BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC ĐÀ NẴNG

PHẠM XUÂN THÀNH

XÂY DỰNG HỆ THỐNG QUẢNG CÁO TRỰC TUYẾN DỰA TRÊN TỪ KHÓA TIẾNG VIỆT

Chuyên ngành : Khoa học máy tính

Mã số : 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2012

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **TS. Nguyễn Thanh Bình**

Phản biện 1: **TS. Huỳnh Hữu Hưng**

Phản biện 2: **PGS.TS. Đoàn Văn Ban**

Luận văn được bảo vệ tại Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 15 tháng 12 năm 2012

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng;
- Trung tâm Học liệu, Đại học Đà Nẵng;

MỞ ĐẦU

1. Lý do chọn đề tài

Ngày nay Word Wide Web đã trở thành một kho tài nguyên dữ liệu khổng lồ về mọi lĩnh vực. Lượng truy cập và trao đổi thông tin qua Word Wide Web diễn ra liên tục tạo ra mạng lưới truyền thông bao phủ khắp toàn cầu khiến kênh truyền thông này trở thành một mảnh đất màu mỡ cho hoạt động quảng cáo trực tuyến.

Ở Việt Nam hiện nay, tốc độ tăng trưởng người dùng Internet tăng cao nhưng tổng doanh thu quảng cáo trực tuyến ở Việt Nam vẫn ở mức khá khiêm tốn - 480 tỷ đồng năm 2010, chiếm 0,4% tổng chi cho quảng cáo.

Chỉ số ngân sách quảng cáo trực tuyến hàng năm trên mỗi người sử dụng của Việt Nam hiện chỉ có 0,5 USD, kém xa so với chỉ số này ở các nước phát triển như Mỹ là 171,5 USD hoặc Trung Quốc 10 USD. Dự kiến đến năm 2015 thị trường quảng cáo trực tuyến Việt Nam mới phát triển ổn định.

Hình thức quảng cáo trực tuyến phổ biến ở Việt Nam thường dành một phần lớn diện tích trang web để hiển thị quảng cáo gây trở ngại đến việc khai thác thông tin của bạn đọc. Hình thức quảng cáo này cũng không phù hợp với các thiết bị duyệt web, có kích thước màn hình hạn chế như Smart Phone, máy tính bảng hay thiết bị giải trí truy nhập Internet khác.

Luận văn đề xuất hướng khai thác quảng cáo trực tuyến bằng cách sử dụng các từ khóa tiếng Việt ở phần văn bản của nội dung chính trang web chuyển tải quảng cáo. Hình thức là xu hướng mới, cải thiện những hạn chế quảng cáo trực tuyến hiện nay ở nước ta.

2. Mục đích nghiên cứu

Nghiên cứu, tìm hiểu kỹ thuật khai phá dữ liệu web nhằm xác định phần nội dung chính của trang web thuộc mạng quảng cáo; tiến hành nghiên cứu tách từ khóa ở nội dung đó nhằm xây dựng máy xử lý từ khóa tiếng Việt tự động, nâng cao mục tiêu hiệu quả của hệ thống quảng cáo trực tuyến sẽ xây dựng.

3. Đối tượng và phạm vi nghiên cứu

- Nghiên cứu tìm hiểu lĩnh vực quảng cáo trực tuyến và mô hình dịch vụ quảng cáo trực tuyến.
- Thực hiện khai phá dữ liệu web để xác định bóc tách nội dung chính của trang web.
- Xử lý tách từ tiếng Việt và xác định từ khóa của văn bản.
- Thiết kế, xây dựng hệ thống quảng cáo trực tuyến.

4. Phương pháp nghiên cứu

5. Ý nghĩa khoa học và thực tiễn của đề tài

Đề tài vận dụng các nghiên cứu, đề xuất phương pháp xây dựng hệ thống quảng cáo trực tuyến nhằm khai thác quảng cáo ở khía cạnh các từ khóa của nội dung văn bản trang web, là một trong những hướng đi mới của công nghệ quảng cáo trực tuyến hiện nay.

