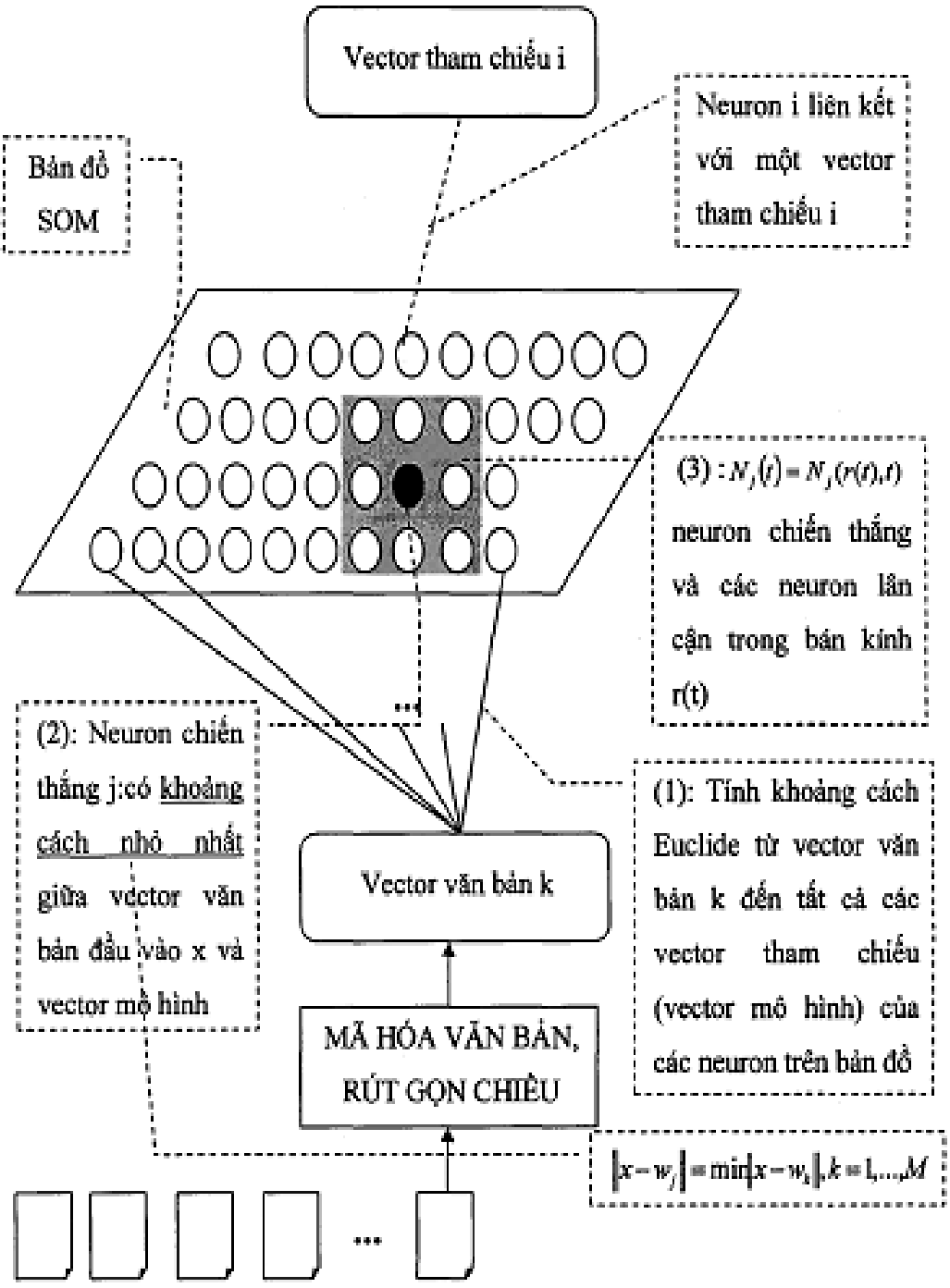
Các bước:

1. Đặt $\eta(t)= \eta_{max}$.
2. Đặt bước học t=0.
3. Chọn giá trị khởi gán ngẫu nhiên cho w_k, k=1,…,M.
4. Chọn ngẫu nhiên vector đầu vào x_i.
5. Tính toán tỷ lệ học $\eta(t)$ ở bước t, với hàm tỷ lệ học cho trước.
6. Tính khoảng cách Euclide: $|| xi - w_k(t) ||, k=1,....M$
   Hoặc tính tích điểm: $y_k= w_k x_i, k=1,...M$
7. Chọn ngẫu nhiên *chiến thắng j:*
   i. $||x_i - w_j(t)|| = min ||x_i(t)- w_k(t)|| , k=1...M$
   ii. Hoặc: $y_i = max(y_{max})$, k=1,…M
8. Định nghĩa tập hợp các neuron lân cận $N_j(t)$ của neuron chiến thắng, với hàm lân cận của nhân $h_j(N_j(t),t)$ cho trước.
9. Hiệu chỉnh trọng só của các neuron trong tập $N_j(t)$:
   1. $w_p(t+1)= w_p(t)+ \eta(t)(x_i - w_k(t)), p \in N_j(t)$
   2. Hoặc $w_p(t+1)= (w_p(t)+ \eta(t) x_i) / (|| w_p(t)+ \eta(t)x_i ||)$, $i \in N_j(t)$
10. Tăng $t=t+1$. Nếu t>T thì dừng ; ngược lại , trở về bước 4.

