ĐỘ ĐO GOOGLE TRONG TÍCH HỢP DỮ LIỆU

Vũ Ngọc Trình[1], Hà Quang Thụy[2], Trần Trọng Hiếu[2,3],

[1] Viện Dầu khí Việt Nam

[2] Trường Đại học Công nghệ, ĐHQG Hà Nội

[3] Trường Đại học Khoa học Tự nhiên, ĐHQG Hà Nội

trinhvn@vpi.pvn.vn, thuyhq@vnu.edu.vn, hieutt@vnu.edu.vn

TÓM TẮT *- Lý thuyết về độ đo đang là một trong những vấn đề được bàn đến nhiều trong các công trình nghiên cứu trong lĩnh vực khoa học máy tính bởi tính ứng dụng sâu rộng của nó từ thu hồi dữ liệu, khai phá dữ liệu đến tích hợp tri thức, nhận dạng và học máy. Việc tìm kiếm các độ đo phản ánh được sự khác biệt một cách tinh tế của các khái niệm, thuật ngữ và thực thể trong một ngữ cảnh nào đó là hết sức cần thiết và có tính ứng dụng thực tiễn cao. Trong bài báo này chúng tôi giới thiệu về một trong các độ đo như vậy, độ đo Google. Bài báo giới thiệu và bàn luận đầy đủ và chi tiết về cơ sở lý thuyết, các tính chất và một số ứng dụng của độ đo Google.*

Từ khóa *- Độ đo Google, tích hợp dữ liệu/tri thức.*

I. GIỚI THIỆU

Khi chữ viết được phát minh, con người có một công cụ tốt để mô tả các đối tượng bằng cách biểu diễn các đối tượng bằng một chuỗi các ký tự. Tuy nhiên do sự linh hoạt của ngôn ngữ nên cùng một đối tượng có thể được biểu diễn bằng nhiều chuỗi ký tự khác nhau và ngược lại một chuỗi ký tự cũng có thể biểu diễn nhiều đối đượng khác nhau. Do đó việc xem xét các đối tượng từ các chuỗi ký tự cần được xem xét trong ngữ cảnh cụ thể. Một bài toán (ngược) được đặt ra là nếu chúng ta có hai chuỗi ký tự, tìm ngữ cảnh mà hai chuỗi ký tự này biểu diễn các đối tượng có quan hệ gần nhau nhất. Bài toán này có nhiều ứng dụng trong xử lý ngôn ngữ tự nhiên, phân cụm dữ liệu, học máy,… Trong bài này chúng ta sẽ xem xét một trong các cách để trả lời cho câu hỏi này.

Hằng ngày có hơn một tỷ lượt người dùng Internet với hàng tỷ comment trên các mạng xã hội, tweeter và hàng triệu các tài liệu được xuất bản trên đó. Internet trở thành một kho dữ liệu khổng lồ về các đối tượng ở tất cả các ngôn ngữ và trong vô vàn các ngữ cảnh khác nhau. Với quy mô cực lớn của Internet, con người không thể tìm kiếm các đối tượng một cách thủ công thông qua các đường link. Thay vào đó chúng ta sử dụng các máy tìm kiếm để hỗ trợ cho việc này, chúng ta chỉ cần gửi các truy vấn và máy tìm kiếm sẽ trả lại các kết quả tìm được. Một trong các máy tìm kiếm mạnh mẽ nhất trên Internet hiện nay là Google. Google hỗ trợ cho cả người dùng qua giao diện người dùng và các nhà phát triển thông qua giao diện API. Các thông tin kết quả mà Google trả về khá chi tiết và đủ cơ sở để chúng ta có thể xây dựng một độ đo như sẽ trình bày trong các mục phía sau.

Trong các công trình nghiên cứu trước đây, một trong các hướng nghiên cứu chính để so sánh các các từ hay cụm từ là sử dụng tần suất xuất hiện của chúng trong các văn bản để xây dựng các độ đo sự tương đồng [6], [7], [8]. Một tiếp cận khác là sử dụng các độ đo giữa các đối tượng được biểu diễn bởi các chuỗi ký tự như [1], [3], [4], [2], [9]. Tuy nhiên các tiếp cận này đều có một điểm yếu chung là chúng phân tích các đối tượng một cách độc lập với các đặc điểm của chúng, tức là chúng phân tích đồng thời tất cả các đặc điểm của đối tượng và xác định sự tương đồng giữa các cặp đối tượng thông qua sự giống nhau nhiều nhất trong số các cặp đặc điểm mà hai đối tượng cùng chia sẻ. Với cách tiếp cận này, các đối tượng được so sánh trực tiếp với nhau và do đó chỉ phù hợp để đạt được các tri thức về chính các đối tượng đó mà không quan tâm đến thông tin chung về sự tương đồng này. Trong bài báo này chúng tôi giới thiệu một cách tiếp cận mới nhằm lấy được các thông tin ý nghĩa hơn về sự tương đồng giữa các đối tượng thông qua ngữ nghĩa Google. Cụ thể là chúng tôi sử dụng các tên của đối tượng và thông qua máy tìm kiếm Google để thu được thông tin về đối tượng từ vô số các người dùng web trong các ngữ cảnh khác nhau, qua đó thống kê tần suất xuất hiện của các tên đối tượng khi chúng xuất hiện riêng rẽ và xuất hiện cùng nhau để có thể xác định một cách định lượng sự tương đồng giữa các đối tượng này.

Trong bài báo cơ sở lý thuyết được trình bày trong Mục II, trong đó các khái niệm chính được đề cập gồm có: Độ phức tạp Kolmogorov, khoảng cách thông tin, độ đo sự tương đồng dựa trên hàm nén. Tiếp theo một mô tả ngắn gọn về phân phối Google, khoảng cách Google và bàn luận về các tính chất của khoảng cách này được trình bày trong Mục III. Mục IV trình bày về một số ứng dụng tiêu biểu của độ đo Google bao gồm xây dựng các cây phân lớp và canh các mục của các ontology. Kết luận và các công việc tương lai được trình bày trong Mục V.

II. CƠ SỞ LÝ THUYẾT

Cơ sở lý thuyết của bài báo này xuất phát từ độ phức tạp Kolmogorov [5]. Dựa trên độ phức tạp này chúng ta sẽ lần lượt xây dựng các khoảng cách thông tin được chuẩn hóa, khoảng cách nén được chuẩn hóa và đi đến xây dựng khoảng cách Google. Nội dung chi tiết của phần này như sau.