6. Cấu trúc của luận văn

Nội dung luận văn bao gồm phần mở đầu, ba chương và phần kết luận. Cuối mỗi chương có phần kết chương, cụ thể:

Chương 1: **QUẢNG CÁO TRỰC TUYẾN**. Luận văn trình bày tổng quan về lĩnh vực quảng cáo trực tuyến, các số liệu thống kê liên quan, những đặc điểm và mô hình hoạt động của hệ thống quảng cáo trực tuyến. Cũng trong chương này luận văn đề xuất mô hình xây dựng hệ thống quảng cáo trực tuyến dựa trên nền tảng là các từ khóa

ở nội dung chính của trang web, trình bày những ưu điểm hệ thống này mang lại.

Chương 2: **TÁCH NỘI DUNG CHÍNH VÀ TỪ KHÓA TIẾNG VIỆT TRÊN WEB**. Luận văn tập trung nghiên cứu kỹ thuật khai phá dữ liệu web ở lĩnh vực khai thác nội dung thông tin. Chương này thực hiện ba nhiệm vụ chính: nghiên cứu và đề xuất phương pháp bóc tách nội dung chính của trang web, thực hiện tách từ tiếng Việt và xác định từ khóa trên nội dung chính này. Nhóm các từ khóa tách được sẽ phục vụ cho phân hệ Engine tách từ khóa thuộc hệ thống quảng cáo trực tuyến. Engine này cung cấp cho người đăng quảng cáo dễ dàng chọn từ khóa liên quan đến trang web mà họ quảng cáo cũng như hệ thống quảng cáo phát mẩu quảng cáo chính xác vào phần nội dung chính trên trang web có từ khóa đã được thiết lập.

Chương 3: **XÂY DỰNG HỆ THỐNG QUẢNG CÁO TRỰC TUYẾN**. Luận văn tiến hành xây dựng hệ thống quảng cáo trực tuyến với từ khóa tiếng Việt. Hệ thống bao gồm hai thành phần chính: xây dựng Engine xử lý tách từ khóa tiếng Việt với các phương pháp đã đề xuất ở chương 2, hệ thống quản lý (Portal AdServer) và chuyển phát quảng cáo (Ad Script) lên mạng quảng cáo. Luận văn đề xuất mô hình hệ thống xây dựng, trình bày các thiết kế chức năng, sơ đồ hoạt động, cơ sở dữ liệu và mô hình triển khai hệ thống quảng cáo trực tuyến. Cuối chương là phần thử nghiệm và đánh giá kết quả quá trình thực hiện chức năng các thành phần của hệ thống quảng cáo trực tuyến.

Phần kết luận nêu những kết quả đạt được, hướng nghiên cứu trong đề xuất từ khóa tiếng Việt và phát triển hoàn thiện hệ thống quảng cáo trực tuyến đã xây dựng

CHƯƠNG 1 - QUẢNG CÁO TRỰC TUYẾN

1.1. Giới thiệu chung về quảng cáo

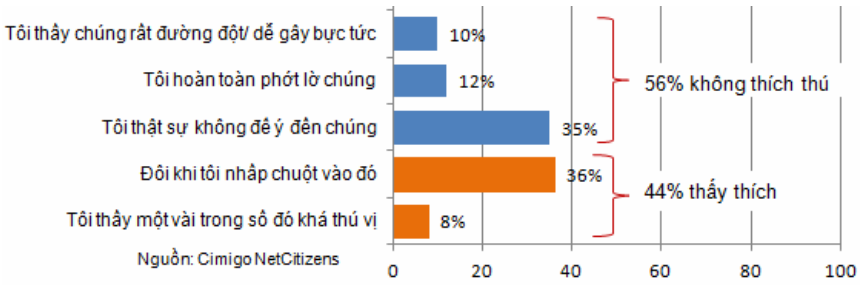
Quảng cáo là hình thức tuyên truyền được trả tiền để thực hiện việc giới thiệu thông tin về sản phẩm, dịch vụ, công ty hay ý tưởng. Quảng cáo là hoạt động truyền thông phi trực tiếp giữa người với người mà trong đó người muốn truyền thông phải trả tiền cho các phương tiện truyền thông đại chúng để đưa thông tin đến thuyết phục hay tác động đến người nhận thông tin.

1.2. Quảng cáo trực tuyến

Quảng cáo trực tuyến khác hẳn quảng cáo trên các phương tiện thông tin đại chúng khác, nó giúp người tiêu dùng có thể tương tác với quảng cáo. Nó không bị giới hạn bởi vị trí địa lý hay thời gian; truyền đạt thông tin quảng cáo ở mức độ toàn cầu tới một lượng lớn người dùng với một chi phí rất thấp.