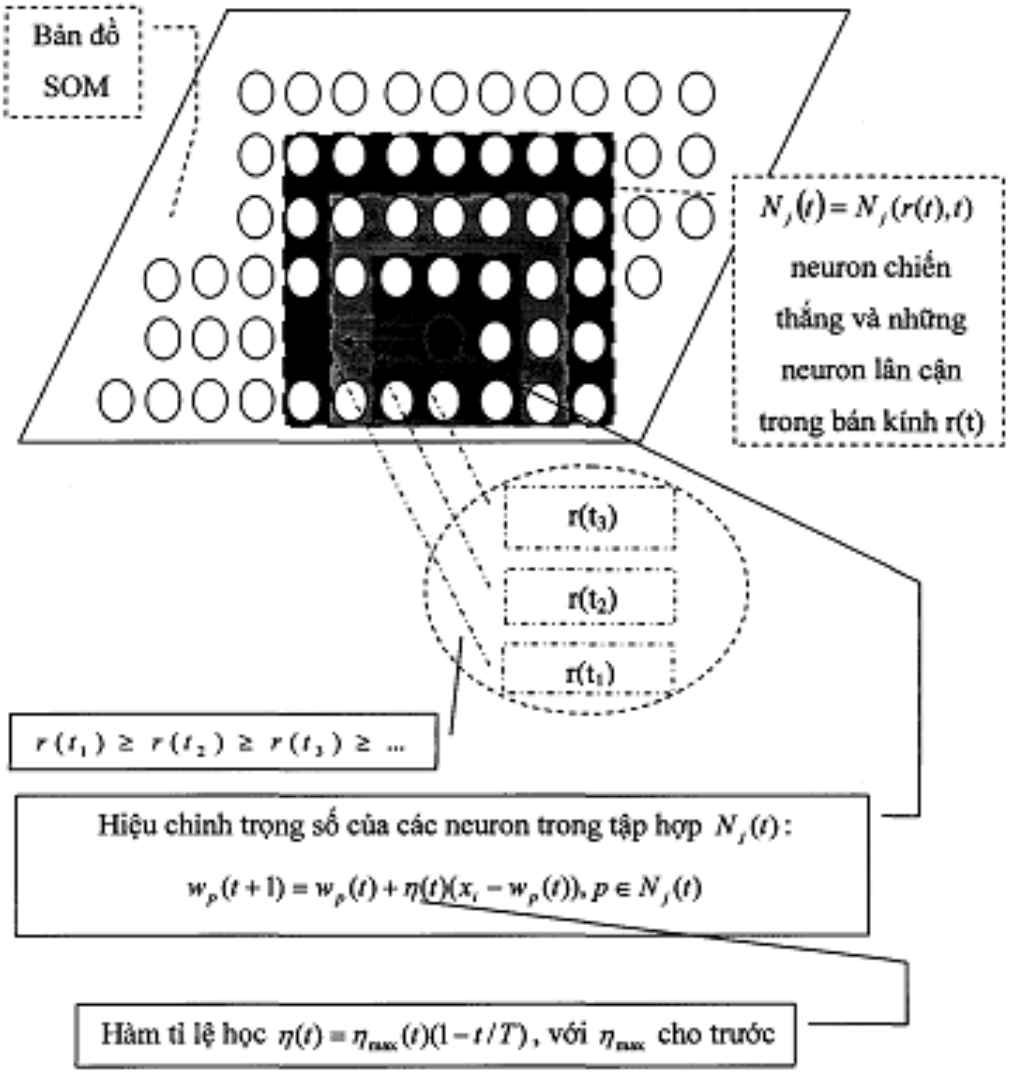
Kết quả: Mạng SOM sau quá trình học.



NỘI DUNG THUẬT TOÁN BẢN ĐỒ TỰ TỔ CHỨC



CƠ CHẾ XÁC ĐỊNH NEURON CHIẾN THẮNG VÀ NHỮNG NEURON LÂN CẬN



CƠ CHẾ HỌC CỦA NEURON CHIẾN THẮNG VÀ NHỮNG NEURON LÂN CẬN

3. PHƯƠNG PHÁP PHÂN TÍCH NGỮ ĐOẠN.

Để tìm kiếm một số vốn ngữ đoạn đặc trưng cho các văn bản trong ngữ liệu chúng ta cần xác định những dạng trung tâm của ngữ đoạn phổ quát nhất.

3.1 Cơ sở phân tích ngữ đoạn.

3.1 .1 Cấu trúc Đề - Thuyết.

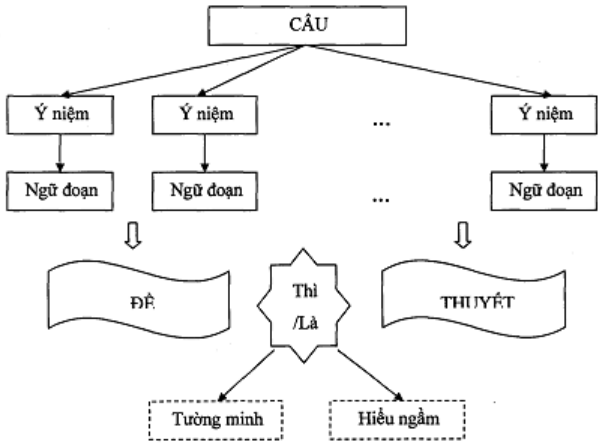
Mathesius (trường phái Prague, 1929) cho rằng ngữ pháp truyền thống chỉ phân tích hình thức chứ không phân tích ngữ nghĩa. Do vậy, trường phái này đã đưa ra khái niệm ngữ pháp chức năng. Mathesius chia câu thành Đề và Thuyết. Đề là cái được nói đến, Thuyết là cái nói về Đề. Lý thuyết này rất có triển vọng trong việc phân tích ngữ pháp tiếng Việt

C. Thompson (1965) phát hiện ra rằng câu trong tiếng Việt được xây dựng trên cấu trúc Đề -Thuyết. Trong tiếng Việt không có chủ ngữ ngữ pháp mà chỉ có logic tương ứng với Sở Đề của câu.

