PHÂN CỤM TRỰC QUAN TẬP CÁC BÀI BÁO KHOA HỌC THEO MÔ HÌNH NGUYÊN TỬ TRONG KHÔNG GIAN BA CHIỀU

Visually clustering research papers using an atom model in 3D space

Nghị Vĩnh Khanh[1]

Tóm tắt

Bài báo này đề xuất một cách tiếp cận mới để xây dựng một hệ thống phân cụm trực quan bằng hình ảnh 3D cho việc phân nhóm các bài báo khoa học bằng cách sử dụng kết hợp hai kỹ thuật trong lĩnh vực trí tuệ nhân tạo là SOM và k-means. Cụ thể, SOM sẽ đóng vai trò cho việc cung cấp một hình ảnh trực quan để quyết định tham số K cho thuật toán k-means tiếp theo. Một thiết kế đồ thị sẽ thể hiện các cụm mà mỗi cụm được đại diện bởi một hạt nhân (tâm của cụm) và các điện tử (các bài báo) bao quanh. Các điện tử sẽ quay quanh hạt nhân bằng các lực hấp dẫn. Bên cạnh đó, chúng tôi sử dụng kỹ thuật ArcBall trong lĩnh vực đồ họa máy tính ba chiều để hỗ trợ sự tương tác người dùng. Dựa trên hệ thống này, người dùng có thể thực hiện đánh giá sự thống nhất về cấu trúc cụm theo cách đơn giản hơn các phương pháp trước đây.

Từ khóa: phân cụm, trực quan hóa, trí tuệ nhân tạo, đồ họa ba chiều, tương tác người dùng.

Abstract

This paper proposes a new approach to construct a visually clustering system with 3D image for scientific papers by using two combined techniques in the field of artificial intelligence, which are SOM and k -means. Specifically, the SOM will play an important role in providing a visual image in order to determine parameter K for k-means algorithm in the next step. A graph layout is designed to show the clusters, each of which is represented by an atomic nucleus (the center of cluster) and electron (the papers) around. The electron will orbit the nucleus by the force of gravity. In addition, ArcBall techniques (in 3D computer graphics field) are used to support user interaction. Based on this system, users are able to evaluate the unification of Cluster's structure in a simpler way than in the previous ones.

Key words: clustering, visualization, artificial intelligence, 3D computer graphic, user interaction.

1. Giới thiệu

Sự phân cụm là một trong những kỹ thuật rất cần thiết cho việc khám phá tri thức nhân loại. Nó giúp cho chúng ta tách các nhóm đối tượng từ tập dữ liệu dựa trên các đặc tính tương đồng trong nhóm. Ngày nay, các kỹ thuật phân cụm được sử dụng rộng rãi trong các ứng dụng như khai phá dữ liệu, xử lý ảnh, nhận dạng mẫu, thống kê, tin sinh học và các lĩnh vực khác. Bên cạnh đó, áp dụng các kỹ thuật trực quan trong việc phân tích cụm dữ liệu rất quan trọng trong việc thể hiện xu hướng của các tập dữ liệu, nó cho chúng ta cái nhìn tổng quan cũng như sự hiểu biết chi tiết về tập dữ liệu. Hiện nay, nhiều nghiên cứu đã và đang tập trung về vấn đề phân tích trực quan các cụm rất thành công như Grand Tour, OPTICS, HD-EYE, H-BLOB, Star Coordinate (Ankerst, Mihael; Grinstein, Georges; Keim, Daniel;, 2002), SOM-based techniques (Kohoren, 1997), HOV3 (Zhang, Ke-Bing; Orgun, Mehmet A; Zhang, Kang;, 2006).

Trong bài báo này, chúng tôi đề xuất một thiết kế đồ họa trực quan ba chiều gọi là mô hình cấu trúc nguyên tử cho việc phân cụm dữ liệu. Mô hình này sử dụng giải thuật SOM để ước lượng số các cụm cần được tách ra từ tập các tài liệu nói chung và các bài báo khoa học nói riêng (được viết bằng tiếng Anh) có định dạng PDF. Sau đó, dựa trên mô hình không gian vector, cụ thể là vector tf-idf, chúng tôi sử dụng thuật toán k-means để tách tập dữ liệu thành k nhóm. Cuối cùng các cụm sẽ được trực quan hóa thành dạng các cấu trúc nguyên tử trong không gian ba chiều. Để đơn giản, chúng tôi tổ chức một nguyên tử có tối đa năm mức năng lượng được tính bằng độ tương đồng Cosin giữa các vector điện tử và hạt nhân.

Cách tiếp cận của bài báo này sẽ giải quyết được các vấn đề sau:

- Thứ nhất, hiển thị được tập các vector nhiều chiều (lớn hơn 1000) trong không gian ba chiều, trong đó thể hiện rõ sự phân phối dữ liệu, mối quan hệ giữa mỗi tâm của các cụm cũng như giữa các bài báo khoa học với nhau.