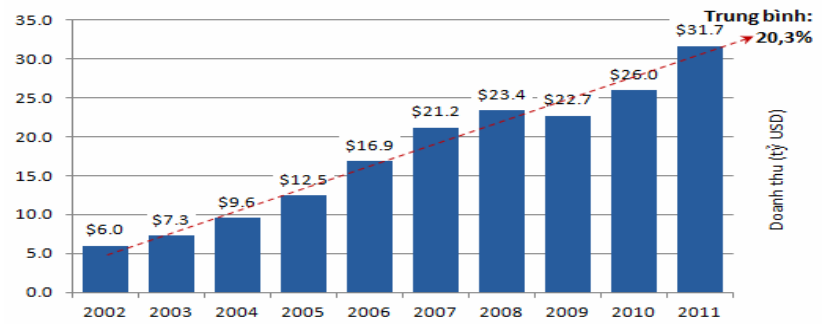
1.2.1. Internet, mạng truyền thông cho quảng cáo trực tuyến

Việt Nam là quốc gia có tỷ lệ tăng trưởng Internet nhanh nhất trong khu vực. Có 26.8 triệu người sử dụng vào thời điểm cuối năm 2010, đại diện cho 31% dân số và thái độ đối với quảng cáo trực tuyến được mô tả như sau:



Hình 1.6. Thái độ người đọc tin với quảng cáo trực tuyến

1.2.2. Sự phát triển của quảng cáo trực tuyến



Hình 1.7. Doanh thu quảng cáo trực tuyến Mỹ qua 10 năm

1.2.3. Quảng cáo trực tuyến ở Việt Nam

1.2.3.1. Số liệu thống kê

1.2.3.2. Các hình thức quảng cáo trực tuyến ở Việt Nam

Hình thức quảng cáo trực tuyến phổ biến ở Việt Nam thường dành một phần lớn diện tích trang web để hiển thị quảng cáo gây trở ngại đến việc khai thác thông tin của bạn đọc.

1.2.3.3. Phát triển quảng cáo trực tuyến ở Việt Nam là cần thiết

Thị trường quảng cáo trực tuyến ở Việt Nam tuy vẫn còn ở giai đoạn mới phát triển. Cần có nghiên cứu, xây dựng các hệ thống quảng cáo có hàm lượng công nghệ mới đáp ứng được xu thế như quảng cáo trên máy tìm kiếm hay quảng cáo theo hành vi, ngữ cảnh, quảng cáo từ khóa tiếng Việt …

1.3. Hệ thống chuyển phát quảng cáo trực tuyến

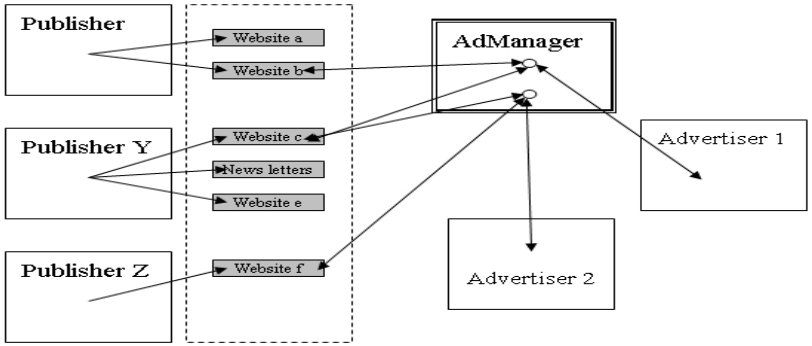
1.3.1. Đặc điểm quảng cáo trực tuyến

- Khả năng nhắm chọn
- Khả năng theo dõi
- Tính linh hoạt và khả năng phân phối

- Tính tương tác

1.3.2. Mô hình hoạt động kinh doanh quảng cáo trực tuyến

Qua nghiên cứu và khảo sát, mô hình hoạt động kinh doanh quảng cáo trực tuyến bao gồm ba thành phần chính, đó là Advertiser, Publisher và Ad Manager.



Hình 1.10. Mô hình tham gia quảng cáo trực tuyến

1.3.3. Các độ đo hiệu quả quảng cáo trực tuyến

Phương pháp đo lường hiệu quả quảng cáo chính là các tiêu chí đánh giá của ngành quảng cáo trực tuyến.

