(12) **BẢN MÔ TẢ GIẢI PHÁP HỮU ÍCH THUỘC BẰNG ĐỘC QUYỀN GIẢI PHÁP HỮU ÍCH**

(19) **Cộng hòa xã hội chủ nghĩa Việt Nam (VN)**
CỤC SỞ HỮU TRÍ TUỆ

(11) **2-0002323**

(51)[7] **G06F 17/30**; G06Q 30/02; G06F 17/21; G06F 17/28

(13) **Y**

(21) 2-2013-00222 (22) 06/09/2013

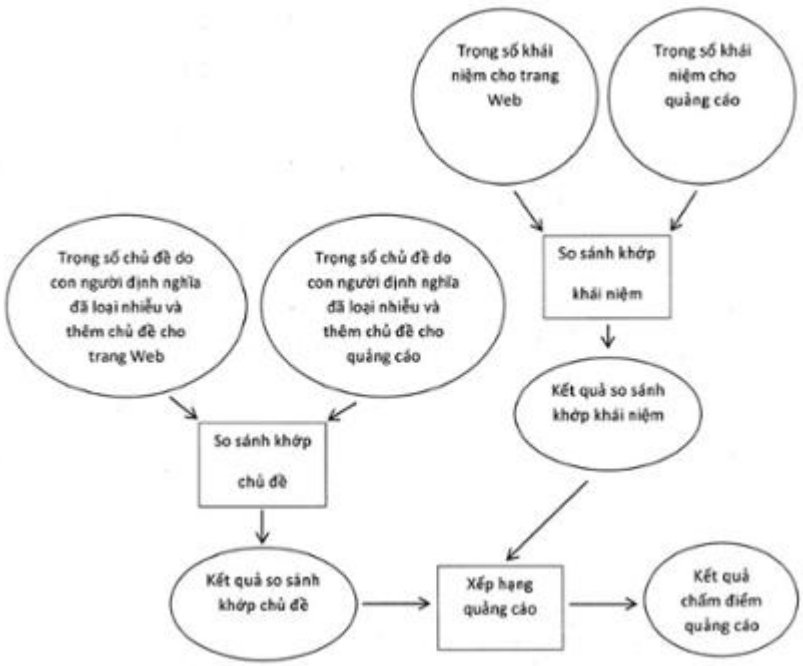
(45) 25/06/2020 387 (43) 25/12/2013 309A

(73) Viện Nghiên cứu Công nghệ FPT (VN)
Số 8 Tôn Thất Thuyết, huyện Từ Liêm, thành phố Hà Nội

(72) Phan Xuân Hiếu (VN).

(54) QUY TRÌNH CHỌN QUẢNG CÁO PHÙ HỢP NHẤT TRONG SỐ CÁC QUẢNG CÁO CÓ SẴN, THEO CHỦ ĐỀ VÀ KHÁI NIỆM VỚI NỘI DUNG VĂN BẢN TIẾNG VIỆT

(57) Sáng chế đề xuất quy trình chọn ra được quảng cáo có nội dung phù hợp nhất, về mặt chủ đề và về mặt khái niệm, với nội dung văn bản mà quảng cáo đó được hiển thị bên cạnh, để người đọc theo dõi cả nội dung văn bản và nội dung quảng cáo bên cạnh đó. Nội dung của cả quảng cáo và văn bản được xử lý để tách thành dãy các từ. Từ dãy các từ này, trọng số của các chủ đề và các khái niệm ứng với quảng cáo và ứng với nội dung văn bản được xây dựng. So sánh khớp trọng số các chủ đề giữa quảng cáo và nội dung văn bản cho ra điểm thể hiện mức tương đồng nội dung về mặt chủ đề của quảng cáo. So sánh khớp trọng số các khái niệm giữa quảng cáo và nội dung văn bản cho ra điểm thể hiện mức tương đồng nội dung về mặt khái niệm của quảng cáo. Kết hợp hai điểm trên cho ra điểm xếp hạng độ phù hợp của quảng cáo. Các quảng cáo có điểm xếp hạng cao nhất được chọn. Việc chọn ra quảng cáo có độ phù hợp cao về mặt nội dung đối với văn bản người đọc đang theo dõi làm tăng khả năng người đọc quan tâm đến nội dung quảng cáo và đọc nội dung quảng cáo, tăng khả năng lan truyền thông tin của quảng cáo tới người dùng.



Lĩnh vực kỹ thuật được đề cập

Sáng chế đề cập đến quy trình lựa chọn nội dung quảng cáo tiếng Việt phù hợp theo cả chủ đề và khái niệm với nội dung văn bản tiếng Việt, ví dụ như nội dung trang Web tiếng Việt.

Tình trạng kỹ thuật của sáng chế

Các nội dung quảng cáo đi kèm với các trang Web có nội dung tiếng Việt hiện nay chủ yếu được đưa ra theo thỏa thuận giữa người cung cấp nội dung trang Web và người muốn quảng cáo. Khi số lượng nội dung cần quảng cáo lớn hơn nhiều so với lượng nội dung trang Web, các trang quảng cáo sẽ được lựa chọn theo nhiều tiêu chí để hiển thị ra trên trang Web. Trong một số mô hình hoạt động, nội dung quảng cáo cần phù hợp với nội dung trang Web mà người dùng đang xem, tức là phù hợp với nhu cầu và sở thích của người dùng, để có được cơ hội cao hơn trong việc người dùng quan tâm đến nội dung quảng cáo và tìm hiểu sâu hơn về thông tin được quảng cáo. Việc tự động chọn nội dung quảng cáo phù hợp sẽ làm giảm nhân lực biên soạn nội dung quảng cáo và giúp tăng năng suất và hiệu quả hoạt động của lĩnh vực quảng cáo trực tuyến.

Một số ít trang Web tiếng Việt đã có hệ thống tự động chọn lọc từ kho chứa sẵn các nội dung cần quảng cáo để xuất ra các nội dung quảng cáo phù hợp nhất với nội dung trang Web đang hiển thị cho người dùng. Tuy nhiên các phương pháp hiện có mới chỉ dừng ở việc so sánh khớp chủ đề được trích chọn tự động từ nội dung trang Web với chủ đề được trích chọn tự động từ mô tả của quảng cáo. Việc chỉ sử dụng chủ đề mà không sử dụng các khái niệm chi tiết hơn có thể làm giảm độ chính xác của sự so sánh khớp, dẫn đến tỷ lệ chọn ra quảng cáo không phù hợp với nội dung trang Web cao hơn.