[1] *Thạc sĩ, Ban Phát triển Hệ thống CNTT, Trường ĐH Trà Vinh*

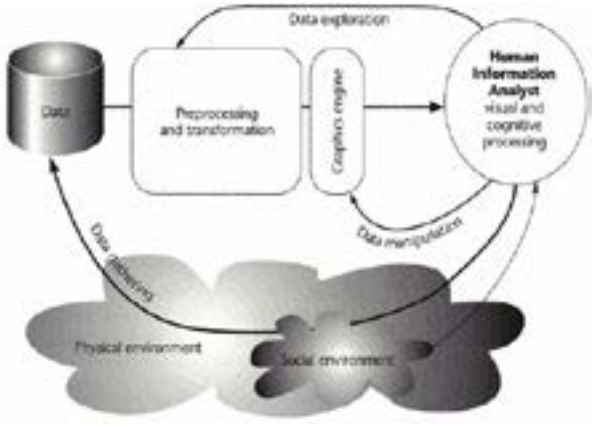
- Thứ hai, phương pháp tiếp cận của chúng tôi tránh được lựa chọn tùy ý các cụm k bởi sự kết hợp của SOM và K-means. Do đó, mô hình này cung cấp cho người dùng một phương pháp trực quan có mục đích và có hiệu quả vào việc phân tích cluster.

- Thứ ba, có thể vượt qua giới hạn của không gian khi so sánh với các phương pháp 2D trước đó bằng cách sử dụng kỹ thuật arcball để tương tác. Nó tạo cho chúng ta cảm giác nhập vai vào hệ thống và thao tác từng đối tượng giống như chơi game 3D hay sử dụng các hệ thống thực tế ảo - ví dụ CAVE (Cruz-Neira C; Sandin D.; DeFanti T.; Kenyon R.; Hart J.;, 1992).

Trong các phần sau, chúng tôi tổ chức cấu trúc bài báo như sau: phần 2 - tổng quan các kết quả nghiên cứu trước đây, phần 3 - phương pháp thực hiện, phần 4 - đánh giá kết quả và phần 5 - kết luận.

2. Tổng quan các kết quả nghiên cứu trước đây

Để xây dựng được bộ công cụ phân tích các cụm trực quan, nhiều kỹ thuật đã được nghiên cứu cho các quá trình biểu thị trực quan các đối tượng từ một tập dữ liệu lên màn hình máy tính. Point-based techniques, line-based technique, region-based technique, hierarchical techniques (Ward, Matthew; Grinstein, Georges; Keim, Daniel;, 2010) là những kỹ thuật phổ biến dùng để trực quan hóa các tập dữ liệu nhiều chiều. Tuy nhiên, hầu hết chúng đều gặp khó khăn khi trực quan các tập dữ liệu khá lớn cũng như dữ liệu có chiều của vector rất cao. Một vài kỹ thuật bị giới hạn trong việc cung cấp một sự nhận thức rõ ràng từ các dạng trực quan cho người dùng.



Hình 1. Quá trình trực quan thông tin (Ware, 2004)

Trong thập niên qua, nhiều kỹ thuật phân tích cụm trực quan đã được phát triển, chẳng hạn như Grand Tour, OPTICS, HD-EYE, H-BLOB, Fastmap, Star Coordinate, SOM-based techniques, HOV3,... Nhìn chung, các kỹ thuật này góp phần quan trọng trong việc phân tích các cụm và có thể giải quyết được các khía cạnh quan trọng của việc nhận thức trực quan:

- Trực quan các dữ liệu lớn và đa chiều;

- Cung cấp một cái nhìn tổng quan rõ ràng và chi tiết về cấu trúc cụm;

- Có độ phức tạp tính toán tuyến tính trên việc ánh xạ dữ liệu từ không gian chiều cao sang chiều không gian thấp hơn;

- Hỗ trợ tương tác động với các đại diện trực quan của cụm;

- Kết nối kiến thức liên quan của các chuyên gia vào lĩnh vực thăm dò vào cụm;

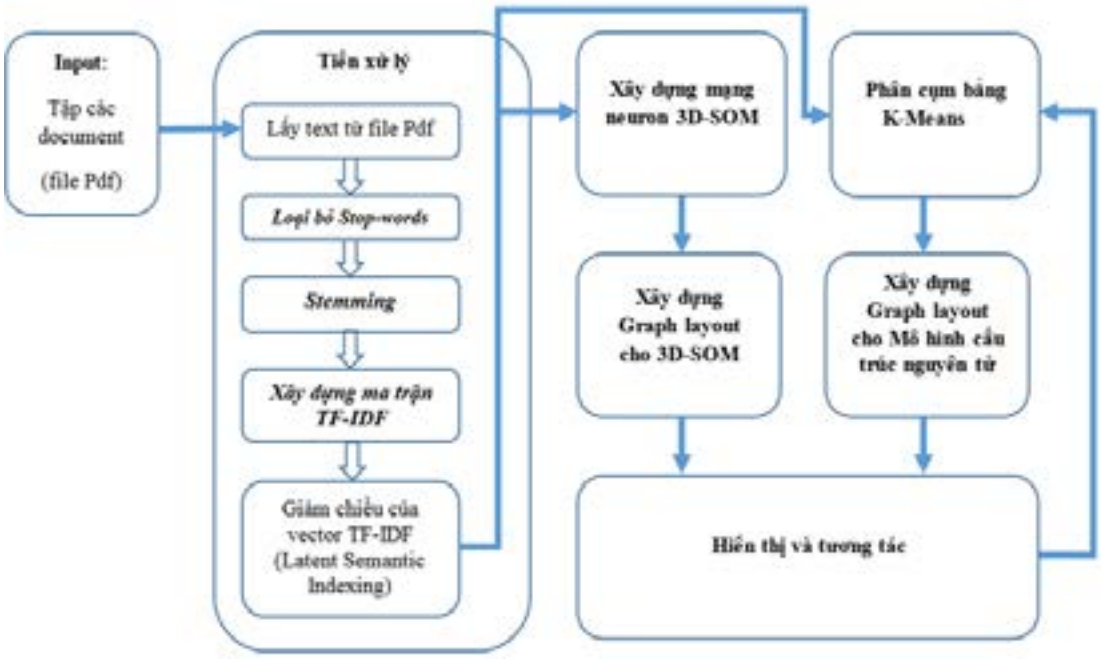
- Cho người dùng các chỉ dẫn có mục đích và chính xác của việc khảo sát/điều tra các cụm cũng như hợp lệ hóa các cụm chứ không phải chỉ đơn giản là thăm dò cụm ngẫu nhiên.