3.1.2 Những phương tiện đánh dấu sự phân chia Đề -Thuyết.

Quan hệ giữa Đề - Thuyết hết sức đa dạng. Đó là những mối quan hệ logic, những mối quan hệ về nghĩa, được đánh dấu bằng những phương tiện ngữ pháp nhưng không thể qui chế hóa vào những khuôn mẫu cứng nhắc.

"Thì" và "là": để dánh dấu chỗ câu phân chia thành hai phần Đề và Thuyết, tiếng Việt dùng trong hai tiểu tố: "thì " và "là". Đây là hai công cụ quan trọng nhất của cú pháp Tiếng Việt. Biên giới giữa Đề và Thuyết của một câu là chỗ nào có hai tiểu tố trên, hoặc có thể hiểu ngầm là hai tiểu tố trên mà cấu trúc cú pháp của câu không bị phá vỡ hay biến đổi, và ý niệm của câu vẫn được giữ nguyên.

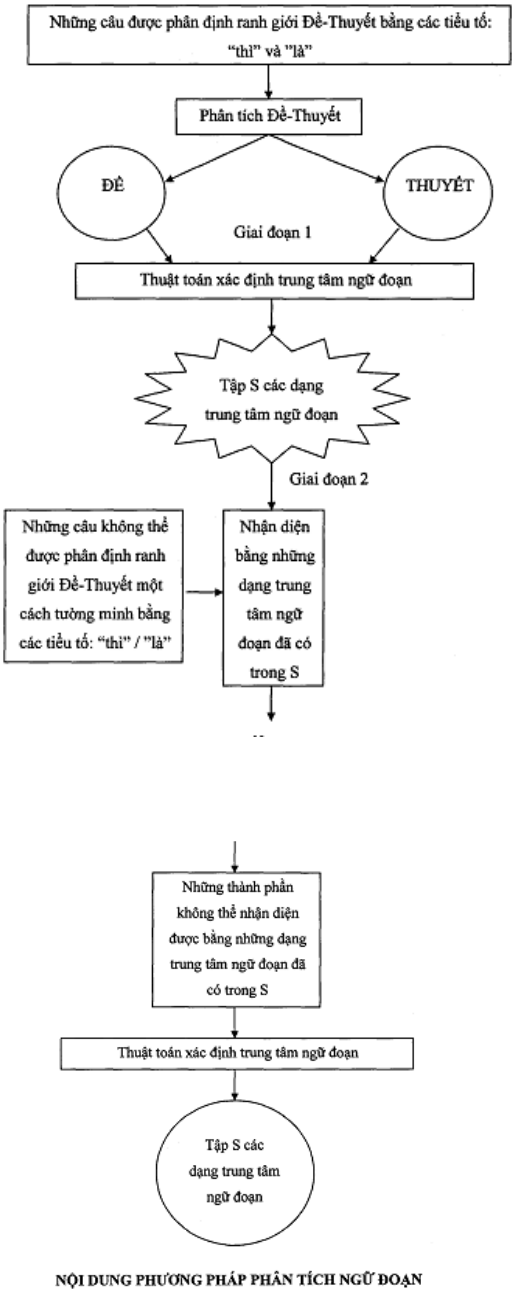


CẤU TRÚC ĐỀ - THUYẾT

3.1.3 Trắc nghiệm lược bỏ- mở rộng văn cảnh.

Trung tâm của ngữ đoạn là yêu tố duy nhất có quan hệ ngữ pháp và ngữ nghĩa vượt ra ngoài biên giới của ngữ đoạn. Do vậy, con đường trực tiếp nhất để xác định trung tâm của ngữ đoạn là tìm xem yếu tố nào của nó có được quan hệ như thế. Để thực hiện điều này, phải thử lược bỏ từng thành phần trong ngữ đoạn, và tìm kiếm sự phân bỗ của ngữ đoạn sau thao tác lược bỏ trong những văn cảnh khác nhau. Trung tâm của ngữ đoạn chính là thành phần không thể lược bỏ được.

3.1.4 Mô hình phân tích ngữ đoạn.



NỘI DUNG PHƯƠNG PHÁP PHÂN TÍCH NGỮ ĐOẠN

3.2 Thuật toán xác định trung tâm ngữ đoạn.

Thuật toán xác định trung tâm ngữ đoạn dựa trên trắc nghiệm lược bỏ và mở rộng văn cảnh được trình bày sau đây chỉ nhằm tìm những dạng trung tâm ngữ đoạn có kết cấu từ hai từ vựng trở nên. Phương pháp này cho kết quả phụ thuộc vào khối lượng ngữ liệu trong đó các văn cảnh hiện diện

Đầu vào: Tập hợp các câu của toàn bộ ngữ liệu văn bản. Các câu này được phân rã sơ bộ dựa trên các dấu phẩy (,) ngăn cách giữa các ngữ đoạn lớn. Tập hợp tất cả những dạng ngữ đoạn được phân rã sẽ là dữ liệu đầu vào cho thuật toán.