1.3.3.1. CPD

1.3.3.2. CPM

1.3.3.3. CPC

1.3.3.4. CPA

1.3.3.5. CTR

1.3.4. Mô hình quảng cáo trực tuyến đề xuất của luận văn

Luận văn đề xuất hướng khai thác quảng cáo trực tuyến bằng cách sử dụng các từ khóa tiếng Việt ở phần văn bản của nội dung chính trang web chuyển tải quảng cáo.

khi rê chuột và từ khóa được ấn định, quảng cáo sẽ xuất hiện

Hình 1.11. Ví dụ về quảng cáo từ khóa trên văn bản web [42]

Có khoảng 0,1 đến 0,2% người lướt web nhấp chuột vào các mẩu quảng cáo trên trang web. Trong khi đó tỷ lệ người đọc rê chuột và nhấp vào các thông tin quảng cáo trên văn bản web lên đến 10%. Đây là con số rất ấn tượng, phản ánh mức độ quan tâm của người đọc với thông tin quảng cáo nhờ vào khả năng nhắm tới khách hàng tiềm năng tốt hơn do quảng cáo trên văn bản web mang lại.

Mục tiêu xây dựng hệ thống cung cấp dịch vụ quảng cáo trực tuyến trên văn bản web dựa trên từ khóa tiếng Việt của luận văn này vì những ưu điểm nổi bật:

- Việc quảng cáo trên văn bản web gồm có ba bên tham gia vào một quá trình quảng cáo, gồm có: bên cung cấp dịch vụ, bên bán quảng cáo và bên mua quảng cáo.
- Thông tin quảng cáo được hiển thị trên nội dung văn bản (text) của trang web, tiếp cận với người đọc một cách tự nhiên. Quảng cáo chỉ hiện ra khi người đọc di chuột qua, họ sẽ không có cảm giác bị "bắt" xem quảng cáo.

- Việc tính chi phí quảng cáo theo CPC hay CPA giúp cho đợt quảng cáo của bên mua quảng cáo hiệu quả hơn rất nhiều so với cách tính chi phí cố định.
- Chủ động trong việc quản lý đợt quảng cáo cho bên mua quảng cáo.
- Hệ thống Engine tách từ tiếng sẽ hỗ trợ người đăng quảng cáo quyết định đặt từ khóa quảng cáo nhằm nâng cao hiệu quả quảng cáo. Engine này tự động tạo ra cơ sở dữ liệu từ khóa tương ứng với các trang web trên mạng quảng cáo của nhà cung cấp dịch vụ.

1.4. Kết chương

Chương 1 trình bày tổng quan về lĩnh vực quảng cáo trực tuyến, các số liệu thống kê cũng như tốc độ phát triển của lĩnh vực này ở Việt nam và thế giới. Cũng trong chương này, luận văn trình bày mô tả hệ thống quảng cáo trực tuyến gồm những đặc điểm, mô hình hoạt động kinh doanh quảng cáo trực tuyến, các độ đo xác định hiệu quả thực hiện quảng cáo.

Cuối cùng là mô hình luận văn đề xuất xây dựng. Hệ thống quảng cáo trực tuyến dựa trên từ khóa tiếng Việt được xây dựng dựa trên *nền tảng là phần văn bản trong khối nội dung chính của trang web*, thông qua từ khóa này, nội dung quảng cáo sẽ được chuyển tải khi người đọc nhắm vào nó.

Ở chương tiếp theo, luận văn trình bày các nghiên cứu, đề xuất phương pháp để xây dựng một Engine (máy xử lý tự động) của hệ thống quảng cáo trực tuyến có khả năng: xác định nội dung chính của trang web, tách từ tiếng Việt và xác định từ khóa.