Bản chất kỹ thuật của sáng chế

Mục đích của sáng chế là đề xuất quy trình lựa chọn quảng cáo khắc phục được các nhược điểm nêu trên, quy trình này phù hợp không chỉ theo chủ đề mà còn theo các khái niệm chi tiết của nội dung trang Web tiếng Việt, nhờ đó tăng khả năng có được quảng cáo có nội dung liên quan tới nội dung trang Web so với phương pháp chỉ sử dụng so sánh khớp chủ đề.

Quy trình theo sáng chế bao gồm các bước:

bước 1: phân tích, suy diễn ra chủ đề và khái niệm của nội dung văn bản trong trang Web tiếng Việt và trong mô tả quảng cáo tiếng Việt;

bước 2: so sánh khớp cả chủ đề lẫn khái niệm của trang Web và nội dung quảng cáo và xếp hạng mức độ phù hợp của nội dung quảng cáo với nội dung trang Web.

Trong bước 1, ngoài việc phân tích, suy diễn ra chủ đề, khái niệm của nội dung văn bản trong trang Web tiếng Việt và trong mô tả quảng cáo tiếng Việt cũng được phân tích, suy diễn. Đây là điểm cải tiến so với các hệ thống quảng cáo đã có trước đó. Bước 1 cũng có điểm cải tiến nữa là có thêm quy trình loại bỏ chủ đề nhiễu khi suy diễn chủ đề, và có thêm bước ánh xạ từ chủ đề do máy sinh ra thành chủ đề do con người định nghĩa, các quy trình này đều có tác dụng mang lại kết quả phân tích chủ đề chính xác hơn.

Trong bước 2, quy trình so sánh khớp các khái niệm đơn và phức, cũng như quy trình thực hiện chấm điểm quảng cáo kết hợp cả so sánh khớp chủ đề và so sánh khớp khái niệm, đây là các điểm mới so với các hệ thống hiện có.

Cụ thể, quy trình chọn quảng cáo phù hợp nhất, trong số các quảng cáo có sẵn, theo chủ đề và khái niệm với nội dung một văn bản tiếng Việt, có thể thực hiện với sự hỗ trợ bởi máy tính điện tử, bước chọn này bao gồm các bước sau:

với mỗi nội dung quảng cáo:

đưa nội dung tiếng Việt của quảng cáo, *Na*, và nội dung tiếng Việt của văn bản, *Nd*, qua các tiến trình dưới đây để thu được trọng số của các chủ đề và khái niệm ứng với chúng:

các nội dung được tách thành từng câu, thành *NCa* và *NCd*, tại các vị trí dấu ngắt câu ở đó có xác suất dấu ngắt này, cùng với các ký tự quanh nó, thể hiện sự ngắt câu cao hơn là không ngắt câu, ví dụ theo kỹ thuật Entropy cực đại của lĩnh vực học máy;

nội dung ở từng câu được ngắt ra thành từng từ, *NTa* và *NTd*, theo kỹ thuật gán nhãn phần bắt đầu và tiếp nối của từ, với nhãn được chọn theo xác suất cao nhất ứng với đoạn văn bản được gán nhãn và đoạn văn bản ngay trước đó và các thông tin liên quan đến các đoạn văn bản này, ví dụ theo kỹ thuật trường ngẫu nhiên điều kiện của lĩnh vực học máy;

dãy các từ thu được ở tiến trình trên, *NTa* và *NTd*, được đưa vào bổ sung cho lượng lớn dãy từ có sẵn trong văn bản tiếng Việt để thu được trọng số chủ đề sinh tự động cho từng dãy từ, bao gồm trọng số chủ đề sinh tự động cho dãy từ thu được ở tiến trình trên, gọi là *TAa* và *TAd*, theo kỹ thuật phân bổ Dirichlet ẩn của lĩnh vực học máy;

trọng số sinh tự động *TAa* và *TAd* được chuyển thành trọng số theo chủ đề do con người định nghĩa, gọi là *Ta* và *Td*, bằng cách nhóm các trọng số sinh tự động thành từng nhóm theo ánh xạ đã có sẵn, rồi đưa từng nhóm trọng số sinh tự động qua hàm số f nhất định để thu được trọng số theo chủ đề do con người định nghĩa, ví dụ f là hàm chọn ra giá trị lớn nhất trong các giá trị đầu vào;

các trọng số ứng với chủ đề do con người định nghĩa có giá trị từ nhỏ đến lớn dần, trong *Ta* và *Td*, được đặt lại giá trị bằng 0%, lần lượt, cho đến khi gặp một trong 3 điều kiện sau:

tổng trọng số đã bị loại vượt quá một ngưỡng nhất định, chẳng hạn 70%;

tổng số các trọng số bị loại vượt quá ngưỡng nhất định, chẳng hạn 10;

trọng số đang được xem xét loại bỏ lớn hơn r lần so với trọng số vừa bị loại trước đó, với r là một ngưỡng nhất định lớn hơn 1, chẳng hạn $r = 1,1$;

với từng trọng số đã được thiết lập bằng 0%, ứng với chủ đề do con người định nghĩa thứ j, đặt lại giá trị trọng số mới bằng giá trị cực đại trong số các giá trị $t_i x_{ij}$ với t_i là trọng số ứng với các chủ

đề do con người định nghĩa thứ i chưa bị loại và x_{ij} là một hệ số có sẵn trong một ma trận vuông kích thước $M \times M$ với M là tổng số các chủ để do con người định nghĩa;

chia các trọng số trong Ta và Td thu được sau các hoạt động nêu trên cho tổng của chúng;

thu nhận các trọng số khái niệm Ka và Kd, lần lượt là tần xuất mà các danh từ hoặc bộ danh từ, trong số các danh từ riêng và danh từ chung và và bộ danh từ riêng và bộ danh từ chung có sẵn, xuất hiện trong các dãy các từ NTa và NTd;

trọng số ứng với các khái niệm là danh từ riêng hoặc bộ danh từ riêng được nhân thêm với một hệ số lớn hơn 1;

bước thứ hai: so sánh khớp trọng số chủ đề và trọng số khái niệm giữa quảng cáo và văn bản, theo các tiến trình dưới đây:

thu lấy giá trị x thể hiện sự tương đồng về mặt chủ đề giữa nội dung quảng cáo và nội dung văn bản, theo công thức:

$$x = \Sigma_i \ln(Ta_i / Td_i)\, Ta_i + \Sigma_i \ln(Td_i / Ta_i)\, Td_i$$

với Σ_i là tổng với i chạy từ 1 đến tổng số chủ đề, Ta_i là trọng số trong Ta ứng với chủ đề thứ i, Td_i là trọng số trong Td ứng với chủ đề thứ i, và nếu với một i nào đó mà Ta_i=0 hoặc Td_i=0 thì số hạng tương ứng trong hai tổng đều quy ước bằng 0;