Đầu ra: Tập hợp S tất cả những dạng trung tâm ngữ đoạn.

- ***Bước 1***: S={}.
- ***Bước 2***: Dùng 2 tiểu tố "thì " và "là" phân tích thành hai phần Đề và Thuyết tất cả những dạng ngữ đoạn có thể .

Gọi R là tập hợp tất cả những dạng ngữ đoạn đầu vào còn lại chưa phân tích được.

Gọi D là tập hợp tất cả những dạng ngữ đoạn làm Đề phân tích được, đây là những danh ngữ hoặc kết cấu có chức năng tương đương danh ngữ. Gọi T là tập hợp tất cả những dạng ngữ đoạn làm Thuyết phân tích được. Gọi C=R + T

- ***Bước 3***:Với mỗi dạng ngữ đoạn s $\in$ D, Thực hiện:

- ***B3.a*** Mở rộng văn cảnh cho dạng ngữ đoạn s', với s' được dẫn xuất từ s bằng cách lược bỏ một từ cuối trong cấu trúc.

Mở rộng văn cảnh cho s' có nghĩa là tìm sự phân bố của s' trong tất cả mọi văn cảnh của ngữ liệu.

- ***B3.b*** Nếu số lượng văn cảnh chứa s' tìm được lớn hơn một ngưỡng nào đó (trong đề tài sử dụng ngưỡng là 10) thì coi như s' là một dạng trung tâm ngữ đoạn S=S+{s'}. Dừng bước 3 đối với s hiện hành, quay trở lại bước 3 với s khác.
- ***B3.c*** Quay lại bước 3.a, cho đến khi s' không còn có thể được cấu trúc bởi 2 từ trở lên thì dừng bước 3 đối với s hiện hành.

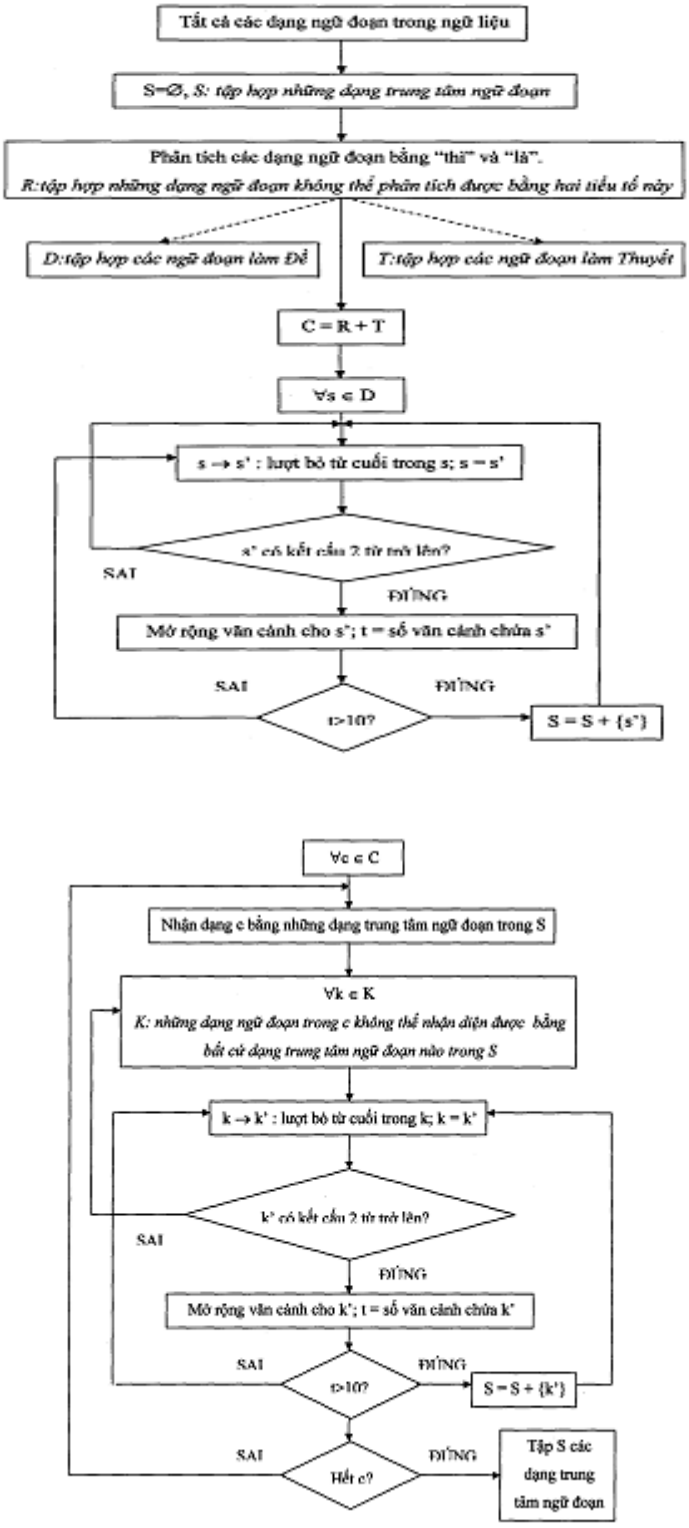
  Quay trở lại bước 3 đối với s khác.