CHƯƠNG 2 - TÁCH NỘI DUNG CHÍNH VÀ TỪ KHÓA TIẾNG VIỆT TRÊN WEB

2.1. Tổng quan chung về khai phá dữ liệu web

2.1.1. Khái niệm

2.1.2. Đặc điểm của khai phá web

2.1.2.1. Những khó khăn trong khai phá web

2.1.2.2. Thuận lợi

2.1.3. Phân loại khai phá web

2.1.3.1. Khai phá nội dung web (web content mining)

2.1.3.2. Khai phá cấu trúc web (web structure mining)

2.1.3.3. Khai phá sử dụng web (web usage mining)

2.1.4. Hướng khai phá web của luận văn

Luận văn nghiên cứu và triển khai ứng dụng thử nghiệm xử lý bóc tách thành phần chính nội dung của trang web, xử lý tách từ khóa tiếng Việt phục vụ cho hệ thống chuyển phát quảng cáo trực tuyến theo thiết kế của tác giả.

2.2. Bóc tách nội dung web

2.2.1. Tổng quan xử lý trích xuất nội dung trang web

Hình 2.3. Khối dữ liệu cần được xử lý phục vụ mục đích bài toán

2.2.2. Các phương pháp xử lý

2.2.2.1. Loại bỏ các tag HTML

2.2.2.2. Phương pháp dựa trên tỷ lệ văn vản và thẻ HTML

2.2.2.3. Phân đoạn trang web VIPS

2.2.3. Đề xuất phương pháp tách nội dung chính của luận văn

Luận văn sử dụng phương pháp phân tích cây DOM kết hợp xử lý văn bản tiếng Việt tại các node với thuộc tính mật độ câu, từ tiếng Việt, và các liên kết như sau:

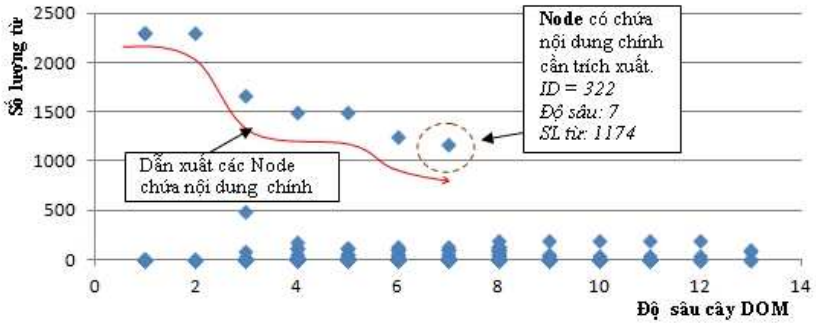
- Phân tích các tag HTML, tiến hành xây dựng cây DOM trong đó các Node được lưu giữ những thông tin đặc trưng của tag HTML mà nó chứa đựng.
- Nội dung chính của trang web bằng nội dung chính của $Node_i$ thỏa mãn:

$$Max\ \{(Số\ lượng\ từ_{Node_i} - Số\ lượng\ từ\ có\ liên\ kết_{Node_i}) x Độ\ sâu_{Node_i},\ i=1..n\}$$

- Tiến hành loại bỏ một số tag HTML bên trong Node, lưu dữ liệu được bóc tách.

Giải thuật cài đặt tách nội dung chính của luận văn

Tác giả lập trình thử nghiệm thực hiện trích xuất nội dung trên báo một trang web báo Tuổi trẻ Online, phân tích kết quả thu được:



Hình 2.10. Phân tích cây DOM với trang tin báo Tuổi trẻ Online

Kết quả phương pháp đề xuất

Bảng 2.1. Kết quả thử nghiệm trích xuất nội dung chính của trang web

Các trang web	Độ chính xác trung bình	Độ bao phủ trung bình	Độ đo F1
10 trang tin vnexpress.net	0.9871	0.9784	0.9827
10 trang tin dantri.vn	0.9717	0.9242	0.9474
10 trang tin báo vnmedia.vn	0.9736	0.9836	0.9786
10 trang tin NewYork Times	0.9867	0.9748	0.9790
10 trang tin báo tuoitre.vn	0.9826	0.9716	0.9771

Sau khi có kết quả trích xuất nội dung chính, luận văn tiến hành nghiên cứu xử lý tách từ tiếng Việt từ nội dung đó.

2.3. Xử lý tách từ khóa tiếng Việt

Mục tiêu xử lý tách từ khóa tiếng Việt của luận văn nhằm thực hiện tìm kiếm tập hợp các từ khóa có thể có trong tập dữ liệu các nội dung chính được trích xuất từ tập hợp tất cả các trang web của mạng quảng cáo.