thu lấy giá trị y thể hiện sự tương đồng về mặt khái niệm giữa nội dung quảng cáo và nội dung văn bản, theo công thức:

$$y = \log_2(N+1)\, \Sigma_i\, Ka_i\, Kd_i$$

với Σ_i là tổng với i chạy từ 1 đến tổng số chủ đề, Ka_i là trọng số trong Ka ứng với khái niệm thứ i, Kd_i là trọng số trong Kd ứng với khái niệm thứ i, N là số lượng các chỉ số i mà ứng với đó cả Ka_i và Kd_i đều khác 0;

chọn lấy các quảng cáo, trong số các quảng cáo có sẵn, có giá trị điểm xếp hạng, $đ$, cao nhất, với điểm $đ$ của mỗi quảng cáo được tính theo:

$$đ = \alpha\, x + (1 - \alpha)\, (\langle x \rangle / \langle y \rangle)\, y$$

với α là một hệ số không âm nhỏ hơn hoặc bằng 1, và $<x>$ là giá trị trung bình của tất cả các giá trị x trong toàn bộ các quảng cáo có sẵn và $<y>$ là giá trị trung bình của tất cả các giá trị y trong toàn bộ các quảng cáo có sẵn.

Mô tả vắn tắt các hình vẽ

Hình 1 là sơ lưu đồ thể hiện các bước trong quy trình chọn lọc quảng cáo phù hợp với trang Web;

Hình 2 là lưu đồ thể hiện bước hai trong quy trình chọn lọc quảng cáo phù hợp với trang Web.

Mô tả chi tiết sáng chế

Phần mô tả dưới đâytrình bày chi tiết phương án được ưu tiên cho quy trình lựa chọn quảng cáo phù hợp không chỉ theo chủ đề mà còn theo các khái niệm chi tiết của nội dung trang Web tiếng Việt. Để lựa chọn được quảng cáo phù hợp, nội dung của các quảng cáo được so sánh với nội dung của trang Web, qua hai bước chính dưới đây, để thu được điểm xếp hạng cho mỗi quảng cáo. Điểm xếp hạng quảng cáo được thực hiện theo tiêu chí: quảng cáo có điểm cao là quảng cáo phù hợp với nội dung trang Web hơn cả và sẽ được chọn để trình bày cùng với nội dung văn bản của trang Web. Hai bước chính của quy trình là:

bước 1: phân tích, suy diễn ra chủ đề và khái niệm của nội dung văn bản trong trang Web tiếng Việt và trong mô tả quảng cáo tiếng Việt;

bước 2: so sánh khớp cả chủ đề lẫn khái niệm của trang Web với nội dung quảng cáo và xếp hạng mức độ phù hợp của nội dung quảng cáo với trang Web

Hình 1 thể hiện lưu đồ của tiến trình xử lý thông tin trong bước 1. Trước hết, nội dung văn bản tiếng Việt, ví dụ như của trang Web, hoặc của mô tả quảng cáo, được đưa qua tiến trình thứ nhất, gọi là tiến trình tiền xử lý về mặt ngôn ngữ.

Tiến trình tiền xử lý văn bản tiếng Việt được thực hiện qua hai giai đoạn.

Ở giai đoạn thứ nhất, văn bản được tách thành từng câu. Có thể sử dụng kỹ thuật đơn giản là tìm kiếm các dấu ngắt câu, như dấu chấm, để ngắt văn bản tại những vị trí này. Tuy nhiên thực tế sự tồn tại dấu ngắt câu đôi khi không nhất thiết ứng với việc câu tiếng Việt được ngắt tại đó, ví dụ như câu "*TS. Nguyễn Đức Thành đã tham dự hội nghị này.*" không bị ngắt tại dấu chấm sau chữ "*TS*". Do đó, có thể thực hiện theo kỹ thuật học máy có độ chính xác cao hơn, trong đó, trước hết xây dựng cho mỗi loại dấu ngắt câu hàm xác suất $p(ngắt, x)$ mà dấu ngắt câu đó trong ngữ cảnh x đúng là vị trí ngắt câu, và hàm xác suất $p(không\_ngắt, x)$ mà dấu ngắt câu đó trong ngữ cảnh x không là vị trí ngắt câu; sau đó lần lượt xem xét từng dấu ngắt câu trong văn bản và ngữ cảnh quanh nó rồi tính hai xác suất trên, nếu xác xuất ngắt cao hơn thì ngắt câu tại vị trí của dấu ngắt câu đang xét. Ngữ cảnh x có thể là các ký tự nằm trước và sau dấu ngắt câu và/hoặc các tính chất của các ký tự đó – như việc ký tự đó có được viết hoa không. Việc xây dựng hàm xác suất $p(ngắt ,x)$ và $p(không\_ngắt, x)$ có thể dựa trên một số lượng lớn mẫu có dấu ngắt câu và ngữ cảnh quanh nó, trích từ các câu tiếng Việt sẵn có, đã được người am hiểu tiếng Việt xác định mẫu nào đúng là vị trí ngắt câu và mẫu nào không phải là vị trí ngắt câu, theo kỹ thuật, chẳng hạn như, kỹ thuật Entropy cực đại được mô tả trong bài báo "*A maximum entropy approach to natural language processing*" của các tác giả A. Berger, A. Della Pietra, và J. Della Pietra đăng trong tạp chí *Computational Linguistics*, trang 39-71, No.1, Vol.22, năm 1996.

Sau khi có được từng câu, các câu tiếng Việt được đưa vào giai đoạn tách thành các từ. Trong tiếng Việt, một từ có thể được cấu tạo từ một tiếng hay nhiều tiếng. Do đó, không thể sử dụng kỹ thuật đơn giản là tìm kiếm các dấu cách, để tách từ tại các vị trí này. Ví dụ như từ "công đoạn" gồm hai tiếng, thường không tách được thành hai từ là "công" và "đoạn" có nghĩa riêng và đứng cạnh nhau. Có thể dùng kỹ thuật gán nhãn từ trên kết quả cắt rời bằng dấu cách và các dấu câu khác, để tách từ tiếng Việt, sử dụng ba nhãn B, I và O; trong đó B là nhãn cho biết phần văn bản tương ứng bắt đầu một từ, I là nhãn cho biết phần văn bản tương ứng vẫn nằm trong từ đã có nhãn B hoặc I trước đó, O là nhãn cho biết phần văn bản tương ứng không phải là từ. Ví dụ, đoạn văn bản "*Các câu tiếng Việt được đưa vào công đoạn tách từ.*" sẽ có nhãn:

Các	*câu*	*tiếng*	*Việt*	*được*	*đưa*	*vào*	*công*	*đoạn*	*tách*	*từ*	.
B	B	B	I	B	B	B	B	I	B	I	O