- ***Bước 4***: Dùng những dạng trung tâm ngữ đoạn của S để phân rã các dạng ngữ đoạn trong tập C.

$\forall_c \in C$, phân rã c thành những dạng ngữ đoạn dựa trên các dạng trung tâm ngữ đoạn đã có trong S. Sự phân rã c thực hiện như sau:

- c được xem như chứa những dạng trung tâm ngữ đoạn đã biết hoặc chưa biết

❖ Những dạng ngữ đoạn thành phần trong kết cấu của c không thể nhận diện được bằng bất cứ dạng trung tâm ngữ đoạn nào đã biết trong S thì sử dụng những thao tác ở bước 3 đối với những dạng ngữ đoạn thành phần chưa biết này.



THUẬT TOÁN XÁC ĐỊNH TRUNG TÂM NGỮ ĐOẠN

3.3 Minh họa thuật toán.

- ***Bước 1***: Đầu vào của thuật toán là 84,343 dạng ngữ đoạn thu được từ sự phân rã sơ bộ các câu của 5,325 văn bản toàn văn .Thuật toán sẽ tiến hành tìm kiếm những dạng trung tâm ngữ đoạn của những dạng ngữ đoạn này.
- ***Bước 2***:Tập D gồm các dạng ngữ đoạn làm Đề.
- ***Bước 3***: ví dụ chọn một dạng ngữ đoạn s trong D là "xây dựng chủ nghĩa xã hội", đây là một dạng ngữ đoạn làm Đề có thể phân tích từ những câu như:

  "Xây dựng chủ nghĩa xã hội là một cuộc đấu tranh cách mạng phức tạp"

  "Xây dựng chủ nghĩa xã hội là xây dựng cuộc sống ấm no và hạnh phúc cho nhân dân" …

  - ***Bước 3a***: s'= "xây dựng chủ nghĩa xã ".
  - ***Bước 3b***: không tìm ra văn cảnh nào chứa s'.
  - ***Bước 3c***: quay lại bước 3a.
  - ***Bước 3a***: s'= " Xây dựng chủ nghĩa".
  - ***Bước 3b***: tìm ra 2 văn cảnh chứa s', Ví dụ: "Nhân dân Liên xô vừa xây dựng chủ nghĩa cộng sản ở nước mình". Số văn cảnh chứa s' tìm được ít hơn ngưỡng là 10, vậy s' không phải là một dạng trung tâm ngữ đoạn.
  - ***Bước 3c:*** quay lại bước 3a.
  - ***Bước 3a***: s'="Xây dựng chủ".
  - ***Bước 3b***: tìm ra 3 văn cảnh chứa s', ví dụ: " Xây dựng chủ trương chung". Số văn cảnh chứa s' tìm được ít hơn ngưỡng là 10, vậy s' không phải là một dạng trung tâm ngữ đoạn.
  - ***Bước 3c***: dừng bước 3 đối với s. Thực hiện bước 3 đối với s khác.

    …

- ***Bước 4***: giả sử S có một dạng trung tâm ngữ đoạn là " xây dựng" .Dùng những dạng trung tâm ngữ đoạn này để phân rã những dạng ngữ đoạn khác trong C.

  Ví dụ: c= " Xây dựng xã hội mới" có chứa một dạng trung tâm ngữ đoạn đã biết là "xây dựng " và một dạng ngữ đoạn chưa nhận diện được là " xã hội mới"

- ***Bước 3a***: s'= "xã hội",
- ***Bước 3b***: tìm ra 3, 101 văn cảnh chứa s'. Ví dụ: "lịch sử phát triển xã hội" .Số văn cảnh chứa s' tìm được nhiều hơn ngưỡng là 10, vậy s' là một dạng trung tâm ngữ đoạn.
- ***Bước 3c***: dừng bước 3 đối với s.

CHƯƠNG 4: QUẢN LÝ VÀ KHAI THÁC TRI THỨC TRÊN BẢN ĐỒ VĂN BẢN TỰ TỔ CHỨC.

4.1 GOM NHÓM TRÊN BẢN ĐỒ VĂN BẢN TỰ TỔ CHỨC.

Khi số lượng đơn vị trên bản đồ SOM lớn, tiến trình gom nhóm ngay trên bản đồ sẽ được thực hiện nhằm phục vụ các mục đích khai thác sau đó.

Như đã trình bày ở những phần trước, SOM tỏ ra đặc biệt thích hợp cho mục đích xây dựng bản đồ bởi vì những đặc tính nổi trội của nó trong việc trình bày dữ liệu. SOM tạo ra một tập hợp các vector nguyên mẫu biểu diễn tập dữ liệu và thực hiện một phép chiếu bảo toàn topo cho những mẫu không gian đầu vào n-chiều lên một bảng ít chiều hơn, thông thường là bản đồ 2- chiều. Bản đồ này là một mặt phẳng hiển thị thích hợp để trình bày những đặc trưng khác nhau của SOM, chẳng hạn cho những cấu trúc nhóm.

Tuy nhiên, các hiển thị trực quan như vậy chỉ có thể được dùng để cảm nhận về những thông tin định tính. Để tạo ra những thông tin tóm lược- những mô tả định lượng về đặc tính của dữ liệu- các đơn vị bản đồ cần được gom nhóm để xử lý một cách có hiệu quả. Ở đây không ngoài mục đích tìm kiếm những cách gom nhóm tốt nhất cho dữ liệu mà là thực hiện một sự gom nhóm có thể, để làm bộc lộ những đặc trưng về cấu trúc của dữ liệu, để phục vụ cho mục đích Khai phá dữ liệu văn bản.