2.3.1. Tách từ tiếng Việt

2.3.1.1. Phương pháp tách từ tiếng Việt dựa trên thống kê Internet

2.3.1.2. Phương pháp khớp tối đa (Maximum Matching)

2.3.1.3. Phương pháp học dựa trên sự cải biến

2.3.2. Tách từ khóa tiếng Việt

2.3.2.1. Hướng tiếp cận dựa vào thống kê

Phương pháp tần số từ

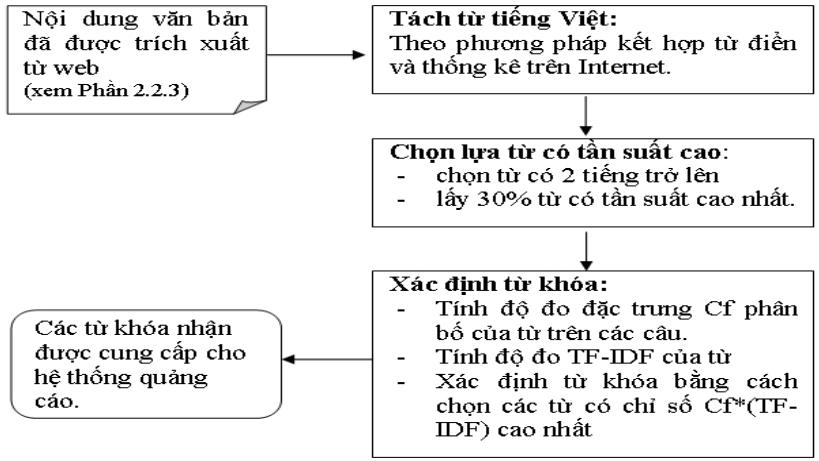
Phương pháp lấy trọng số từ dựa vào các thông tin khác

2.3.2.2. Phương pháp dựa trên máy học

2.3.3. Đề xuất phương pháp của luận văn

Phương pháp tách từ tiếng Việt của luận văn theo hướng kết hợp từ điển tiếng Việt và độ đo sự liên quan từ của từ vựng dựa vào thống

kê trên Internet. Kế tiếp, để xác định từ khóa, luận văn tiếp cận dựa vào thống kê phân bố các từ tiếng Việt trên các câu với độ đo TF-IDF để xác định từ khóa. Mô hình thực hiện như sau:



2.3.3.1. Tách từ tiếng Việt

Luận văn cài đặt giải thuật tách từ tiếng Việt dựa vào phương pháp khớp tối đa để so sánh tập các từ tạo ra và dữ liệu các từ tiếng Việt có số lượng tiếng tương ứng trong từ điển Việt-Việt [41]. Số token các tiếng của văn bản còn lại sau khi tách được (hoặc không có trong từ điển) được chuyển sang xác định dựa trên độ đo sự liên quan từ vựng thông qua Internet với trọng số NGD theo công thức:

$$NGD = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

Trọng số NGD được luận văn áp dụng vào thử nghiệm theo nghiên cứu của Alberto J.Evangelista [26]:

$$NGD\#(x,y) = \frac{NGD(x,y)}{0{,}7}$$

Kết quả thử nghiệm phương pháp trên:

Bảng 2.3. Kết quả áp dụng độ đo NGD khi tách từ tiếng Việt

Từ/cụm từ	x	y	NGD#	Kết quả
nhà hàng hải sản	"nhà hàng"	"hải sản"	0,673	Chấp nhận
hợp tác xã	"hợp"	"tác xã"	0,775	Chấp nhận
biệt động sài gòn	"biệt động"	"sài gòn"	0.670	Chấp nhận
biệt động Hà Nội	"biệt động"	"Hà Nội"	1.323	~~Chấp nhận~~
chiến hạm tàng hình	"chiến hạm"	"tàng hình"	0.523	Chấp nhận
điện thoại di động	"điện thoại"	"di động"	0.393	Chấp nhận
điện thoại di chuyển	"điện thoại"	"di chuyển"	1.233	~~Chấp nhận~~
điện toán di động	"điện toán"	"di động"	0.995	~~Chấp nhận~~

Giải thuật cài đặt tách từ tiếng Việt của luận văn

Sự kết hợp tách từ thông qua từ điển và thống kê từ Internet thật sự mang lại hiệu quả về tốc độ xử lý và khả năng phát hiện những từ/cụm từ tiếng Việt không có trong từ điển. Phương pháp này có thể tự làm phong phú thêm danh sách từ tiếng Việt và giảm thiểu sự phụ thuộc vào Internet sau một thời gian thực thi.