Hầu hết các kỹ thuật trên giải quyết được các yêu cầu này nhưng chúng vẫn còn hạn chế khi kích thước và chiều của tập dữ liệu khá lớn. Hơn nữa, một vài kỹ thuật trên gặp khó khăn khi cung cấp một cái nhìn tổng quan sáng sủa của cấu trúc của cụm cũng như mức độ dễ sử dụng dành cho các người dùng.

3. Phương pháp thực hiện

Như đã giới thiệu, để giải quyết các hạn chế trên, chúng tôi đề xuất giải pháp trực quan bằng một thiết kế đồ thị trực quan ba chiều có hỗ trợ tương tác dựa theo cấu trúc nguyên tử.

Các bước thực hiện như sau:



Hình 2. Quy trình xây dựng hệ thống

Bước tiền xử lý:

Sau khi tách các ký tự từ file định dạng PDF, nhiệm vụ tiếp theo là chế biến từ vựng. Về cơ bản, chúng ta cần ba hoạt động: loại bỏ các từ vô nghĩa hoặc từ không mang thông tin trong ngữ cảnh cần xem xét (stop-word), chuyển các từ về dạng gốc (steamming), và tính trọng số của từng từ so với các từ khác (term weighting).

Các bước loại bỏ "Stop Words" và "Steamming" sẽ giúp chúng ta giảm kích thước của tập từ vựng, do đó sẽ tiết kiệm được nguồn tài nguyên tính toán. Bởi vì tập các bài báo khoa học đầu vào được viết bằng tiếng Anh nên nó không khó để áp dụng giải thuật tìm gốc từ; cụ thể giải thuật Porter stemming (Porter, 1980) hiện được sử dụng rất hiệu quả cho một số ngôn ngữ như tiếng Anh mặc dù chưa hỗ trợ được nhiều ngôn ngữ trên thế giới. Chúng ta sẽ thật sự gặp khó khăn nếu tập các bài báo được viết bằng tiếng Việt vì cần có nhiều nghiên cứu chuyên sâu về ngôn ngữ tự nhiên của tiếng Việt để tìm được từ gốc của chúng.

Kết quả sau khi loại bỏ "Stop Words" và "Steamming" của tất cả các từ vựng trong tất cả các văn bản, ta sẽ xác định được một tập hợp duy nhất các từ vựng, gọi là Bag-Of-Word. Tiếp theo, chúng ta sẽ tính trọng số của các từ này (term weighting). Để xác định trọng số của mỗi từ vựng, chúng tôi sử dụng một công thức rất phổ biến để tính đại lượng Term Frequency Invert Document Frequency (TFIDF) (Manning, Christopher D.; Raghavan, Prabhakar; Schutze, Hinrich;, 2009)

$$tfidf(w) = tf \bullet \log\left(\frac{N}{df(w)}\right)$$

Trong đó: tf(w): term frequency (số lần từ w này xuất hiện trong một tài liệu), df(w): document frequency (số lượng tài liệu chứa đựng từ w này), N: Tổng số tài liệu.

Đại lượng tfidf(w) nói lên sự quan trọng của từ w trong tài liệu. Từ công thức này, chúng ta tiến hành tính giá trị của ma trận TFIDF. Trong đó, mỗi hàng là đại diện một tài liệu, các cột là giá trị tfidf của các từ trong tập Bag-Of-Words.