A. Độ phức tạp Kolmogorov

Để xem xét về độ phức tạp Kolmogorov, chúng ta trước tiên cần xem xét về khái niệm hệ thống lập trình. Một cách không hình thức, một hệ thống lập trình được hiểu là hệ thống mà qua đó chúng ta có thể xây dựng các chương

trình nhằm thực hiện các yêu cầu công việc khác nhau. Mỗi hệ thống lập trình thường sử dụng một ngôn ngữ lập trình nào đó, chẳng hạn shell, C/C++, LIPS,… Ngược lại, khi nói về các chương trình, chúng ta cần tham chiếu đến hệ thống lập trình để sinh ra chúng.

Định nghĩa 1. Độ phức tạp Kolmogorov của một xâu *x* là độ dài tính theo bít của chương trình ngắn nhất để sinh ra *x* trên một hệ thống lập trình được tham chiếu.

Gọi độ phức tạp Kolmogorov của một xâu *x* là *K(x)*. Từ định nghĩa trên chúng ta có nhận xét rằng việc lựa chọn các hệ thống lập trình khác nhau sẽ làm thay đổi giá trị của *K(x)* bằng cách cộng thêm một hằng số cố định. Một cách hiểu đơn giản của độ phức tạp Kolmogorov của xâu *x* là các độ dài nhỏ nhất của các xâu nén mà qua đó chúng ta có thể sinh ra *x* thông qua các chương trình giải nén khác nhau. Một ví dụ là khi nén cùng xâu *x* bằng thuật toán *gzip* chúng ta thu được xâu x_g và bằng thuật toán tốt hơn là *rar* chúng ta thu được x_r. Dùng ký hiệu |*s*| để biểu diễn độ dài theo bít của xâu *s*, ta có $K(x) \leq |x_r| \leq |x_g| \leq |x|$.

Độ phức tạp Kolmogorov cung cấp giá trị giới hạn dưới của các chương trình sinh ra *x*. Đó là độ dài của chương trình "lý tưởng" sinh ra xâu *x* trong một hệ thống lập trình cụ thể. Trở lại ví dụ trên, *K(x)* là giá trị độ dài nhỏ nhất của xâu kết quả khi nén *x* bằng mọi thuật toán nén có thể.

B. Khoảng cách thông tin được chuẩn hóa

Định nghĩa 2. Cho hai xâu *x* và *y*, *δ* là chương trình ngắn nhất sao cho *δ(x) = y* và *δ(y) = x*, độ dài của *δ* được gọi là *khoảng cách thông tin* giữa *x* và *y*.

Khoảng cách thông tin giữa *x* và *y* và được ký hiệu là *E(x, y)* và được tính theo công thức là:

$E(x, y) = K(x, y) + min\{K(x), K(y)\}$

Trong đó *K(x, y)* là độ dài của chương trình nhỏ nhất sinh ra cặp *x, y* và cách để phân tác chúng. Rõ ràng khoảng cách *E(x, y)* là một metric, tức là nó có các tính chất sau:

Với mọi xâu *x, y* và *z* ta có:

1. **E(x, y) > 0 với x ≠ y;**
2. **E(x, x) = 0;**
3. **E(x, y) = E(y, x);**
4. **E(x, y) + E(y, z) ≥ E(x, z).**

Vì *E* là một metric, hiển nhiên nó là một độ đo tốt. Tuy nhiên, chúng ta có nhận xét như sau: Do *E* không quan tâm đến độ dài của các xâu đầu vào nên nếu có cùng một khoảng cách thông tin, hai xâu nhỏ sẽ rất khác nhau trong khi hai xâu lớn lại có thể rất giống nhau. Do đó, khoảng cách thông tin không phản ánh đầy đủ được về sự tương đồng giữa các xâu. Do vậy việc chuẩn hóa khoảng cách thông tin là cần thiết.

Khoảng cách thông tin được chuẩn hóa có giá trị nằm trong khoảng 0 và 1 là hàm khoảng cách thông tin có xét đến độ dài của các xâu đầu vào. Công thức để tính khoảng cách này như sau:

$$NID(x,y) = \frac{K(x,y) - \min(K(x), K(y))}{max(K(x), K(y))}$$

Khoảng cách thông tin được chuẩn hóa có một số tính chất thú vị và nó cũng được chứng minh là một metric (chi tiết xem tại [4]).

C. Khoảng cách nén được chuẩn hóa

Mặc dù *NID* là một là một độ đo tốt nhưng nó được dựng dựa trên độ phức tạp Kolmogorov. Điều này dẫn tới *NID* không thể tính được trong thực tế vì độ phức tạp Kolmogorov là không thể tính được. Để khắc phục chúng ta cần xấp xỉ các độ phức tạp Kolmogorov trong công thức nói trên bằng cách sử dụng hàm nén. Mỗi hàm nén nhận vào một xâu ký tự và trả lại một xâu kết quả nén. Xâu kết quả này có độ dài (theo bít) nhỏ hơn xâu đầu vào và là cận trên của các của độ phức tạp Kolmogorov đối với xâu đầu vào. Nói cách khác, độ phức tạp Kolmogorov của xâu đầu vào sẽ nhỏ hơn hay bằng độ dài của xâu kết quả nén mà chúng ta đã chỉ ra được. Gọi *C* là một hàm nén và *C(x)* trả kết quả là xâu được nén của *x*, khi đó khoảng cách nén được chuẩn hóa được định nghĩa như sau:

$$\mathrm{NCD}_C(x,y) = \frac{C(xy) - \min(C(x), C(y))}{max(C(x), C(y))}$$

trong đó để thuận tiện chúng ta thay *C(x, y)* bằng *C(xy)* với *xy* có được bằng cách nối xâu *x* với xâu *y*. Rõ ràng NCD_C xấp xỉ *NID* khi *C* xấp xỉ *K*. Bây giờ *NCD* là một lớp các hàm khoảng cách nén được tham số hóa bởi hàm nén *C*. Nếu có *C* và *C'* là hai hàm nén và *C* là "tốt hơn" *C'*, tức là $C(x) \leq C'(x)$ thì chúng ta cũng sẽ có $NCD_C(x, y) \leq NCD_{C'}(x, y)$.