Với mỗi một đoạn văn bản, gọi $p(y, x)$ là xác suất xuất hiện đồng thời nhãn y – với y có thể nhận giá trị B, I hoặc O – và ngữ cảnh x của đoạn văn bản này, với ngữ cảnh x có thể bao gồm đoạn văn bản này cùng với các đoạn văn bản đã được gán nhãn ngay trước nó và các nhãn hoặc tính chất của chúng. Nhãn y được gán vào một đoạn văn bản là nhãn có xác suất $p(y, x)$ cao nhất, trong ngữ cảnh x đã biết. Việc xây dựng hàm xác suất $p(y, x)$ có thể dựa vào một số lượng lớn các mẫu câu tiếng Việt có đã được gán nhãn và ngữ cảnh quanh nó, bởi những người am hiểu tiếng Việt, theo kỹ thuật, chẳng hạn như, trường ngẫu nhiên điều kiện – *Conditional Random Field* – được mô tả trong bài báo "*Conditional random fields: Probabilistic models for segmenting and labeling sequence data*" của các tác giả J. Lafferty, A. McCallum, và F. Pereira đăng trong kỷ yếu hội nghị *International Conference on Machine Learning*, trang 282-289, năm 2001.

Văn bản tiếng Việt, sau khi hoàn thành tiến trình trên, là dãy các từ tiếng Việt. Kết quả này được đưa vào hai chuỗi tiến trình, có thể tiến hành song song là: chuỗi tiến trình suy diễn chủ đề và tiến trình suy diễn khái niệm.

Trong chuỗi tiến trình suy diễn chủ đề, đầu vào là dãy từ tiếng Việt của văn bản và đầu ra là tập các trọng số ứng với các chủ đề đã được những người có hiểu biết về tiếng Việt xây dựng từ trước. Ví dụ, một văn bản viết về điện thoại di động Iphone 5 sẽ có trọng số cao ứng với những chủ đề như "Thiết bị di động thông minh", "Viễn thông", "Điện tử" và có thể có trọng số thấp ứng với những chủ đề như "Tham nhũng", "Mại dâm". Kết quả đầu ra của ví dụ trên có thể ở dạng { {"Thiết bị di động thông minh", 70%}, {"Viễn thông", 10%}, {"Điện tử", 11%}, ... , {"Tham nhũng", 0.1%}, {"Mại dâm", 0.2%}, ...}. Chuỗi tiến trình để thu được đầu ra như đã nêu gồm có 4 tiến trình con: suy diễn chủ đề sinh tự động, ánh xạ sang chủ đề do con người định nghĩa, loại chủ đề nhiễu, và cuối cùng là thêm chủ đề liên quan.

Tiến trình con suy diễn chủ đề sinh tự động nhận đầu vào là dãy từ tiếng Việt của văn bản và đầu ra là tập các trọng số ứng với các chủ đề đã được xây dựng từ trước không phải bởi con người mà bởi máy tính. Các chủ đề này được máy tính sinh ra từ một lượng lớn các văn bản tiếng Việt theo phương pháp phân bổ Dirichlet ẩn – *Latent Dirichlet Allocation* hay LDA như mô tả trong bài báo *Latent Dirichlet Allocation* của các tác giả David Blei, Andrew Ng và Michael Jordan in trong tạp chí *Journal of Machine Learning Research* số 3 năm 2003,

trang 993 đến 1022. Chẳng hạn, sử dụng khoảng 200 nghìn bài báo của báo điện tử VnExpress, có khoảng 300 chủ đề được máy tính sinh ra tự động theo phương pháp LDA. Phương pháp LDA được chọn vì ngoài việc sinh được ra các chủ đề tự động, chẳng hạn 300 chủ đề trong ví dụ trên, nó còn cùng lúc cho ra được trọng số của mỗi văn bản huấn luyện, tức mỗi bài báo trong 200 nghìn bài báo ở ví dụ trên, ứng với từng chủ đề. Do đó, việc sử dụng các chủ đề do máy tạo ra tự động theo phương pháp LDA, tiện lợi hơn so với việc sử dụng chủ đề do con người tạo ra, trong tiến trình con này, vì để ra tập các trọng số ứng với các chủ đề này cho một dãy từ tiếng Việt đầu vào, chỉ cần lặp lại phương pháp LDA, trong đó dãy từ đầu vào được coi như văn bản bổ sung cho, chẳng hạn, 200 nghìn văn bản đã được huấn luyện trước đó, và thu được trọng số đầu ra của văn bản bổ sung này ứng với từng chủ đề sinh tự động.

Tuy nhiên các chủ đề tự động sinh ra của máy tính không nhất thiết mang ý nghĩa hoàn toàn độc lập với nhau, thậm chí đôi khi có thể không ứng với ngữ nghĩa thường dùng trong tiếng Việt. Tiến trình con tiếp theo ánh xạ từ chủ đề sinh tự động sang chủ đề do con người định nghĩa. Tập chủ đề do người có kiến thức tiếng Việt định nghĩa ra, có thể có số lượng khác, thường là nhỏ hơn, số lượng các chủ đề do máy tự động sinh ra. Cách xây dựng các chủ đề do con người định nghĩa, có thể là đọc lại từng nhóm văn bản, mỗi nhóm có trọng số cao ứng với một chủ đề sinh tự động, nắm bắt được ý chính chung của nhóm, gọi tên ý chính này và gán cho chủ đề sinh tự động. Hai hoặc nhiều hơn các chủ đề sinh tự động có thể mang cùng một tên ý chính. Ví dụ: các chủ đề sinh tự động số 30, 52, 116, 117, 256, 272 trong tập 300 chủ đề sinh tự động ở ví dụ nêu trên đều có ý chính là chủ đề "Bóng đá". Ánh xạ tương ứng từ tập các chủ đề sinh tự động tới tập các chủ đề do con người định nghĩa được lưu lại và dùng lại trong tiến trình con này. Trong tiến trình con này, trọng số của dãy từ tiếng Việt ứng với các chủ đề do con người định nghĩa được tính bằng cách lần lượt đi qua từng chủ đề do con người định nghĩa. Với mỗi chủ đề do con người định nghĩa, gọi là C, xác định tất cả các chủ đề sinh tự động, chẳng hạn k chủ đề sinh tự động, ứng với C theo ánh xạ đã lưu; đồng thời xác định các trọng số của dãy từ đầu vào ứng với các chủ đề sinh tự động này, gọi là t_i, với i là chỉ số chạy từ 1 đến k. Trọng số của dãy từ đầu vào ứng với chủ đề C, gọi là T, có thể được tính theo một trong hai cách. Cách 1, T là trọng số lớn nhất trong các trọng số t_i. Cách 2, T là tổng tất cả các trọng số t_i rồi chia cho $\log(k+1)$.