Bởi vì kích thước của ma trận TFIDF có thể rất lớn (bằng M tài liệu x N term trong Bag-Of-Words). Thực tế, nếu ta có một tập gồm 100 bài báo khoa học của cùng một lĩnh vực nghiên cứu và mỗi bài báo khoảng 10 trang thì ma trận TFIDF có thể có kích thước là 100x10000. Cần chú ý rằng nếu ta chỉ dùng các keyword hay chỉ xác định Bag-Of-Words từ trong phần Abstract của bài báo (nhằm rút gọn kích thước ma trận này) thì về mặt thống kê cũng như ngữ nghĩa sẽ đem lại một kết quả không chính xác cho sự khác biệt nội dung giữa các bài báo. Có nhiều phương pháp được dùng cho việc giảm kích thước của ma trận TFIDF, trong đó kỹ thuật Latent semantic indexing analysis - LSI

(Manning, Christopher D.; Raghavan, Prabhakar; Schutze, Hinrich;, 2009) được sử dụng khá phổ biến. Nó là một kỹ thuật thống kê nhằm cố gắn ước lượng cấu trúc nội dung được ẩn bên trong văn bản bằng cách sử dụng kỹ thuật đại số tuyến tính Singular-Value-Decomposition.

LSI rất hiệu quả trong việc giảm chiều của tập dữ liệu. Tuy nhiên, việc sử dụng kỹ thuật này sẽ gặp trở ngại khi ta muốn thực hiện truy vấn tìm kiếm từ trong ma trận TFIDF. Ví dụ từ hình trên, xét trường hợp ta có 6 văn bản (d1, d2, d3, d4, d5, d6) với Bag-Of-Words có chiều là 5 (ship, boat, ocean, voyage, trip). Sau khi sử dụng kỹ thuật LSI để giảm chiều từ 5 thành 2 thì các chiều mới là "1" và "2" sẽ không còn mang ý nghĩa tương ứng của "ship", "boat", "ocean", "voyage", "trip". Điều đó có nghĩa là chúng ta không thể thực thi truy vấn để tìm thuộc tính ban đầu, ví dụ từ "ship" trong ma trận đã giảm chiều. Thực tế, thì đã có nhiều nghiên cứu để giải quyết vấn đề này, tuy nhiên chúng tôi không đề cập đến do nằm ngoài phạm vi nghiên cứu của bài báo này.

	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1

the term-document matrix C

	d_1	d_2	d_3	d_4	d_5	d_6
1	−1.62	−0.60	−0.44	−0.97	−0.70	−0.26
2	−0.46	−0.84	−0.30	1.00	0.35	0.65

illustrates the documents in $(V^T)^T$ in two dimensions.

Hình 3. Rút gọn ma trận C từ 5D thành 2D bằng LSI

Sau quá trình tiền xử lý, chúng ta tiếp tục tiến hành mô hình hóa và trực quan các tài liệu dựa trên ma trận TFIDF. Cụ thể theo trình tự như sau:

- Xây dựng mạng Nơ ron nhân tạo Self-ogranizing Map ba chiều (3D-S.O.M).

- Xây dựng Graph layout cho 3D-S.O.M và hiển thị, từ đó giúp người dùng cân nhắc chọn số nhóm cần phân cụm, đây chính là thông số k dùng cho giải thuật phân cụm k-means.

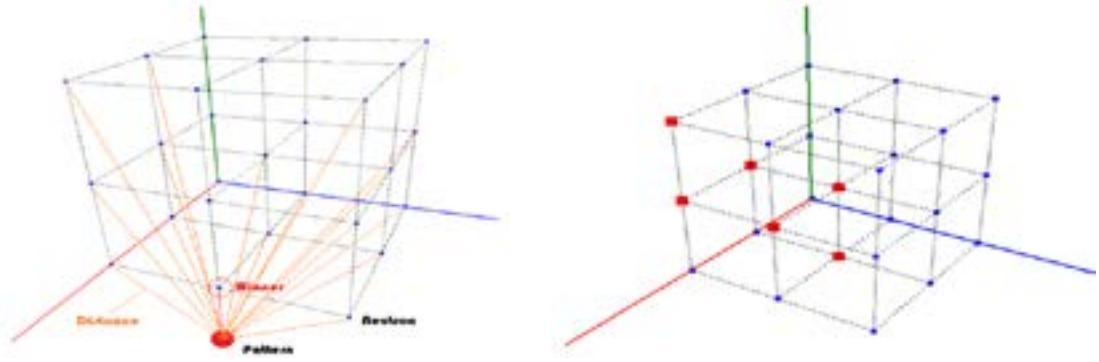
- Áp dụng giải thuật k-means để phân các vector trong ma trận TFIDF (đại diện cho mỗi bài báo khoa học) thành K cụm. Các bài báo trong mỗi cụm thì có ý nghĩa là chúng gần gũi, tính tương đồng gần nhau.

- Xây dựng graph layout cho mô hình.

Chi tiết sẽ được trình bày sau đây.

Xây dựng 3D-SOM

Chúng tôi xây dựng một lưới ba chiều để thể hiện các nơron. Mỗi nơron có tọa độ (X,Y,Z) , vector (có cùng chiều với các vector TFIDF của tập dữ liệu, cụ thể là cùng chiều với vector Bag-Of-Word) và trọng số. Ví dụ hình dưới là một lưới có 27 nơron (3x3x3).