III. ĐỘ ĐO GOOGLE

Trong mỗi văn bản, tần suất xuất hiện của các từ hay cụm từ phản ánh mối quan hệ (về tần suất) giữa từ hay cụm từ này. Trên môi trường Internet hiện nay có hàng tỉ trang web (web page) chứa nội dung do hàng triệu người

dùng tạo ra và đã được Google lập chỉ mục tìm kiếm[1]. Mỗi trang web được lập chỉ mục có một phân bố xác suất riêng. Với số lượng vô cùng lớn các trang web như vậy, chúng ta có thể coi như là tập vũ trụ và nó đã bao quát (gần như) toàn bộ các ngữ cảnh có thể. Trong phần này chúng ta xem xét các nội dung chính gồm có: phân bố xác suất Google, ngữ nghĩa Google, mã Google và độ đo Google.

A. Phân bố xác suất Google

Gọi S là tập các từ khóa tìm kiếm đơn, tập các cặp khóa tìm kiếm là $\{(x,y): x,y \in S\}$. Tập các trang web được lập chỉ mục bởi Google là Ω. Số lượng các trang web được Google lập chỉ mục là $M = |\Omega|$. Một cách lý tưởng chúng ta giả sử rằng các trang web này có xác suất được tìm thấy là như nhau[2] và bằng $1/M$. Mỗi tập con của tập Ω được gọi là một *sự kiện*. Với mỗi từ khóa tìm kiếm đơn x, Google sẽ trả lại tập các trang chứa x trong một sự kiện đơn e_x. Xác suất của sự kiện e_x là: $L(e_x) = |e_x|/M$. Với hai từ khóa tìm kiếm x và y, xác suất của cặp sự kiện e_x và e_y được tính là $L(e_x \cap e_y)=|e_x \cap e_y|/M$, là xác suất tìm trang web mà có cả x và y xuất hiện.

B. Ngữ nghĩa Google

Ngữ nghĩa của Google được phát biểu ngắn gọn như sau: Sự kiện e_x chứa tập các trang web mà từ khóa tìm kiếm x có xuất hiện một hay nhiều lần, nó thể hiện tất cả các ngữ cảnh có liên quan trực tiếp đến x.

Lưu ý rằng trong một số trường hợp, ngữ cảnh của trang web chứa từ khóa tìm kiếm x liên quan trực tiếp đến các trang web khác mà x không xuất hiện. Các trang web như vậy gọi là có ngữ cảnh gián tiếp liên quan đến x. Ngữ cảnh gián tiếp cũng quan trọng trong tìm kiếm liên quan ngữ nghĩa. Tuy nhiên trong bài báo này chúng ta tạm thời chưa xét đến.

C. Mã Google

Trong bài báo này chúng ta sử dụng xác suất của các sự kiện để định nghĩa hàm khối xác suất trên tập $\{(x,y): x,y \in S\}$ của các từ khóa tìm kiếm đơn cũng như các cặp từ khóa tìm kiếm. Chúng ta có $|S|$ các từ khóa tìm kiếm đơn và $C^2_{|S|}$ các cặp từ khóa tìm kiếm mà các từ khóa trong mỗi cặp là khác nhau. Chúng ta định nghĩa

$$\mathrm{N} = \sum_{\{x,y\}\subseteq S} |e_x \cap e_y|$$

là tổng số các trang web tìm được từ hai từ khóa x và y. Với $\{x,y\} \subseteq S$ và $x \neq y$, mỗi trang web $z \in e_x \cap e_y$ sẽ được đếm đến ba lần trong các tập e_x, $e_x \cap e_y$ và e_y. Mỗi trang web được Google đánh chỉ mục phải chứa ít nhất một từ khóa tìm kiếm, do đó $N \geq M$. Ngược lại mỗi trang web chứa trung bình α từ khóa nên chúng ta cũng có $N \leq \alpha.M$.

Tiếp theo chúng ta định nghĩa hàm phân phối Google g như sau:

- $g(x) = g(x,x)$, (1)
- $g(x,y) = L(e_x \cap e_y)M/N = |e_x \cap e_y|/N$. (2)

Ta có: $\sum_{\{x,y\}\subseteq S} g(x,y) = 1$. Từ hàm phân phối g này chúng ta định nghĩa mã Google G như sau:

- $G(x) = G(x,x)$, (3)
- $G(x,y) = \log 1/g(x,y)$. (4)

D. Độ đo Google

Như đã trình bày ở các phần trên, với một xâu x, độ phức tạp $C(x)$ sẽ trả lại độ dài của kết quả nén xâu x bởi hàm nén C. Trong khi đó mã Google của độ dài $G(x)$ biểu diễn độ dài từ tiền mã ngắn nhất được mong đợi của sự kiện e_x. Giá trị kỳ vọng này có được từ phân phối Google g. Do vậy ta có thể dùng phân phối Google như bộ nén cho ngữ nghĩa Google. Kết hợp với họ các hàm khoảng cách nén được chuẩn hóa ở trên ta được khoảng các Google được chuẩn hóa như sau:

$$NCD_G(x,y) = \frac{G(x,y) - \min(G(x),G(y))}{max(G(x),G(y))} \quad (5)$$

Kết hợp công thức (5) với các công thức (1), (2), (3) và (4) ở trên và thực hiện một số biến đổi đơn giản, chúng ta có:

$$NCD_G(x,y) = \frac{max(\log|e_x|,\ \log|e_y|) - \log|e_x \cap e_y|}{\log N - min(\log|e_x|,\log|e_y|)} \quad (6)$$

E. Các tính chất của độ đo Google

Mệnh đề 3. Khoảng giá trị của NCD_G từ 0 đến +∞.

- Nếu $x = y$ hoặc $x \neq y$ nhưng $|e_x| = |e_x \cap e_y| = |e_y| > 0$ thì $NCD_G(x,y)=0$, tức là x và y có cùng ngữ nghĩa Google.
- Nếu $|e_x| = 0$ thì với mọi từ khóa tìm kiếm y ta luôn có $|e_x \cap e_y| = 0$, do đó $NCD_G(x,y)=\infty/\infty$. Trong trường hợp này ta gán cho nó giá trị là 1.