4.1.1 Những khoảng cách tiêu chuẩn dùng trong gom nhóm.

1. Những khoảng cách bên trong nhóm

- Khoảng cách trung bình: $||x_i - x_j||$

$$S_a = \frac{\sum_{i,j} || xi - xj ||}{Nk(Nk-1)}$$

- Khoảng cách lân cận gần nhất:

$$S_{nn} = \frac{\sum_i \min i\{|| xi - xj ||\}}{Nk}$$

- Khoảng cách tâm:

$$S_c = \frac{\sum_i || xi - ck ||}{Nk}$$

2. Những khoảng cách giữa các nhóm:

- Liên kết đơn:

$$d_s = \min_{i,j} \{|| x_i - x_j ||\}$$

- Liên kết hoàn toàn:

$$d_{co} = \max_{i,j} \{|| x_i - x_j ||\}$$

- kết trung bình:

$$d_a = \frac{\sum_{i,j} || xi - xj ||}{NkNl}$$

- Liên kết tâm:

$$d_{ce} = || c_k - c_j ||$$

Các thuật toán gom nhóm:

Những thuật toán gom nhóm được phân thành hai loại chính: *gom nhóm phân cấp* và *gom nhóm phân hoạch*. Những thuật toán gom nhóm phân cấp lại được chia thành hai loại: gom nhóm tích hợp (agglomerative algorithms) và gom nhóm chia nhỏ (divisive algorithms). Những thuật toán gom nhóm tích tụ thường bao gồm các bước sau:

1. Khởi tạo: gán mỗi vector cho một nhóm.
2. Tính toán khoảng cách giữa tất cả các nhóm.
3. Trộm hai nhóm gần nhau lại.
4. Trở lại bước 2 cho đến khi chỉ còn một nhóm duy nhất.

Nói cách khác, các mục dữ liệu được trộn với nhau để hình thành nên cây phân cấp nhóm. Cây phân cấp nhóm có thể dùng để diễn giải cho cấu trúc của dữ liệu và xác định số lượng nhóm.

Những thuật toán gom nhóm phân hoạch thì chia một tập dữ liệu thành một số các nhóm và tìm cách tối thiểu hóa một số tiêu chuẩn hoặc hàm lỗi.

Thuật toán dựa trên các bước sau:

1. Xác định số lượng nhóm.
2. Khởi tạo các trung tâm nhóm.
3. Tính toán (cập nhật) các trung tâm nhóm.
4. Nếu tình trạng phân hoạch không còn thay đổi thêm được nữa thì dừng; ngược lại, trở về bước 3.

Nếu không tìm thấy trước số lượng nhóm, thuật toán phân hoạch có thể đưa ra giả sử về số lượng nhóm này, thường thì từ 2 nhóm đến $\sqrt{N}$ nhóm, với N

là số lượng mẫu trong tập dữ liệu. Trong trường hợp này thuật toán sẽ lặp đi lặp lại để tìm số lượng nhóm tốt nhất cho sự gom nhóm phân hoạch.

4.1.2 Gom nhóm trên SOM.

Giả sử ban đầu rằng mỗi đơn vị bản đồ là một nhóm. Áp dụng thuật toán gom nhóm tích tụ với phép trộn được xác định bởi một trong hai tiêu chuẩn sau:

A. Chỉ số Davies- Bouldin: tính chỉ số này cho hai nhóm quan tâm, nếu chỉ số này lớn hơn 1 thì tiến hành trộn hai nhóm.

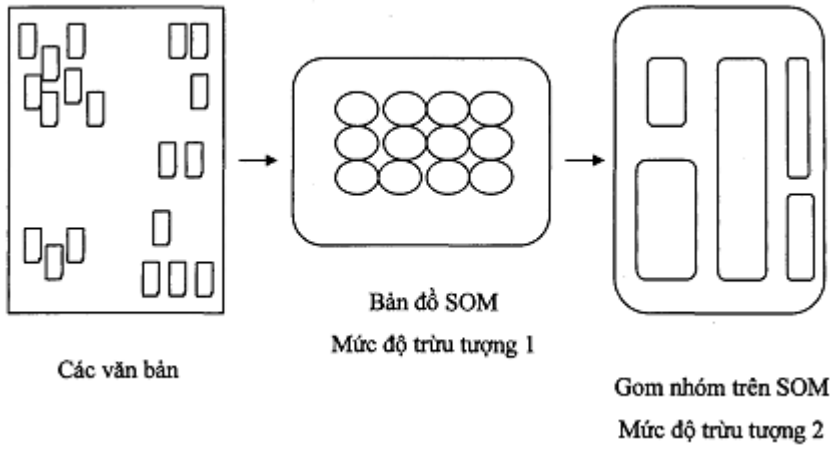
Chỉ số Davies-Bouldin được tính như sau:

$$\frac{1}{C}\sum_{k=1}^{C}\max_{l\neq k}\left\{\frac{S_c(Q_k)+S_c(Q_l)}{d_{ce}(Q_k,Q_l)}\right\}$$

trong đó C là số lượng nhóm.

B. Khoảng cách giữa hai nhóm: nếu khoảng cách $d_s(Q_k,Q_l)$ lớn hơn tổng của các khoảng cách trung bình $S_{nn}(Q_k) + S_{nn}(Q_l)$ giữa các điểm trong hai nhóm thì tiến hành trộn hai nhóm.