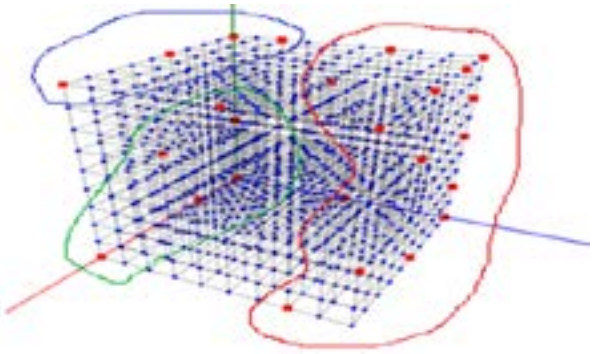
Hình 4. 3D SOM với 3x3x3 nơ ron (trái) và 6 winners (màu đỏ-phải)

Đầu tiên, chúng ta gán tọa độ duy nhất (lấy từ 3D-Grid) cho mỗi nơron. Trọng số của nơron W_{ijk}(d1,d2,d3,….,dn) có giá trị ngẫu nhiên trong khoảng (0,1).

Tiếp theo, chúng ta sẽ huấn luyện mỗi nơron từ tập hợp các mẫu (ma trận TFIDF) theo giải thuật SOM (Kohoren, 1997).

3D-SOM sẽ gặp hạn chế về vấn đề thời gian thực thi khi chiều của vector khá lớn cũng như độ ổn định của Mạng phụ thuộc vào số lần lặp. Tuy nhiên, nó có lợi thế là Mạng Self-Ogranizing Map có thể phân loại dữ liệu mà không cần phải huấn luyện lại Mạng khi mà nó đã ổn định.

Sau khi xây dựng graph layout và hiển thị 3D-SOM, dựa vào hình ảnh trực quan nhìn thấy được, chúng ta có thể ước lượng được giá trị K (số nhóm cần được phân cụm) cho giải thuật phân cụm K-means trong bước tiếp theo. Theo hình trên, ta có thể dễ dàng chỉ ra K = 3 cho tập dữ liệu.



Hình 5. Lưới 3D-SOM 10x10x10 của 25

Mô hình của chúng tôi sử dụng dạng cấu trúc nguyên tử cho việc hiển thị và tương tác, vì vậy chúng ta cần biết tâm của Cụm mà nó sẽ thể hiện như là hạt nhân của nguyên tử. Trong các phương pháp phân cụm, chúng ta thấy phương pháp phân hoạch bằng k-means (MacQueen, 1967) là phù hợp nhất bởi vì ta sau khi phân cụm, ta có thêm giá trị vector là tâm của cụm. Cần nói rõ thêm là hiện nay đã có nhiều giải thuật cải tiến của K-means (ví dụ k-means ++) nhưng vì để đơn giản nên chúng tôi chỉ sử dụng k-means.

Mô hình dạng cấu trúc nguyên tử cho việc phân cụm tập tài liệu:

Sau khi xây dựng 3D-SOM và K-means, chúng ta sẽ tiến hành xây dựng mô hình. Chúng ta qui ước như sau:

a) Mỗi cluster ω là một nguyên tử.

Mỗi nguyên tử là đại diện của một Cụm. Cụ thể, hạt nhân là Centroil $\vec{\mu}$ của Cụm W, các electron bao quanh hạt nhân là các vector $\vec{x}$ (vector TFIDF của một bài báo) trong cùng nhóm.

Centroil $\vec{\mu}$ của cluster W:

$$\vec{\mu}(w) = \frac{1}{|w|} \sum_{\vec{x} \in W} \vec{x}$$

- Khoảng cách giữa electron và hạt nhân của nó được đo bằng sự giống nhau về ngữ nghĩa giữa chúng (là hệ số cosin giữa 2 vector), còn gọi là năng lượng của electron

$$Engergy(\vec{x}, \vec{\mu}) = Distance(\vec{x}, \vec{\mu}) = \frac{\vec{x} \cdot \vec{\mu}}{|\vec{x}||\vec{\mu}|}$$

- Những electron có cùng mức năng lượng sẽ nằm trên cùng quỹ đạo và được phân bố đều trên bề mặt của một khối cầu có bán kính so với tâm của hạt nhân bằng mức năng lượng của nó khi hệ thống ở trạng thái không chuyển động.

- Mỗi electron sẽ có cùng kích thước quy ước

- Kích thước của nguyên tử = số electron * kích thước electron

b) Nếu chúng ta có k nguyên tử (clusters):

Sự phân bố của chúng sẽ được tính dựa trên kích thước của nó (cụ thể là số lượng electron – là các vector TFIDF của tài liệu) theo sau:

- Tất cả hạt nhân của các nguyên tử được bố trí trên cùng một mặt phẳng.

- Tạo ra một vòng tròn tưởng tượng, chia vòng này thành k góc – mỗi góc sẽ chứa một nguyên tử, độ lớn mỗi góc tương ứng tỉ lệ với kích thước nguyên tử của nó. Vector $\vec{c}$ của tâm vòng tròn này được tính theo công thức:

$$\vec{c} = \frac{1}{|w|} \sum_{\vec{x} \in W} \vec{x},$$

trong đó W là tập tất cả các vector TFIDF của tập tài liệu, $\vec{x}$ là vector TFIDF của một tài liệu trong W.