[1] Tính đến tháng 6/2015 Google đã lập chỉ mục được $2{,}5.10^{10}$ trang web.

[2] Thực tế thì có một số trang có xác suất được tìm thấy cao hơn do chính sách của Google (quảng cáo, ưu tiên, …).

Mệnh đề 4. NCD_G là một khoảng cách nhưng không là metric.

Thật vậy:

- NCD_G luôn không âm và $NCD_G(x,x)$=0 với mọi khóa tìm kiếm x;
- NCD_G có tính chất đối xứng. Điều này là hiển nhiên vì theo công thức (6) vai trò của x và y là như nhau.
- NCD_G không thỏa mãn tính chất tách biệt, tức là $NCD_G(x,y) > 0$ với mọi cặp $x \neq y$.
- NCD_G cũng không thỏa mãn bất phương trình tam giác, tức là $NCD_G(x,z) \leq NCD_G(x,y) + NCD_G(y,z)$.

IV. CÁC ỨNG DỤNG

Trong mục này chúng ta xem xét một số ứng dụng của độ đo Google bao gồm xây dựng các cây phân lớp và canh các mục của các ontology.

A. Cây phân lớp

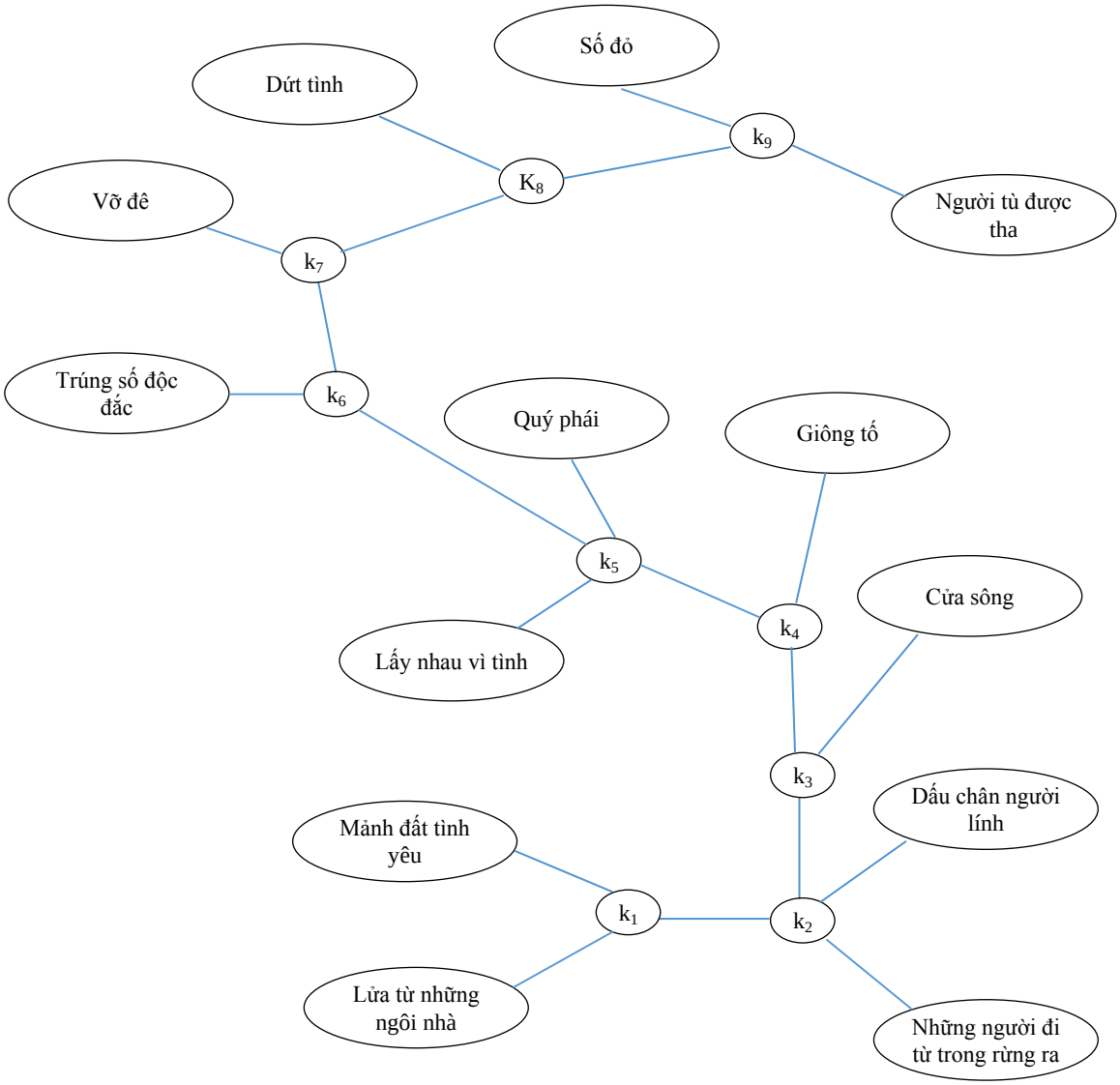
Trong việc phân lớp các đối tượng, một độ đo được sử dụng để xác định khoảng cách giữa các đối tượng. Độ đo này sẽ xác định ma trận khoảng cách giữa các đối tượng trong tập các đối tượng cần phân lớp. Sau đó một thuật toán phân lớp được áp dụng để phân lớp các đối tượng và xây dựng lên cây phân lớp. Ở đây chúng ta xét một tập các tiểu thuyết của hai tác giả Vũ Trọng Phụng và Nguyễn Minh Châu. Tập các tiểu thuyết này gồm có:

- Nguyễn Minh Châu: *Cửa sông, Dấu chân người lính, Mảnh đất tình yêu, Lửa từ những ngôi nhà, Những người đi từ trong rừng ra*.
- Vũ Trọng Phụng: *Dứt tình, Giông tố, Lấy nhau vì tình, Người tù được tha, Quý phái, Số đỏ, Trúng số độc đắc, Vỡ đê*.

Bảng 1. Ma trận khoảng cách giữa các đối tượng.