Diễn giải theo cách khác nhưng tương đương, các trọng số sinh tự động được nhóm lại thành từng nhóm. Mỗi nhóm ứng với một chủ đề do con người định nghĩa. Ví dụ: các chủ đề sinh tự động số 30, 52, 116, 117, 256, 272 trong tập 300 chủ đề sinh tự động ở ví dụ nêu trên được nhóm thành một nhóm ứng với chủ đề "Bóng đá" do con người định nghĩa. Sau đó, các trọng số trong mỗi nhóm được đưa vào hàm số nhất định để cho ra trọng số ứng với nhóm đó, tức là ứng với chủ đề do con người định nghĩa. Cụ thể, hàm số có thể là hàm lấy ra giá trị cực đại trong số các giá trị đầu vào, hoặc bằng tổng các giá trị đầu vào chia cho $\log(k+1)$ với k là số các giá trị đầu vào.

Sau tiến trình con trên, mặc dù đã thu được trọng số ứng với các chủ đề do con người định nghĩa, nhiều trọng số có thể bị sinh ra do những từ ngữ gây nhiễu trong văn bản đầu vào, hoặc không cần thiết cho các tính toán tiếp theo. Việc loại bỏ những chủ đề nhiều hoặc thừa cho các tính toán tiếp theo sẽ giúp tăng tốc độ xử lý và tăng độ chính xác cho kết quả. Tiến trình con tiếp theo, trong chuỗi tiến trình suy diễn chủ đề, là loại chủ đề nhiễu. Trong tiến trình con này, các chủ đề do con người định nghĩa có trọng số từ nhỏ đến lớn dần được loại bỏ, hoặc đặt trọng số tương ứng bằng 0%, lần lượt, cho đến khi gặp một trong 3 điều kiện sau:

- tổng trọng số của các chủ đề đã bị loại vượt qua một ngưỡng nhất định, chẳng hạn 70%;

- tổng số các chủ đề bị loại vượt quá ngưỡng nhất định, chẳng hạn 10;

- chủ đề đang xem xét để loại bỏ có trọng số lớn hơn r lần so với trọng số của chủ đề vừa bị loại trước đó, với r là một ngưỡng nhất định lớn hơn 1, chẳng hạn $r = 1,1$.

Tiến trình con cuối cùng trong chuỗi tiến trình suy diễn chủ đề là thêm chủ đề liên quan. Chủ đề liên quan là những chủ đề mà người đọc có xu hướng quan tâm, khi đọc nội dung một văn bản, mặc dù không hề được nhắc đến trong nội dung văn bản. Tiến trình con loại nhiễu nêu trên có thể loại bỏ những chủ đề không thấy có trong nội dung của văn bản, tuy nhiên, với mục đích chọn quảng cáo phù hợp, tiến trình con này sẽ thêm vào những chủ đề đã bị loại mà có thể gây mối quan tâm với người đọc văn bản này. Tương ứng với các chủ đề do con người định nghĩa, ví dụ có M chủ đề gồm C_1, C_2, ... C_M, có bảng ma trận $M \times M$ đã lưu sẵn thể hiện sự liên quan giữa những chủ đề này, với các hệ số x_{ij} thể hiện sự liên quan

giữa chủ đề C_i với chủ đề C_j. Giá trị x_{ij} chạy từ 0% đến 100%, càng lớn càng thể hiện sự liên quan – người đọc chủ đề C_i có xác suất là x_{ij} cũng quan tâm đến chủ đề C_j. Các hệ số này có thể do người hiểu biết về tiếng Việt xây dựng từ trước. Việc thêm chủ đề liên quan được thực hiện bằng cách: với từng chủ đề C_j đã bị loại hoặc bị đặt trọng số bằng 0% sau tiến trình loại nhiễu ở trên, đặt lại trọng số mới bằng giá trị cực đại trong số các giá trị $t_i x_{ij}$ với t_i là trọng số ứng với các chủ đề C_i chưa bị loại.

Các trọng số thu được sau chuỗi tiến trình suy diễn chủ đề có thể được chuẩn hóa, tức là chia với một hệ số dương để tổng các trọng số bằng 100%. Hệ số dương cần chia cho chính là tổng các trọng số trước khi được chuẩn hóa.

Tiến trình suy diễn khái niệm nhận đầu vào là dãy từ tiếng Việt của văn bản và đầu ra là tập các trọng số ứng với các khái niệm đã được xây dựng từ trước theo thống kê từ các nội dung quảng cáo và các văn bản khác. Các khái niệm là các danh từ riêng, khi đó khái niệm là *khái niệm riêng*, hoặc danh từ chung tiếng Việt, khi đó khái niệm là *khái niệm chung*, xuất hiện với tần xuất lớn hơn một ngưỡng nhất định trong các quảng cáo và các văn bản tiếng Việt khác. Tiến trình suy diễn khái niệm có thể thực hiện theo phương án đơn giản là gán trọng số ứng với mỗi khái niệm bằng với tần suất xuất hiện của khái niệm trong dãy từ tiếng Việt đầu vào.

Theo một phương án khác, trong tiến trình suy diễn khái niệm, mỗi khái niệm không chỉ đơn giản là tương ứng với một danh từ, gọi là *khái niệm đơn*, mà tương ứng với một bộ danh từ, gọi là *khái niệm phức*. Nếu bộ danh từ là các danh từ riêng thì khái niệm ứng với bộ này gọi là *khái niệm phức riêng*; nếu bộ danh từ là các danh từ chung thì khái niệm với bộ này gọi là *khái niệm phức chung*. Chẳng hạn, khái niệm phức riêng "IPad 3" tương ứng với bộ {Ipad 3, IPad 3, Ipad3, IPad3, The New Ipad, The new IPad, The new Ipad, The New IPad}. Khi đó tiến trình suy diễn khái niệm có thể thực hiện theo phương án là gán trọng số ứng với mỗi khái niệm, đơn hoặc phức, bằng với tần suất xuất hiện của nó trong dãy từ tiếng Việt đầu vào.

Cuối cùng, trọng số của các khái niệm riêng và khái niệm phức riêng được nhân tăng thêm với một hệ số lớn hơn 1, trong khi trọng số của các khái niệm chung và khái niệm phức chung không được nhân thêm với hệ số nào; do các khái

niệm riêng thường hướng tới các sản phẩm, dịch vụ cần quảng cáo cụ thể hơn, giúp cho việc chọn quảng cáo phù hợp hơn trong các tiến trình tiếp theo.