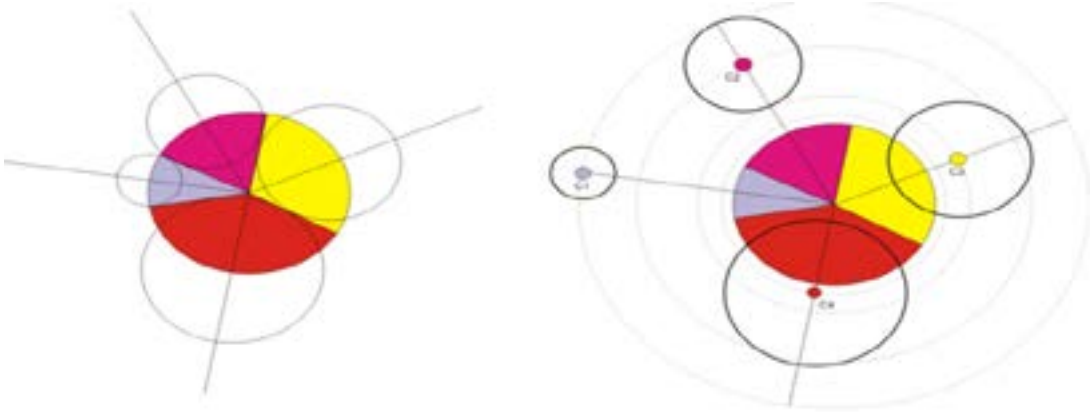
- Hạt nhân của mỗi nguyên tử (tâm $\vec{n}$ của mỗi cụm) sẽ nằm trên đường phân giác của góc; khoảng cách của hạt nhân so với tâm $\vec{c}$ của đường tròn được tính theo công thức:

$$Distance(\vec{n}, \vec{c}) = \frac{\vec{n} \cdot \vec{c}}{|\vec{n}||\vec{c}|}$$

Lưu ý, để đảm bảo ngữ nghĩa về "tính tương tự" nên sẽ không dùng công thức khoảng cách Euclid.

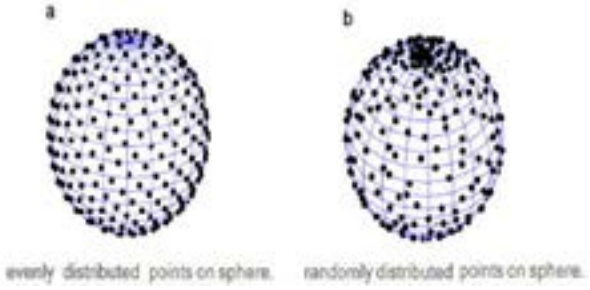
- Dựa trên vị trí của các hạt nhân vừa tính được, ta sẽ xác định được sự phân bố của các electron so với hạt nhân của nó dựa trên công thức tính mức năng lượng như đã đề cập phần trên.

Ví dụ minh họa Hình 6, chúng ta có bốn Cụm C1, C2, C3, C4 với số electron tương ứng n1 < n2 < n3 < n4. Trong hình bên trái, chúng ta có thể thấy rõ các vùng sẽ chứa các nguyên tử được tính toán dựa trên số lượng electron của nó. Trong Hình 6, chúng ta định vị trí các hạt nhân của các cụm C1, C2, C3, C4 sau khi tính được vị trí của hạt nhân nguyên tử so với tâm của hình tròn. Lưu ý, khi xảy ra sự chồng chéo, đan xen giữa các quỹ đạo các Cụm, một hệ số tỉ lệ cần được thêm vào để dịch chuyển các vị trí hạt nhân ra xa tâm vòng tròn nhằm tạo vùng không gian cách ly rộng hơn.

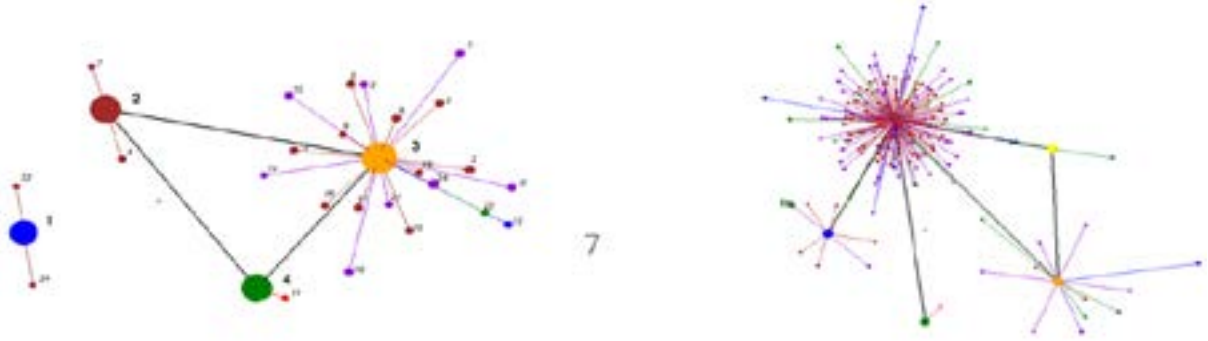


Hình 6. Tính vị trí cho bốn clusters dựa trên kích thước

Như đã trình bày, mỗi nguyên tử sẽ có tối đa năm mức năng lượng để phân bố các electron của chúng. Sau khi chuẩn hóa các giá trị khoảng cách từ một electron đến hạt nhân theo khoảng cách tối đa qui ước, chúng sẽ là một số thực nằm trong khoảng [0, max_distance]. Trong một mức năng lượng cụ thể ở trạng thái tĩnh, các electron sẽ được phân bố đều trên một quả cầu mà có tâm là vị trí của hạt nhân và bán kính chính là giá trị mức năng lượng của nó. Để giải quyết bài toán phân bố đều các điểm trên một quả cầu, chúng tôi tham khảo các giải pháp từ các diễn đàn thảo luận (Bulatov, 1996).