4.1.3 Thuật toán gom nhóm.



4.2. GÁN NHÃN BẢN ĐỒ.

Khám phá tri thức trên bản đồ văn bản về bản chất là một quá trình khai thác nhãn được gán cho những đơn vị và những vùng bản đồ. Các nhãn bản đồ này là những mô tả nội dung được xây dựng ở cấp độ khái quát cao, trên cơ sở các nhãn của văn bản.

Giả sử rằng mỗi văn bản được kết hợp với một tập hợp các nhãn, và mỗi nhãn tương ứng với một từ khóa trong văn bản.

Phương pháp LabelSOM để gán nhãn cho các đơn vị bản đồ, phương pháp này phân tích những thành phần của vector tham chiếu và chọn làm nhãn những

từ tương ứng với những thành phần của vector tham chiếu có độ lệch nhỏ nhất theo định nghĩa.

Phương pháp gán nhãn cho các đơn vị và các vùng bản đồ văn bản trong mô hình WEBSOM dựa trên việc chọn lựa những từ vựng theo các độ đo tỉ lệ tần số xuất hiện.

Việc ứng dụng ngữ đoạn vào gán nhãn bản đồ đã được nhiều tác giả tiên liệu trong thời gian dài, xuất phát từ những nghiên cứu về vấn đề khám phá và phát hiện các cụm từ trong văn bản. (Turney, 1999) đã chỉ rõ việc ứng dụng ngữ đoạn trong năm lĩnh vực quan trọng, trong đó có lĩnh vực gán nhãn cho các bản đồ văn bản. (Feldman, 1998) đưa ra phương pháp gán nhãn bằng cách phát sinh tự động một số ngữ đoạn dựa trên các từ khóa và những từ vựng hiện diện trong văn bản theo một số qui tắc cú pháp đơn giản.

Thuật toán:

1. Gọi tập hợp các văn bản trong ngữ liệu là K_0
2. Đối với một đơn vị bản đồ (hay một vùng bản đồ) i, gọi tập hợp những văn bản của nó là khối ngữ liệu K_i.
3. Áp dụng thuật toán phân tích ngữ đoạn để tìm các dạng trung tâm ngữ đoạn K_0. (Thông thường không cần thực hiện bước này do có thể sử dụng lại kết quả đã có từ giai đoạn mã hóa văn bản, nếu mã hóa được dựa trên ngữ đoạn).
4. $\forall$ s, Tính giá trị đại lượng Z của s trên K_1 so với K_0 . Nếu Z >2.58, s là trung tâm ngữ đoạn khóa của K_1. Sử dụng s làm nhãn của i.
5. Quay lại bước 2, thực hiện gán nhãn cho những đơn vị (vùng) bản đồ khác. Thuật toán dừng khi đã gán nhãn cho tất cả các đơn vị (vùng) bản đồ.

4.3 CƠ CHẾ TRÌNH BÀY BẢN ĐỒ VĂN BẢN.

Đề tài dùng các kỹ thuật web để trình bày bản đồ văn bản trong mục đích minh họa. Việc xây dựng những phương pháp đồ họa hiệu quả để trình bày bản đồ không nằm trong phạm vi của đề tài.

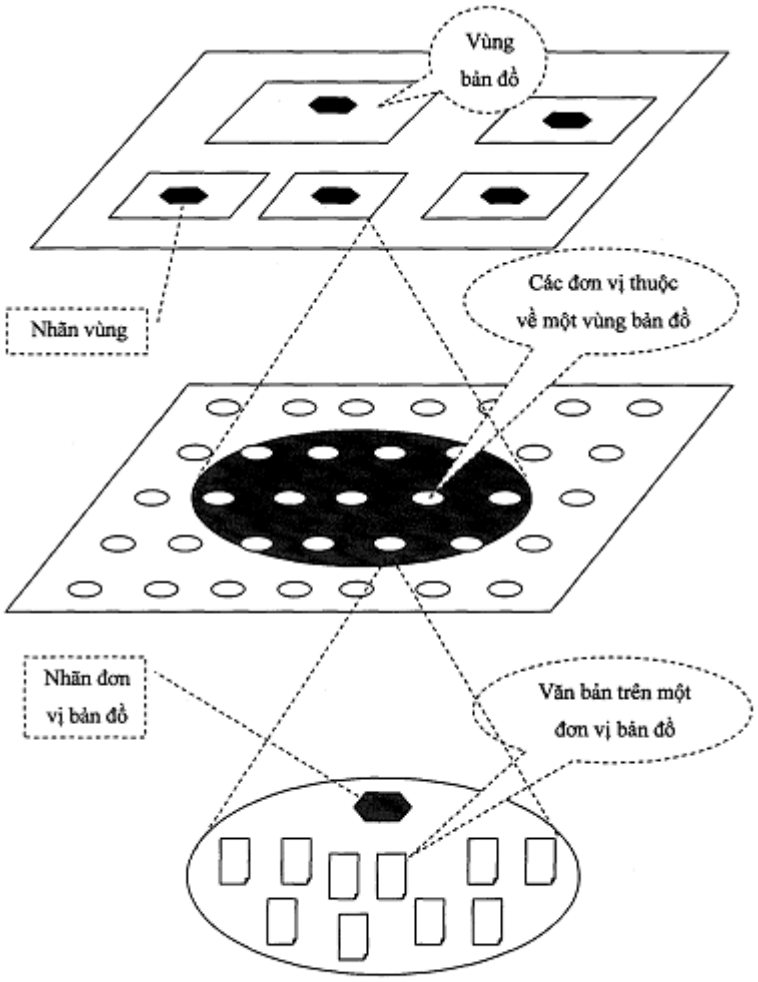
Bản đồ được trình bày theo hai dạng: một cách nhìn bao quát ghi nhận những đơn vị bản đồ có sự phân bố dữ liệu, bản đồ đã được gom nhóm thành những vùng lớn nhỏ khác nhau.