Hình 2 là lưu đồ thể hiện tiến trình xử lý thông tin trong bước 2. Đầu vào cho các tiến trình của bước 2 là các kết quả cuối cùng của bước 1, gồm: các trọng số chủ đề do người định nghĩa, đã lọc nhiễu và thêm chủ đề liên quan, cho bài viết hoặc nội dung trang Web và cho các quảng cáo có thể dùng để đăng cùng bài viết hoặc nội dung trang Web; các trọng số của các khái niệm chung và riêng cho bài viết hoặc nội dung trang Web và cho các quảng cáo có thể dùng để đăng cùng bài viết hoặc nội dung trang Web. Các đầu vào trọng số chủ đề được đưa vào tiến trình so sánh khớp chủ đề; và song song với đó, các đầu vào trọng số khái niệm được đưa vào tiến trình so sánh khớp khái niệm. Kết quả của hai tiến trình trên được đưa vào tiến trình cuối cùng là tiến trình xếp hạng quảng cáo. Kết quả cuối cùng là điểm chấm ứng với các quảng cáo; quảng cáo có điểm càng cao càng phù hợp về mặt nội dung – gồm chủ đề và khái niệm – với bài viết hoặc nội dung trang Web đang xét. Biên tập viên hoặc hệ thống quản lý nội dung trang Web có thể lựa chọn các quảng cáo có điểm cao để hiển thị ra cùng với trang Web, với kỳ vọng người đọc nội dung trang Web cũng quan tâm đến các nội dung quảng cáo được hiển thị cùng.

Tiến trình so sánh khớp chủ đề giữa một quảng cáo với nội dung trang Web nhận đầu vào là trọng số chủ đề của quảng cáo, gọi là Ta_i, và trọng số chủ đề của trang Web, gọi là Td_i - với i là chỉ số chạy từ 1 đến tổng số chủ đề - và cho ra một số thực, gọi là x; x càng lớn thì độ tương đồng, về mặt chủ đề, giữa quảng cáo và nội dung trang Web càng lớn. x được tính theo công thức:

$$x = \Sigma_i \ln(Ta_i / Td_i)\, Ta_i + \Sigma_i \ln(Td_i / Ta_i)\, Td_i$$

với Σ_i là tổng với i chạy từ 1 đến tổng số chủ đề, và nếu với một i nào đó mà Ta_i=0 hoặc Td_i=0 thì số hạng tương ứng trong hai tổng đều quy ước bằng 0.

Tiến trình so sánh khớp khái niệm giữa một quảng cáo với nội dung trang Web nhận đầu vào là trọng số khái niệm của quảng cáo, gọi là Ka_i, và trọng số chủ đề của trang Web, gọi là Kd_i - với i là chỉ số chạy từ 1 đến tổng số khái niệm - và cho ra một số thực, gọi là y; y càng lớn thì độ tương đồng, về mặt khái niệm, giữa quảng cáo và nội dung trang Web càng lớn. Số y được tính theo hai quá trình. Trong quá trình thứ nhất, thực hiện tính:

$y' = \Sigma_i Ka_i Kd_i$

với Σ_i là tổng với i chạy từ 1 đến tổng số khái niệm, hoặc, theo cách tính nhanh hơn, chỉ chạy trong số các chỉ số mà ứng với đó cả Ka_i và Kd_i đều khác 0. Trong quá trình thứ hai, gọi N là số lượng các chỉ số i mà ứng với đó cả Ka_i và Kd_i đều khác 0, $y = \log_2(N+1)\, y'$.

Trong tiến trình cuối cùng, tiến trình xếp hạng quảng cáo, đầu vào là tất cả các bộ kết quả $\{x,y\}$ ứng với tất cả các quảng cáo được dùng để chọn lọc ra cho đăng cùng nội dung trang Web, của hai tiến trình ngay trên, và đầu ra là điểm chấm của từng trang web, gọi là $đ$. Điểm $đ$, cho mỗi quảng cáo, được tính bằng:

$đ = \alpha\, x + (1 - \alpha)\, (\langle x \rangle / \langle y \rangle)\, y$

với α là một hệ số, thể hiện mức độ ưu tiên cho chủ đề, chẳng hạn $\alpha = 90\%$ nghĩa là kết quả so sánh khớp chủ đề được ưu tiên hơn trong xếp hạng quảng cáo, và $\langle x \rangle$ là giá trị trung bình của tất cả các giá trị x và $\langle y \rangle$ là giá trị trung bình của tất cả các giá trị y.

Yêu cầu bảo hộ

1. Quy trình chọn quảng cáo phù hợp nhất trong số các quảng cáo có sẵn, theo chủ đề và khái niệm với nội dung một văn bản tiếng Việt, quy trình có thể được thực hiện với sự hỗ trợ của máy tính điện tử, quy trình này bao gồm các bước sau:

với mỗi nội dung quảng cáo:

đưa nội dung tiếng Việt của quảng cáo, *Na*, và nội dung tiếng Việt của văn bản, *Nd*, qua các tiến trình con dưới đây để thu được trọng số của các chủ đề và khái niệm ứng với chúng:

nội dung được tách thành từng câu, thành *NCa* và *NCd*, tại các vị trí dấu ngắt câu ở đó có xác suất dấu ngắt này, cùng với các ký tự quanh nó, thể hiện sự ngắt câu cao hơn là không ngắt câu;

nội dung ở từng câu được ngắt ra thành từng từ, thành *NTa* và *NTd*, theo kỹ thuật gán nhãn phần bắt đầu và tiếp nối của từ, với nhãn được chọn theo xác suất cao nhất ứng với đoạn văn bản được gán nhãn và đoạn văn bản ngay trước đó và các thông tin liên quan đến các đoạn văn bản này;

dãy các từ thu được ở tiến trình trên, *NTa* và *NTd*, được đưa vào bổ sung cho một lượng lớn có sẵn các dãy từ trong văn bản tiếng Việt để thu được trọng số chủ đề sinh tự động cho từng dãy từ, bao gồm trọng số chủ đề sinh tự động cho dãy từ thu được ở tiến trình trên, gọi là *TAa* và *TAd*, theo kỹ thuật phân bổ Dirichlet ẩn của lĩnh vực học máy;

trọng số sinh tự động *TAa* và *TAd* được chuyển thành trọng số theo chủ đề do con người định nghĩa, gọi là *Ta* và *Td*, bằng cách nhóm các trọng số sinh tự động thành từng nhóm theo ánh xạ đã có sẵn, rồi đưa từng nhóm trọng số sinh tự động qua hàm số *f* nhất định để thu được trọng số theo chủ đề do con người định nghĩa;

các trọng số ứng với chủ đề do con người định nghĩa có giá trị từ nhỏ đến lớn dần, trong *Ta* và *Td*, được đặt lại giá trị bằng 0%, lần lượt, cho đến khi gặp một trong 3 điều kiện sau:

tổng trọng số đã bị loại vượt qua một giá trị ngưỡng đã chọn trước;

tổng số các trọng số bị loại vượt một giá trị ngưỡng đã chọn trước;

trọng số đang được xem xét loại bỏ lớn hơn r lần so với trọng số vừa bị loại trước đó, với r là một ngưỡng nhất định lớn hơn 1;

với từng trọng số đã bị đặt bằng 0%, ứng với chủ đề do con người định nghĩa thứ j, đặt lại giá trị trọng số mới bằng giá trị cực đại trong số các giá trị $t_i x_{ij}$ với t_i là trọng số ứng với các chủ đề do con người định nghĩa thứ i chưa bị loại và x_{ij} là một hệ số có sẵn trong một ma trận vuông kích thước $M \times M$ với M là tổng số các chủ để do con người định nghĩa;

chia các trọng số trong Ta và Td thu được sau các hoạt động nêu trên cho tổng của chúng;

thu lấy các trọng số khái niệm Ka và Kd, lần lượt là tần xuất mà các danh từ hoặc bộ danh từ, trong số các danh từ riêng và danh từ chung và và bộ danh từ riêng và bộ danh từ chung có sẵn, xuất hiện trong các dãy các từ NTa và NTd;

trọng số ứng với các khái niệm là danh từ riêng hoặc bộ danh từ riêng được nhân thêm với một hệ số lớn hơn 1;

bước thứ hai: so sánh khớp trọng số chủ đề và trọng số khái niệm giữa quảng cáo và văn bản, theo các tiến trình dưới đây:

thu nhận giá trị x thể hiện sự tương đồng về mặt chủ đề giữa nội dung quảng cáo và nội dung văn bản, theo công thức:

$$x = \Sigma_i \ln(Ta_i / Td_i)\, Ta_i + \Sigma_i \ln(Td_i / Ta_i)\, Td_i$$

với Σ_i là tổng với i chạy từ 1 đến tổng số chủ đề, Ta_i là trọng số trong Ta ứng với chủ đề thứ i, Td_i là trọng số trong Td ứng với chủ đề thứ i, và nếu với một i nào đó mà Ta_i=0 hoặc Td_i=0 thì số hạng tương ứng trong hai tổng đều quy ước bằng 0;

thu nhận giá trị y thể hiện sự tương đồng về mặt khái niệm giữa nội dung quảng cáo và nội dung văn bản, theo công thức:

$$y = \log_2(N+1)\, \Sigma_i\, Ka_i\, Kd_i$$

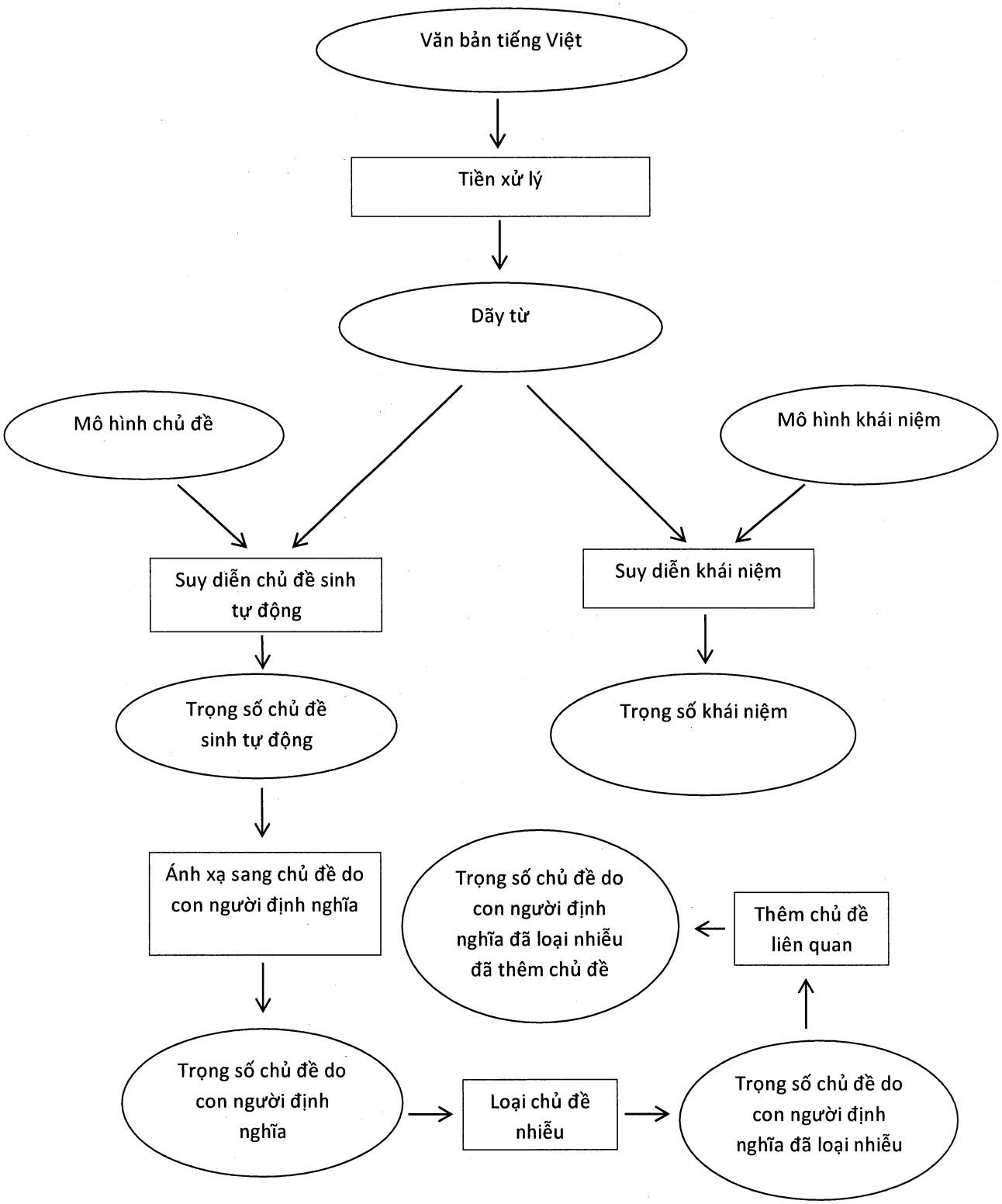
với Σ_i là tổng với i chạy từ 1 đến tổng số chủ đề, Ka_i là trọng số trong Ka ứng với khái niệm thứ i, Kd_i là trọng số trong Kd ứng với khái niệm thứ i, N là số lượng các chỉ số i mà ứng với đó cả Ka_i và Kd_i đều khác 0;

chọn ra các quảng cáo từ trong số các quảng cáo có sẵn, có giá trị điểm xếp hạng, *đ*, cao nhất, với điểm *đ* của mỗi quảng cáo được tính theo:

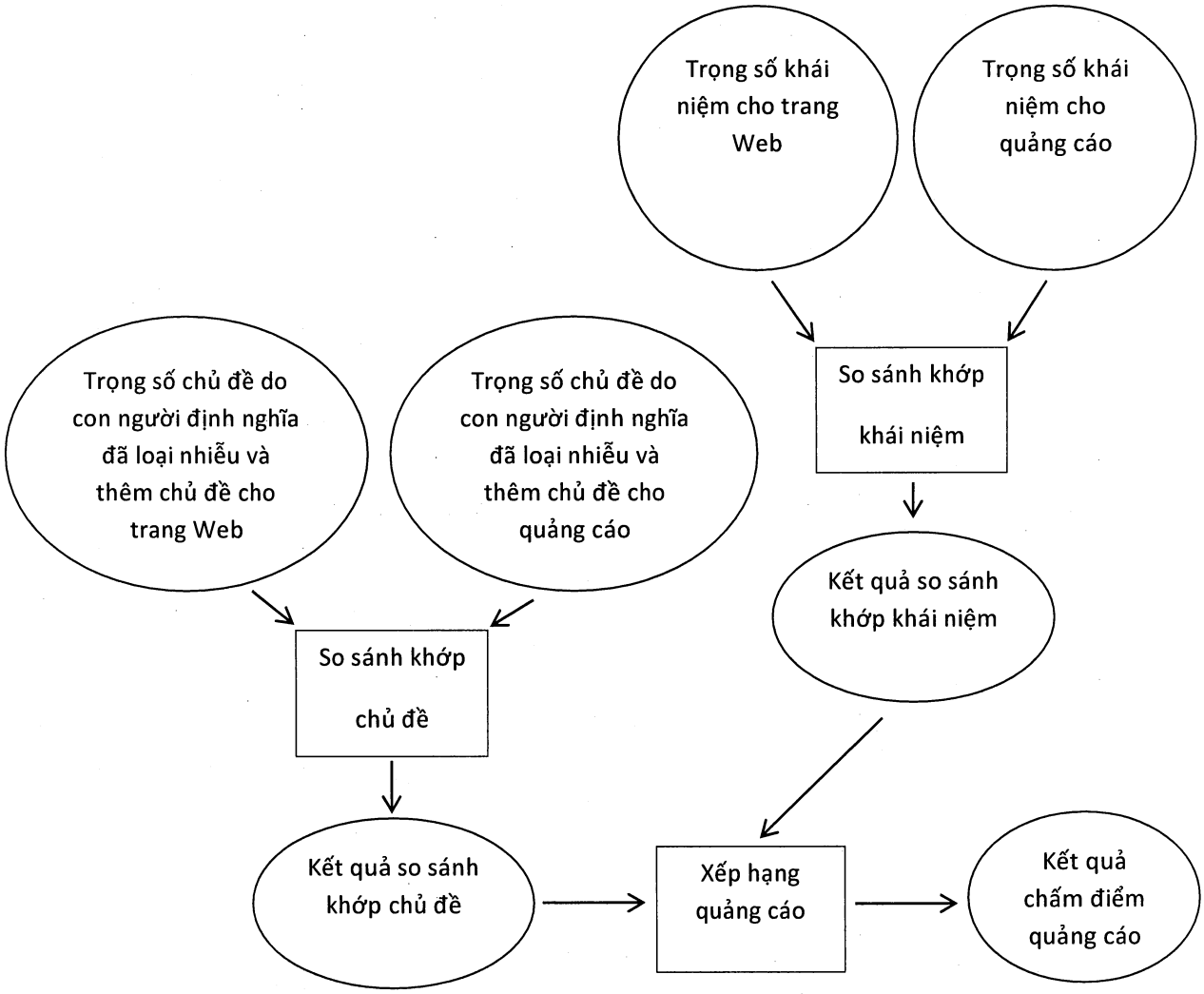
$đ = \alpha\, x + (1 - \alpha)\, (\langle x \rangle / \langle y \rangle)\, y$

với α là một hệ số không âm nhỏ hơn hoặc bằng 1, và $\langle x \rangle$ là giá trị trung bình của tất cả các giá trị x trong toàn bộ các quảng cáo có sẵn và $\langle y \rangle$ là giá trị trung bình của tất cả các giá trị y trong toàn bộ các quảng cáo có sẵn.

2. Quy trình theo điểm 1, khác biệt ở chỗ, bước nội dung được tách thành từng câu tại các vị trí dấu ngắt câu ở đó có xác suất dấu ngắt này, cùng với các ký tự quanh nó, thể hiện sự ngắt câu cao hơn là không ngắt câu, được thực hiện theo kỹ thuật Entropy cực đại của lĩnh vực học máy.
3. Quy trình theo điểm 1, khác biệt ở chỗ, bước nội dung ở từng câu được ngắt ra thành từng từ, theo kỹ thuật gán nhãn phần bắt đầu và tiếp nối của từ, với nhãn được chọn theo xác suất cao nhất ứng với đoạn văn bản được gán nhãn và đoạn văn bản ngay trước đó và các thông tin liên quan đến các đoạn văn bản này, được thực hiện theo kỹ thuật trường ngẫu nhiên điều kiện của lĩnh vực học máy.


Văn bản tiếng Việt
Tiền xử lý
Dãy từ
Mô hình chủ đề
Mô hình khái niệm
Suy diễn chủ đề sinh tự động
Suy diễn khái niệm
Trọng số chủ đề sinh tự động
Trọng số khái niệm
Ánh xạ sang chủ đề do con người định nghĩa
Trọng số chủ đề do con người định nghĩa đã loại nhiễu đã thêm chủ đề
Thêm chủ đề liên quan
Trọng số chủ đề do con người định nghĩa
Loại chủ đề nhiễu
Trọng số chủ đề do con người định nghĩa đã loại nhiễu

Hình 1


Trọng số khái niệm cho trang Web
Trọng số khái niệm cho quảng cáo
So sánh khớp khái niệm
Trọng số chủ đề do con người định nghĩa đã loại nhiễu và thêm chủ đề cho trang Web
Trọng số chủ đề do con người định nghĩa đã loại nhiễu và thêm chủ đề cho quảng cáo
Kết quả so sánh khớp khái niệm
So sánh khớp chủ đề
Kết quả so sánh khớp chủ đề
Xếp hạng quảng cáo
Kết quả chấm điểm quảng cáo

Hình 2