2.3.3.2. Xác định từ khóa

Phương pháp đề xuất xác định từ khóa của luận văn dựa trên độ đo sự tần suất xuất hiện của từ trên các câu, độ đo tần số từ TF (Tearm Frequency) và độ đo nghịch đảo tần số tài liệu IDF (Inverse Document Frequency) như sau:

- Gọi cf_{ij} là số lượng câu có chứa từ khóa t_i trong tập k_j câu của tài liệu d_j đang xét, thì giá trị tần số từ khóa t_i xuất hiện trong tài liệu được tính:

$$freq\,(cf_{ij}) = \frac{cf_{ij}}{k_j}$$

- Gọi tf_{ij} là số lần xuất hiện của từ khóa t_i, độ đo TF được tính:

$$freq\,(tf_{ij}) = 1 + \log\left(tf_{ij}\right)$$

- Gọi df_i là số lượng tài liệu có chứa từ khóa t_i trong tập m tài liệu đang xét, độ đo IDF được tính:

$$idf_{ij} = \log\left(\frac{m}{df_i}\right) = \log(m) - \log(df_i)$$

Luận văn tính trọng số từ khóa t_i qua độ đo w_{ij}:

$$w_{ij} = freq(cf_{ij}) \times freq(tf_{ij}) \times idf_{ij}$$

Giải thuật xác định từ khóa của luận văn

Cài đặt giải thuật tính độ đo w_{ij} và tiến hành thử nghiệm tách từ tiếng Việt tại một trang tin Báo Tuổi Trẻ Online. Kết quả thu được:

Bảng 2.5. Các độ đo từ khóa được chọn theo phương pháp đề xuất

Từ tách được	Số phổ biến	TF×IDF	W_{ij} đề xuất
sinh viên	11	3.04445	0.15815
cà phê	**13**	2.51629	0.14161
đá bóng	4	***2.38925***	0.04137
thông tin	6	1.2682	0.03294
tấp nập	3	***1.75826***	0.02283
tổ chức	4	1.14261	0.01979
hoạt động	5	0.91255	0.01975
tài khoản	3	***2.20292***	0.01907

Kết quả thử nghiệm:

Các từ khóa có độ đo TFxIDF cao chưa phải là được chọn là từ khóa. Kết quả tính theo W_{ij} đề xuất mang lại rất khả quan và hợp lý.

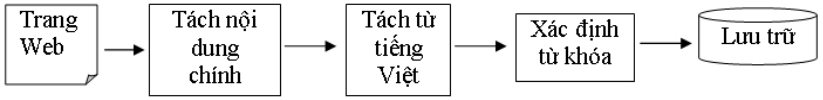
2.4. Kết chương

Chương 2 luận văn đã trình bày tổng quan về khai phá dữ liệu web, một ngành mới mở ra nhiều hướng nghiên cứu phục vụ khai phá text thông qua Internet.

Trong chương 2, luận văn đã lập trình kiểm thử đề xuất phương pháp xác định nội dung trang web thông qua kỹ thuật sử dụng độ sâu cây DOM của trang web kết hợp độ đo mật độ liên kết trong các Node cho kết quả bóc tách tốt.

Nội dung được bóc tách được chuyển sang tách từ tiếng Việt. Luận văn đã nghiên cứu kết hợp tách từ sử dụng từ điển có sẵn kết hợp với xử lý tách từ nhờ thông kê qua Internet, cụ thể là xác định độ đo NGD nhằm tìm ra những từ tiếng Việt chưa có trong từ điển.