Trình bày bản đồ theo cấu trúc phân cấp chủ đề- nội dung:

- Cấp 0: bản đồ
- Cấp 1: vùng bản đồ,
- Cấp 2: đơn vị bản đồ,

- Cấp 3: văn bản.

Ở mỗi cấp trình bày, hiển thị tập nhãn phản ánh chủ đề của nhóm dữ liệu thuộc cấp đó.



MÔ HÌNH HIỂN THỊ TRỰC QUAN BẢN ĐỒ

Chương 5: KẾT LUẬN

Khai phá dữ liệu văn bản với bản đồ tự tổ chức SOM có thể thực hiện trong thực tiễn để giải quyết những vấn đề có liên quan đến các ngữ liệu văn bản lớn. Mô hình tổng quát đã được xác lập và nghiên cứu nhưng cần phải có những đóng góp mới để phù hợp với mỗi ngôn ngữ riêng biệt, đặc biệt đối với Tiếng Việt, một ngôn ngữ đơn lập khác loại hình với các tiếng châu Âu đã được nghiên cứu nhiều trong lĩnh vực này.

Đề tài đã nghiên cứu và triển khai thực nghiệm toàn bộ mô hình Khai phá dữ liệu văn bản, bao gồm tất cả các giai đoạn có liên quan: tiền xử lý –bao hàm năm phương pháp lựa chọn đặc trưng, mã hóa văn bản, giảm chiều vector văn bản, thuật toán bản đồ tự tổ chức SOM, gom nhóm trên bản đồ, gán nhãn các vùng và đơn vị bản đồ, cơ chế hiển thị bản đồ. Các kết quả đạt được có thể cho phép kết luận về tính khả thi của mô hình Khai phá dữ liệu văn bản với bản đồ tự tổ chức trong tiếng Việt

Từ kết quả của đề tài, những hướng nghiên cứu sau có thể tiếp tục:

1. Khám phá và quản lý tri thức trên bản đồ văn bản.
2. Kết hợp sử dụng bản đồ với các hệ thống tìm kiếm thông tin IR và cơ chế tìm kiếm sàng lọc và sắp xếp kết quả tìm kiếm.
3. Xây dựng các kĩ thuật đồ họa cao cấp và thuật toán để tô màu và trình bày trực quan bản đồ có hiệu quả
4. Nghiên cứu các phương pháp gom nhóm trên bản đồ và các bảng phân ngành chủ đề giải quyết vấn đề phân nghành văn bản.
5. Sử dụng bản đồ văn bản như một bộ lọc chủ đề để phân loại các văn bản khi chúng mới xuất hiện, hoặc phát hiện những chủ đề mới đang dần dần hình thành trong kho dữ liệu. Đặc biệt, bộc lọc có thể sử dụng trong các mục đích an ninh để theo dõi thông tin thu thập (Thư điện tử, Fax, …) những thông tin nhạy cảm khi bi sàng lọc sẽ được cảnh báo tự động cho các hệ thống theo dõi, phân loại, và thông báo cho các hệ thống truy tìm nguồn gốc khác.

TÀI LIỆU THAM KHẢO

A.Sách

[1].Cao Xuân Hạo, *Tiếng Việt*: *mấy vấn đề về ngữ âm, ngữ pháp, ngữ pháp, ngữ nghĩa*, NXB Giáo dục, 1998.752 trang.

[2].Cao Xuân Hạo,*Tiếng Việt: sơ thảo ngữ pháp chức năng*, quyển 1, NXB khoa học xã hội, 1991. 254 trang.

[3].Nguyễn Đức Dân, Đặng Thái Minh, *thống kê ngôn ngữ học: một số ứng dụng*, NXB Giáo dục, 1999. 220 trang.

B. Luận văn

[4]. **Nguyễn Thị Thanh Hà, Nguyễn Trung Hiếu.***Hệ thống tìm kiếm tiếng Việt.* Giáo viên hướng dẫn: Thạc Sĩ Trần Thái Minh

[5]. **Võ Hồ Bảo Khanh**, *Xây dựng bộ ngữ liệu Tiếng Việt*.Giáo viên hướng dẫn: Tiến sĩ Hồ Quốc Bảo

[6].**Nguyễn Đức Cường**,*Tổng quan về khai khoáng dữ liệu*,Trường ĐH Bách Khoa Tp Hồ Chí minh, Khoa Công Nghệ Thông Tin

[7]. **Nguyễn Thị Phương Thảo**,*Ứng dụng Data Mining trong phân tích dữ liệu thống kê.*Giáo viên Hướng Dẫn: Thạc sĩ Nguyễn Trọng Tuấn

[8].**Hoàng Hải Xanh**,*Các Kỹ thuật phân cụm dữ liệu trong Data Mining*;Giáo viên hướng dẫn: Hoàng Xuân Huấn.