Cửa sông	0	0.3505	0.3748	0.3943	0.7921	0.2716	0.3250	0.9733	0.2484	0.3788	0.4919	0.3695	0.3087
Dấu chân người lính	0.3505	0	0.2871	0.0936	0.6324	0.1734	0.1687	0.7740	0.0767	0.2467	0.2673	0.3822	0.2911
Dứt tình	0.3748	0.2871	0	0.3270	0.1979	1.1255	0.2417	0.4252	0.7434	0.2626	0.3962	0.4840	0.4141
Giông tố	0.3943	0.0936	0.3270	0	0.1841	0.9502	0.2632	0.3669	0.3769	0.2938	0.2826	0.1859	0.2730
Lấy nhau vì tình	0.7921	0.6324	0.1979	0.1841	0	1.0000	0.7905	0.2719	1.0000	0.1856	0.4477	0.2590	0.1966
Lửa từ những ngôi nhà	0.2716	0.1734	1.1255	0.9502	1.0000	0	0.3325	1.0000	0.2302	1.1493	0.5788	0.9818	0.8816
Mảnh đất tình yêu	0.3250	0.1687	0.2417	0.2632	0.7905	0.3325	0	0.8774	0.2172	0.3016	0.5057	0.5778	0.5009
Người tù được tha	0.9733	0.7740	0.4252	0.3669	0.2719	1.0000	0.8774	0	1.0000	0.4447	0.3785	0.4214	0.3412
Những người đi từ trong rừng ra	0.2484	0.0767	0.7434	0.3769	1.0000	0.2302	0.2172	1.0000	0	0.7605	0.4257	0.6896	0.6126
Quý phái	0.3788	0.2467	0.2626	0.2938	0.1856	1.1493	0.3016	0.4447	0.7605	0	0.4073	0.2157	0.4082
Số đỏ	0.4919	0.2673	0.3962	0.2826	0.4477	0.5788	0.5057	0.3785	0.4257	0.4073	0	0.4053	0.3136
Trúng số độc đắc	0.3695	0.3822	0.4840	0.1859	0.2590	0.9818	0.5778	0.4214	0.6896	0.2157	0.4053	0	0.2261
Vỡ đê	0.3087	0.2911	0.4141	0.2730	0.1966	0.8816	0.5009	0.3412	0.6126	0.4082	0.3136	0.2261	0

Từ ma trận khoảng cách giữa các đối tượng thu được bằng độ đo Google (Bảng 1). Sử dụng phần mềm vẽ cây phân lớp tại địa chỉ: http://www.complearn.org, chúng ta thu được cây phân lớp của các tiểu thuyết như Hình 1.



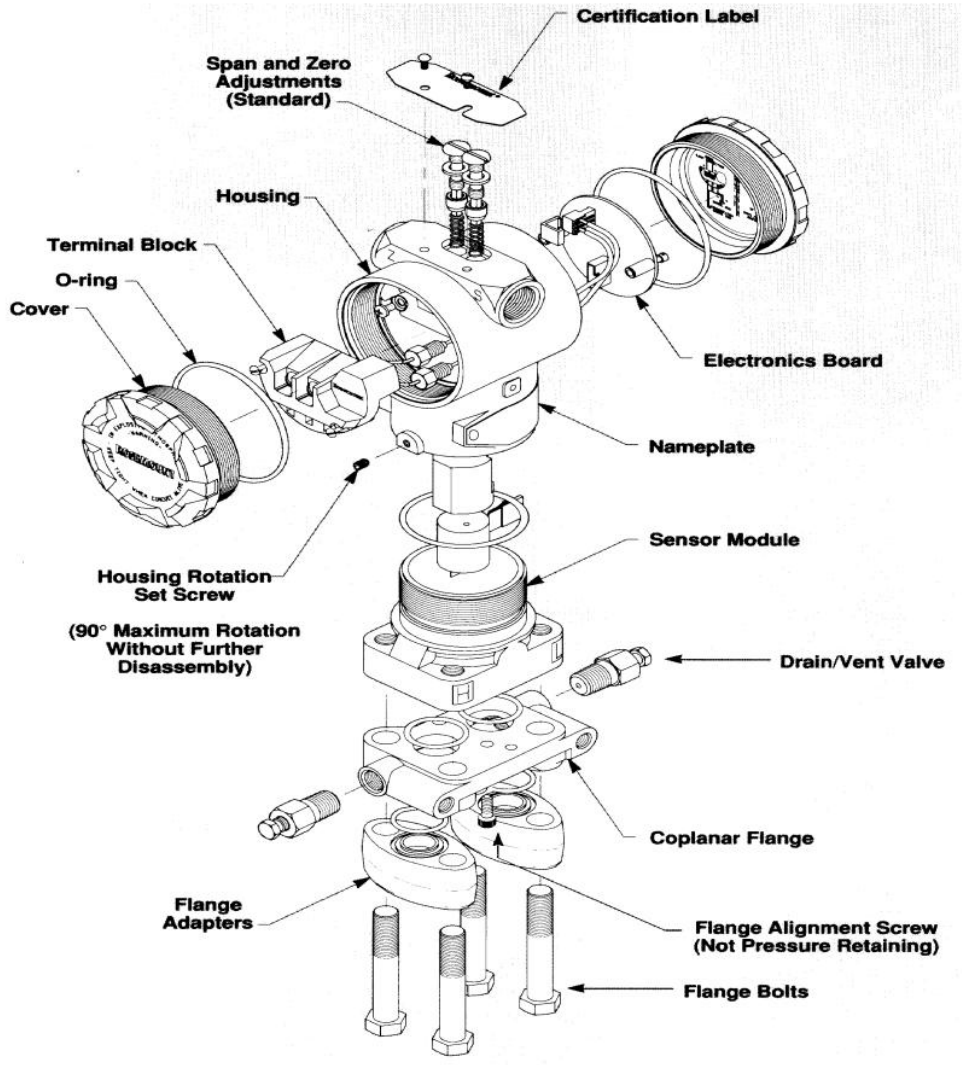
Hình 1. Cây phân lớp của các tiểu thuyết