Hình 7. Phân bố đều các điểm trên quả cầu



Hình 9. 4 nguyên tử với 25 electrons (trái) và 5 nguyên tử với 200

Theo cách tiếp cập này, chúng ta sẽ có được một cái nhìn đầy đủ để phân tích từng Cụm và mối quan hệ giữa các cụm. Dựa trên không gian ba chiều, sự giới hạn về không gian cho việc trực quan hóa đã được xử lý một cách hiệu quả.

Truy vấn thông tin

Dựa vào tính chất thể hiện tài liệu bằng vector (cụ thể là TFIDF), ta có thể xem một truy vấn là một vector. Xét ví dụ với một truy vấn q = (visualization, cluster) từ tập vector tf-idf của ba tài liệu như sau:

Trước tiên, chúng ta sẽ biến đổi truy vấn này thành vector đơn vị:

q = (0, visualization, cluster)

$\vec{q}$ = (0,1,1) → vector đơn vị

$$\vec{q} = \left(0, \frac{1}{\sqrt{0^2+1^2+1^2}}, \frac{1}{\sqrt{0^2+1^2+1^2}}\right) = (0, 0.707, 0.707)$$

Tính điểm Score (q,d) của mỗi tài liệu d ứng với truy vấn q theo công thức độ tương tự cosine: (Manning, Christopher D.; Raghavan, Prabhakar; Schutze, Hinrich;, 2009)

$$\text{Score}(q,d) = \frac{\vec{v}(q).\vec{v}(d)}{|\vec{v}(q)|\,|\vec{v}(d)|}$$

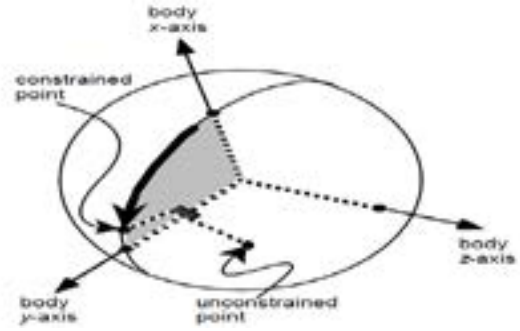
Theo Bảng 1, Doc 3 có Score cao nhất ứng với truy vấn q = (visualization, cluster). Điều này nói lên rằng Doc 3 có mối quan hệ gần gũi nhất với truy vấn này.

Bảng 1. Kết quả tính điểm cho truy vấn q

Term	Doc 1	Doc 2	Doc 3
Computer	0.996	0.993	0.847
Visualization	0.087	0.120	0.466
Cluster	0.017	0	0.254
Score(q,d)	**0.074**	**0.085**	**0.509**

Tương tác:

Chúng tôi hiện thực hệ thống này trong không gian ba chiều với đầy đủ tính năng của một hệ thống trực quan như là overview, zoom, filter, Detail-on-demand, relative and extract (Shneiderman, 2010). Chúng tôi chọn kỹ thuật Arcball (Shoemake, 1992) cho các thao tác trong một thế giới 3D một cách trực quan bởi vì nó không đòi hỏi các thiết bị đặc biệt hỗ trợ tương tác như kỹ thuật 3D ball và Tracer.



Hình 8. Kỹ thuật Arcball

4. Đánh giá Kết quả

Để đánh giá hệ thống, chúng ta xem xét ba tiêu chí:

(1) Tính trực quan: tiêu chí quan trọng đầu tiên là sự tổ chức, sắp xếp các đối tượng trên một màn hình máy tính mà nó phải thỏa mãn các yêu cầu về sự dễ hiểu, sự khả dụng và sự thẩm mỹ.

Dựa trên sự tổ chức các đối tượng theo mô hình nguyên tử, hệ thống dễ dàng thể hiện một cách hiệu quả cái nhìn tổng quan cũng như mối quan hệ giữa các đối tượng riêng rẽ trong tập dữ liệu. Kết hợp với khả năng tương tác tốt, hệ thống sẽ cho chúng ta hiểu được nhiều thông tin hơn. Cụ thể như chúng ta có thể so sánh mức độ tương đồng giữa các Cụm dữ liệu một cách gián tiếp dựa trên khoảng cách của các hạt nhân so với tâm chung. Khi muốn tìm hiểu mối quan hệ trực tiếp, hệ thống này sẽ hỗ trợ chúng ta so sánh ở chế độ lưới 3D-SOM.