B. Canh các mục của ontology

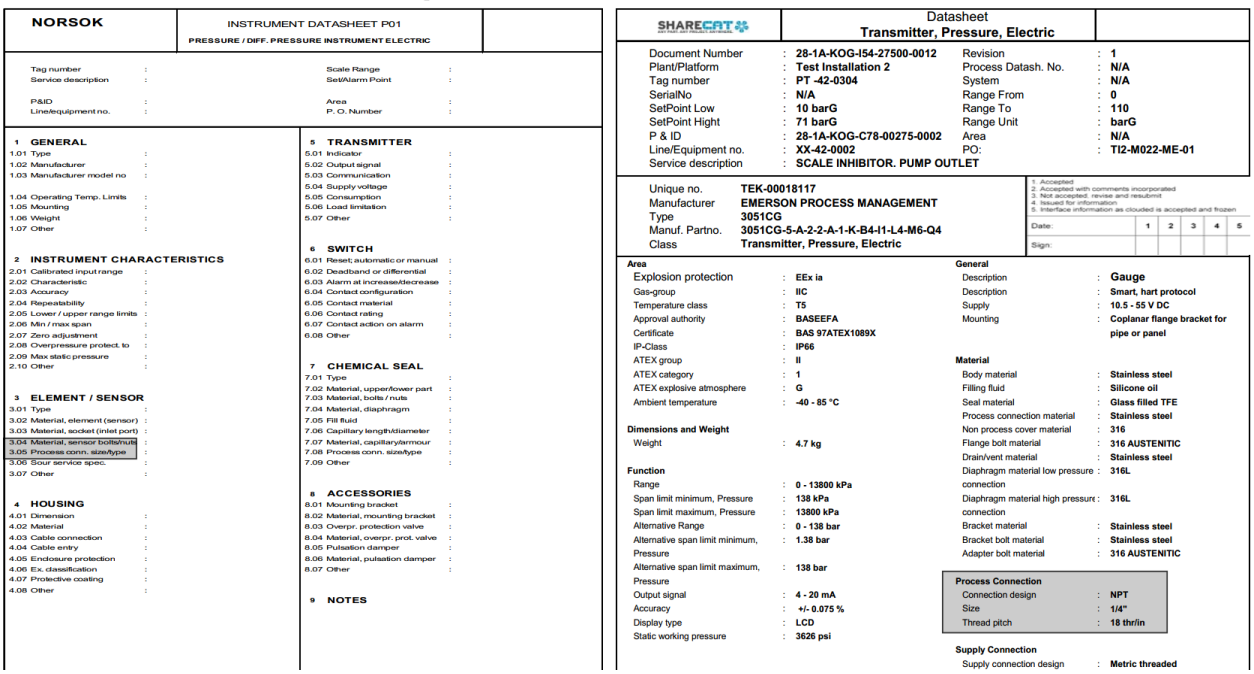
Trong phần này chúng ta xem xét một ứng dụng khác của độ đo Google đó là dùng để canh các mục trong các ontology. Trong thực tế cùng một đối tượng trong thế giới thực có thể được mô hình hóa và biểu diễn bằng các ontology khác nhau trong các hệ thống khác nhau. Để các hệ thống này có thể trao đổi dữ liệu và làm việc được với nhau thì các ontology này cần phải được thống nhất lại về nội dung của các mục mà chúng biểu diễn. Xét một ví dụ, Hình 2 biểu diễn một bộ truyền áp suất được sử dụng trong khai thác dầu khí với các bộ phận chính của thiết bị được mô tả. Hình 3 là bảng các thông số kỹ thuật của cùng thiết bị này được cung cấp bởi hai nhà cung cấp là Norsock và ShareCat. Chúng ta dễ dàng tìm thấy những điểm khác nhau về thông tin của thiết bị trên hai bảng này. Ví dụ, thông tin về Trọng lượng (*Weight*) của thiết bị nằm trong mục Thông tin chung (*General*) trong bảng thứ nhất nhưng lại nằm trong mục Kích cỡ và Trọng lượng (*Dimenssions and Weight*) ở bảng thứ hai, hay thông tin về Kết nối (*Process Connection*) trong bảng thứ hai (phần bôi đậm) lại là các mục con của mục Thành phần/Cảm biến (Element/Sensor) trong bảng thứ nhất.

Quá trình canh các mục của các ontology đòi hỏi nhiều thời gian và công sức của các chuyên gia, nhất là khi số mục của các ontology lên đến con số hàng nghìn, thậm chí hàng vạn. Một chương trình hỗ trợ trong việc canh mục các ontology bằng cách đưa ra các gợi ý cho các chuyên gia là rất cần thiết. Ở đây chúng ta xét một chương trình như vậy sử dụng độ đo Google.

Do hạn chế về số trang của bài báo và mang tính minh họa, chúng ta chỉ xem xét sự canh mục các ontology với thông tin phần tiêu đề của hai bảng thông số kỹ thuật này. Các thông tin của Norsock gồm có: *Tag number, Scale Range, Service description, Set/Alarm Point, P&ID, Area, Line / equipment no., P. O. Number*; của ShareCat gồm có: *Document Number, Revision, Plant/Platform, Process Datash. No. , Tag number, SerialNo, Range From, SetPoint Low, Range To, SetPoint Height, Range Unit, P & ID, Area, Line/Equipment no. , Service description.* Ma trận khoảng cách Google giữa các mục này được tính như trong Bảng 2. Qua đó một gợi ý về canh các mục được trình bày như Hình 4.



Hình 2. Một bộ truyền áp suất trong khai thác dầu khí



NORSOK — INSTRUMENT DATASHEET P01
PRESSURE / DIFF. PRESSURE INSTRUMENT ELECTRIC

Tag number : Scale Range :
Service description : Set/Alarm Point :
P&ID : Area :
Line/equipment no. : P. O. Number :

1 GENERAL
1.01 Type :
1.02 Manufacturer :
1.03 Manufacturer model no :
1.04 Operating Temp. Limits :
1.05 Mounting :
1.06 Weight :
1.07 Other :

2 INSTRUMENT CHARACTERISTICS
2.01 Calibrated input range :
2.02 Characteristic :
2.03 Accuracy :
2.04 Repeatability :
2.05 Lower / upper range limits :
2.06 Min / max span :
2.07 Zero adjustment :
2.08 Overpressure protect. to :
2.09 Max static pressure :
2.10 Other :

3 ELEMENT / SENSOR
3.01 Type :
3.02 Material, element (sensor) :
3.03 Material, socket (inlet port) :
3.04 Material, sensor bolts/nuts :
3.05 Process conn. size/type :
3.06 Sour service spec. :
3.07 Other :