(2) Thời gian thực thi: là một trong những yêu cầu thiết yếu của hệ thống. Chúng ta xem thời gian thực thi của các tác vụ sau đây:

Bảng 2. Thời gian thực thi cho 3 cụm với lần lượt 25x5052, 50x8865, 100x12348 (bài báo x số chiều của vector tfidf)

Tính toán thời gian thực thi từng tác vụ		25 x 5052	50 x 8865	100x12348
Xây dựng TFIDF	O(NxM)	00:00:17.66	00:00:53.93	00:02:36.86
K-means	O(IKNM)	00:00:00.23	00:00:00.91	00:00:03.24
SOM	O(I x N x M x) NumNeuron^3	00:00:24.15	00:00:41.16	00:02:01.55
Phân bố đều trên quả cầu	O(I x N^2)	00:00:00.03	00:00:00.10	00:00:00.44
Đơn vị: giờ:phút:giây.%giây I:số vòng lặp, K: số cụm, N: số lượng bài báo, M: số chiều Vector của bài báo				

Thời gian thực thi hệ thống bằng tổng thời gian các tác vụ. Nhìn chung, có thể thấy thời gian tính toán cho việc phân cụm trực quan 100 tài liệu chạy trên máy CPU Core i5, RAM 8 GB xấp xỉ 5 phút là chấp nhận được. Do không có số liệu đo của các hệ thống khác nên chưa thể đánh giá so sánh chúng.

(3) Tính đúng đắn của giải thuật: như đã đề cập, cách tiếp cận này sử dụng hai giải thuật quá phổ biến là SOM và K-Means vốn đã được chứng minh tính đúng đắn. Trong đó, chúng tôi trực quan hóa giải thuật SOM để tạo tiền đề xác định tham số K cho giải thuật K-means. Như vậy, hệ thống vẫn đảm bảo tính đúng đắn của việc phân cụm.

5. Kết luận

Hệ thống trực quan hóa thông tin 3D này được chúng tôi phát triển nhằm hỗ trợ việc phân cụm trực quan tập các bài báo khoa học nói riêng và các tài liệu, văn bản nói chung theo mô hình nguyên tử trong không gian 3 chiều, nó trợ giúp các nhà khai phá dữ liệu trong việc phân tích Cụm với một tập dữ liệu có chiều rất lớn. Các thí nghiệm đã cho thấy rằng phương pháp tiếp cận của chúng tôi có thể cải thiện hiệu quả của trực quan để phân tích cluster. Như chúng ta đã thấy, hệ thống này

có khả năng giúp người dùng dễ dàng hiểu được mối quan hệ giữa các bài báo khoa học trong một tập hàng ngàn bài báo. Công cụ này sẽ rất hữu ích trong việc hiển thị cụm và phơi bày những khoảng trống trong bộ dữ liệu (xu hướng tiết lộ thông tin từ tập dữ liệu). Kết quả là, với lợi thế của hệ thống này, những người khai phá dữ liệu có thể dễ dàng ước tính số lượng cụm cũng như có một hướng dẫn hiệu quả cho việc phân tích dữ liệu trong những bước tiếp theo với thông tin chính xác hơn.

Tài liệu tham khảo

Ankerst, Mihael, Grinstein, Georges, Keim, Daniel. 2002. *Visual Data Mining: Background, Techniques, and Drug Discovery Application.* Alberta, s.n.

Bulatov, V.. 1996. *The Mathematical Atlas: A Geteway to modern mathematics.* Xem 20.01.2013 <http://www.math.niu.edu/~rusin/known-math/96/repulsion>.

Cruz-Neira C, Sandin D., DeFanti T., Kenyon R., Hart J.. 1992. The CAVE. *Communications of the ACM,* 35(6), pp. 64-72.

Kohoren, T.. 1997. *Seft-Organizing Maps*. In: Second extended Edition ed. Berlin: Springer.

MacQueen, J. B.. 1967. *Some methods for classification and analysis of multivariate observations.* Berkeley, University of California Press, pp. 281-297.

Manning, Christopher D., Raghavan, Prabhakar, Schutze, Hinrich. 2009. "An introduction to Information Retrieval". *Cambridge University Press.*

Porter, M. F.. 1980. "Algorithm for suffix stripping". *Program,* pp. 130-137.

Shneiderman, P.. 2010. *Designing the user interaction interface: strategies for effective Human-Computer Interaction.* 5th ed. s.l.:Addison Wesley.

Shoemake. 1992. "Arcball: a user interface for specifying three-dimensional orientation using a mouse". *Proceedings of Graphics Interface'92,* pp. 151-156.

Ward, Matthew, Grinstein, Georges, Keim, Daniel. 2010. *Interaction Data Visualization: Foundation, Techniques, and Application.* s.l.:A K Peter, Ltd.

Ware, C.. 2004. *Information Visualization: Perception for design.* 2nd ed. s.l.:Morgan Kaufman.

Zhang, Ke-Bing, Orgun, Mehmet A, Zhang, Kang. 2006. HOV3: "An approach for Visual Cluster Analysis". *In Proceedings of The 2nd International Conference on Advanced Data Mining and Application,* Volume LNAI 4093, pp. 316-327.