4 HOUSING
4.01 Dimension :
4.02 Material :
4.03 Cable connection :
4.04 Cable entry :
4.05 Enclosure protection :
4.06 Ex. classification :
4.07 Protective coating :
4.08 Other :

5 TRANSMITTER
5.01 Indicator :
5.02 Output signal :
5.03 Communication :
5.04 Supply voltage :
5.05 Consumption :
5.06 Load limitation :
5.07 Other :

6 SWITCH
6.01 Reset; automatic or manual :
6.02 Deadband or differential :
6.03 Alarm at increase/decrease :
6.04 Contact configuration :
6.05 Contact material :
6.06 Contact rating :
6.07 Contact action on alarm :
6.08 Other :

7 CHEMICAL SEAL
7.01 Type :
7.02 Material, upper/lower part :
7.03 Material, bolts / nuts :
7.04 Material, diaphragm :
7.05 Fill fluid :
7.06 Capillary length/diameter :
7.07 Material, capillary/armour :
7.08 Process conn. size/type :
7.09 Other :

8 ACCESSORIES
8.01 Mounting bracket :
8.02 Material, mounting bracket :
8.03 Overpr. protection valve :
8.04 Material, overpr. prot. valve :
8.05 Pulsation damper :
8.06 Material, pulsation damper :
8.07 Other :

9 NOTES

SHARECAT — Datasheet
Transmitter, Pressure, Electric

Document Number	28-1A-KOG-I54-27500-0012	Revision	1
Plant/Platform	Test Installation 2	Process Datash. No.	N/A
Tag number	PT -42-0304	System	N/A
SerialNo	N/A	Range From	0
SetPoint Low	10 barG	Range To	110
SetPoint Hight	71 barG	Range Unit	barG
P & ID	28-1A-KOG-C78-00275-0002	Area	N/A
Line/Equipment no.	XX-42-0002	PO:	TI2-M022-ME-01
Service description	SCALE INHIBITOR. PUMP OUTLET		

Unique no. TEK-00018117
Manufacturer EMERSON PROCESS MANAGEMENT
Type 3051CG
Manuf. Partno. 3051CG-5-A-2-2-A-1-K-B4-I1-L4-M6-Q4
Class Transmitter, Pressure, Electric

1. Accepted
2. Accepted with comments incorporated
3. Not accepted. revise and resubmit
4. Issued for information
5. Interface information as clouded is accepted and frozen

Date: 1 2 3 4 5
Sign:

Area

Explosion protection	EEx ia
Gas-group	IIC
Temperature class	T5
Approval authority	BASEEFA
Certificate	BAS 97ATEX1089X
IP-Class	IP66
ATEX group	II
ATEX category	1
ATEX explosive atmosphere	G
Ambient temperature	-40 - 85 °C

Dimensions and Weight

Weight	4.7 kg

Function

Range	0 - 13800 kPa
Span limit minimum, Pressure	138 kPa
Span limit maximum, Pressure	13800 kPa
Alternative Range	0 - 138 bar
Alternative span limit minimum, Pressure	1.38 bar
Alternative span limit maximum, Pressure	138 bar
Output signal	4 - 20 mA
Accuracy	+/- 0.075 %
Display type	LCD
Static working pressure	3626 psi

General

Description	Gauge
Description	Smart, hart protocol
Supply	10.5 - 55 V DC
Mounting	Coplanar flange bracket for pipe or panel

Material

Body material	Stainless steel
Filling fluid	Silicone oil
Seal material	Glass filled TFE
Process connection material	Stainless steel
Non process cover material	316
Flange bolt material	316 AUSTENITIC
Drain/vent material	Stainless steel
Diaphragm material low pressure connection	316L
Diaphragm material high pressure connection	316L
Bracket material	Stainless steel
Bracket bolt material	Stainless steel
Adapter bolt material	316 AUSTENITIC

Process Connection

Connection design	NPT
Size	1/4"
Thread pitch	18 thr/in

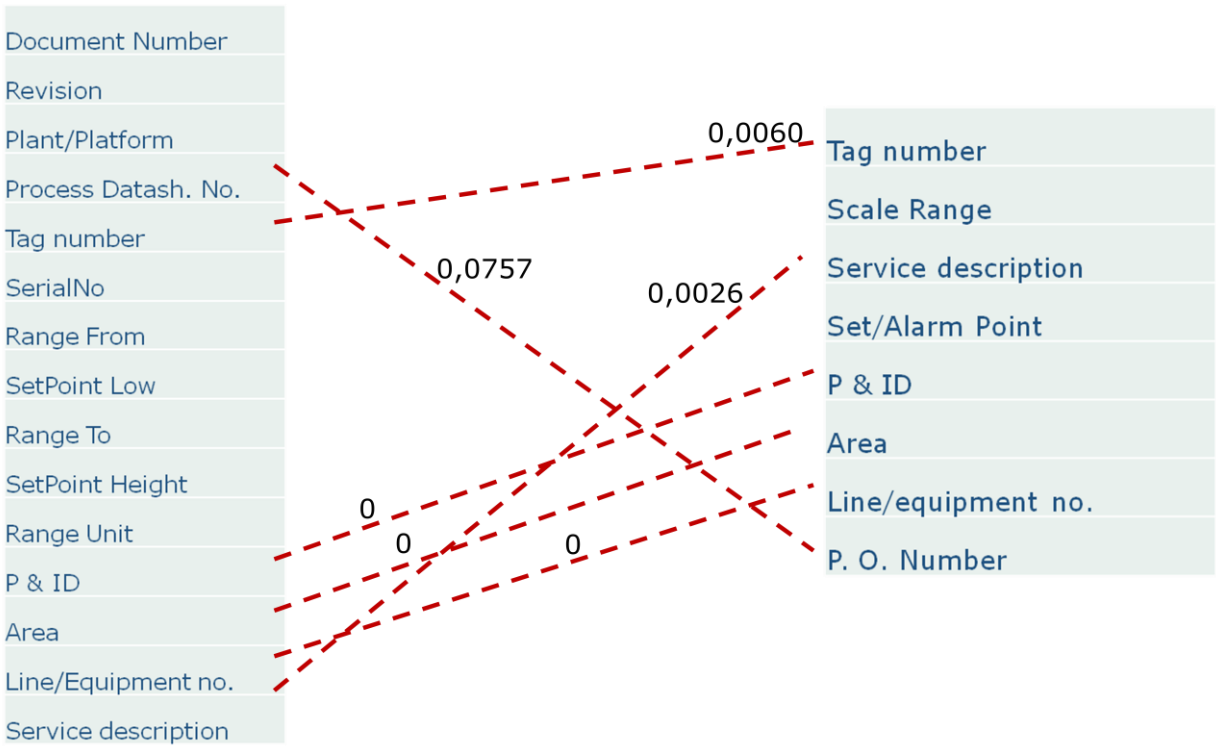
Supply Connection

Supply connection design	Metric threaded

Hình 3. Bảng thông số kỹ thuật của Norsock và ShareCat

Bảng 2. Ma trận khoảng cách giữa các tên mục của các ontology

	Tag number	Scale Range	Service description	Set/Alarm Point	P&ID	Area	Line / equipment no.	P. O. Number
Document Number	0.6630	0.6822	0.6929	0.6998	0.8105	0.9022	0.6877	0.2390
Revision	0.7950	0.7572	0.8154	0.8403	0.8419	0.7957	0.8728	0.4187
Plant/Platform	0.7220	0.7391	0.8054	0.3959	0.4981	0.9233	0.3890	0.3564
Process Datash. No.	0.5032	0.4956	0.5400	0.1678	0.5579	0.8952	0.1532	0.0757
Tag number	0.0060	0.5839	0.6484	0.6011	0.5602	0.8976	0.5682	0.2776
SerialNo	0.7927	0.7961	0.8897	0.5603	1.0096	0.9939	0.5506	0.4692
Range From	0.8289	0.6055	0.7786	0.7736	0.9214	0.7615	0.8323	0.4852
SetPoint Low	0.6397	0.5051	0.7341	0.3176	0.7140	0.9396	0.3121	0.2859
Range To	0.7861	0.5679	0.8279	0.7494	0.9377	0.7194	0.7589	0.4312
SetPoint Height	1.0000	1.0000	1.0000	1.0000	1.0000	0.8975	1.0000	1.0000
Range Unit	0.7310	0.6545	0.8789	0.5524	0.9370	0.9539	0.5428	0.4973
P & ID	0.3860	0.6272	0.6305	0.5017	0.0000	0.6752	0.4706	0.3606
Area	0.9030	0.8708	0.8692	0.9037	0.8895	0.0000	0.9106	0.5048
Line/Equipment no.	0.5717	0.5853	0.6152	0.1982	0.6363	0.9106	0.0000	0.1657
Service description	0.6542	0.8266	0.0026	0.6501	0.8389	0.8710	0.6163	0.3707



Hình 4. Kết quả canh các mục của hai ontology

V. KẾT LUẬN

Bài báo đã trình bày cơ sở lý thuyết bao gồm độ phức tạp Kolmogorov, khoảng cách thông tin được chuẩn hóa, khoảng cách nén được chuẩn hóa và khoảng cách Google. Phân bố xác suất, ngữ nghĩa và công thức tính của độ đo Google cũng như các tính chất của nó cũng đã được đề cập và bàn luận chi tiết. Hai trong số các ứng dụng tiêu biểu của độ đo Google được giới thiệu. Tuy nhiên các kết quả lý thuyết và thực nghiệm trong bài báo này còn nhiều điểm hạn chế. Việc khảo sát, nghiên cứu và bàn luận sâu hơn về mặt lý thuyết và thực hiện các thực nghiệm với nhiều các ứng dụng khác của độ đo Google là các công việc tương lai.

VI. LỜI CẢM ƠN

Công trình được tài trợ bởi Quỹ phát triển Khoa học Công nghệ (nhóm B) của Đại học Quốc gia Hà Nội thông qua đề tài có mã số QG.14.13 (2014-2015).

VII. TÀI LIỆU THAM KHẢO

[1] C. H. Bennett , P. Gacs , Ming Li , P. M.B. Vitanyi, W. H. Zurek, Information distance, IEEE Transactions on Information Theory, v.44 n.4, p.1407-1423, July 1998.

[2] Rudi Cilibrasi , Paul Vitányi , Ronald De Wolf, Algorithmic Clustering of Music Based on String Compression, Computer Music Journal, v.28 n.4, p.49-67.

[3] M. Li, J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An Information-Based Sequence Distance and Its Application to Whole Mitochondrial Genome Phylogeny," Bioinformatics, vol. 17, no. 2, pp. 149-154, 2001.

[4] Ming Li , Xin Chen , Xin Li , Bin Ma , P. M.B. Vitanyi, The similarity metric, IEEE Transactions on Information Theory, v.50 n.12, p.3250-3264, December 2004.

[5] Ming Li , Paul Vitányi, An introduction to Kolmogorov complexity and its applications (2nd ed.), Springer-Verlag New York, Inc., Secaucus, NJ, 1997.

[6] Egidio Terra , C. L. A. Clarke, Frequency estimates for statistical word similarity measures, Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, p.165-172, May 27-June 01, 2003, Edmonton, Canada.

[7] M.E. Lesk, "Word-Word Associations in Document Retrieval Systems," Am. Documentation, vol. 20, no. 1, pp. 27-38, 1969.

[8] Pang-Ning Tan, Vipin Kumar , Jaideep Srivastava, Selecting the right interestingness measure for association patterns, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, July 23-26, 2002, Edmonton, Alberta, Canada.

[9] R. Cilibrasi , P. M.B. Vitanyi, Clustering by compression, IEEE Transactions on Information Theory, v.51 n.4, p.1523-1545, April 2005.

GOOGLE SIMILARITY DISTANCE FOR DATA INTEGRATION

Ngoc Trinh Vu, Quang Thuy Ha, Trong Hieu Tran

ABSTRACT *- Measurement theory has emerged as one of the important issues and discussed in a lot of research work in Computer Science. It is applied in a wide range from Data Retrieval, Data Mining to Knowledge Integration, Recognition and Machine Learning. Obtaining good measures that reflect in a subtle way the difference of the concepts, terminology and entities in a particular context is urgently needed and has high practical applicability. In this paper we introduce such a measure, Google similarity distance. To this end, detailed theoretical basis is discussed, properties are pointed out and some applications are presented.*