Để xác định từ khóa tiếng Việt theo danh sách từ tách được, luận văn đã tiến hành thử nghiệm và đưa ra độ đo trọng số từ dựa trên 3 độ đo chính: độ đo mật độ câu có chứa từ trong tài liệu, độ đo tần số từ và độ đo nghịch đảo tần số. Những từ có w_{ij} cao nhất là những từ khóa tài liệu. Quá trình nghiên cứu đặt thử nghiệm được thực hiện chương hai theo sơ đồ sau:



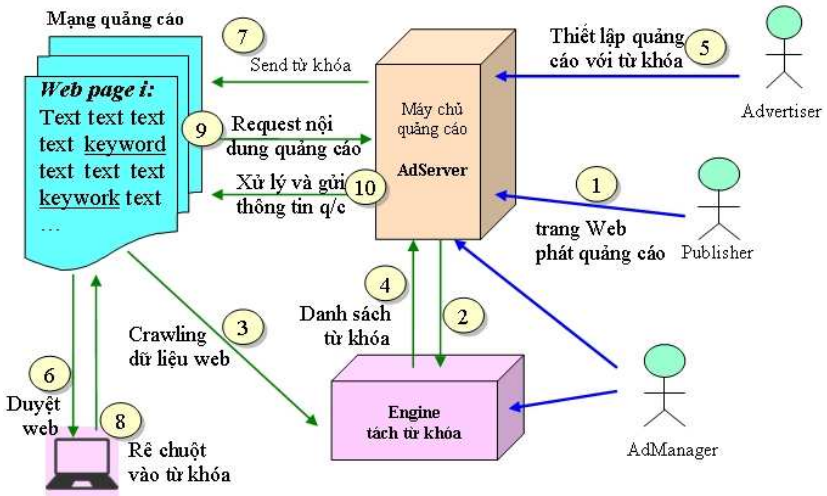
Ở chương tiếp theo, luận văn trình bày xây dựng hệ thống quảng cáo trực tuyến, sử dụng các từ khóa được lưu trữ làm cơ sở để chọn từ cũng như phát quảng cáo trên từ khóa này.

CHƯƠNG 3 - XÂY DỰNG HỆ THỐNG QUẢNG CÁO TRỰC TUYẾN

3.1. Tổng quan hệ thống

3.1.1. Các thành phần

3.1.2. Mô hình nghiệp vụ hệ thống xây dựng



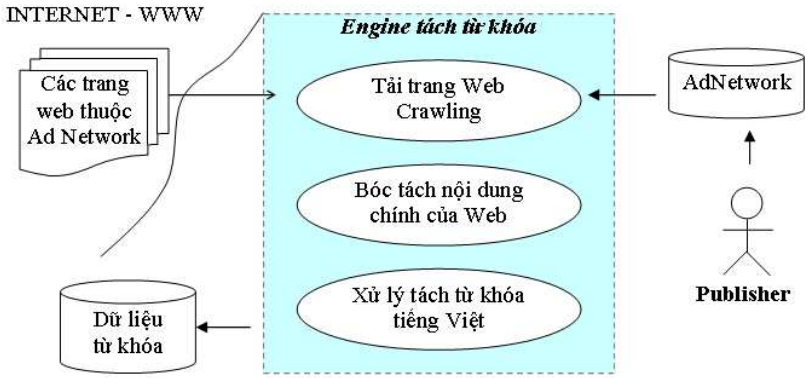
Hình 3.2. Mô hình hoạt động của hệ thống quảng cáo đề xuất

3.2. Phân tích và thiết kế

3.2.1. Thành phần mạng quảng cáo (Ad Network)

3.2.2. Thành phần Engine tách từ khóa

Engine tách từ khóa cung cấp danh sách những từ khóa tương ứng với trang web mà nó xử chuyển được nhập vào cơ sở dữ liệu máy chủ quảng cáo trực tuyến.



Hình 3.4. Mô hình chức năng của Engine tách từ khóa

3.2.2.1. Mô-đun tách nội dung chính của trang web

Mô-đun tách nội dung chính của trang web được thực hiện dựa trên phương pháp đề xuất của luận văn ở phần 2.3.3, chương 2.

Biểu đồ hoạt động tách nội dung chính của trang web

3.2.2.2. Mô-đun tách từ khóa tiếng Việt

Mô-đun tách từ khóa tiếng Việt bao gồm hai thành phần chính: tách từ tiếng Việt và tính toán lựa chọn từ khóa của nội dung cần tách.

Biểu đồ hoạt động mô-đun tách từ khóa tiếng Việt

3.2.3. Tác nhân tham gia hệ thống

3.2.3.1. Chức năng của Advertiser

Biểu đồ ca sử dụng của Advertiser

Biểu đồ hoạt động mô-đun đăng mẩu quảng cáo

3.2.3.2. Chức năng của Publisher

Biểu đồ ca sử dụng của Publisher

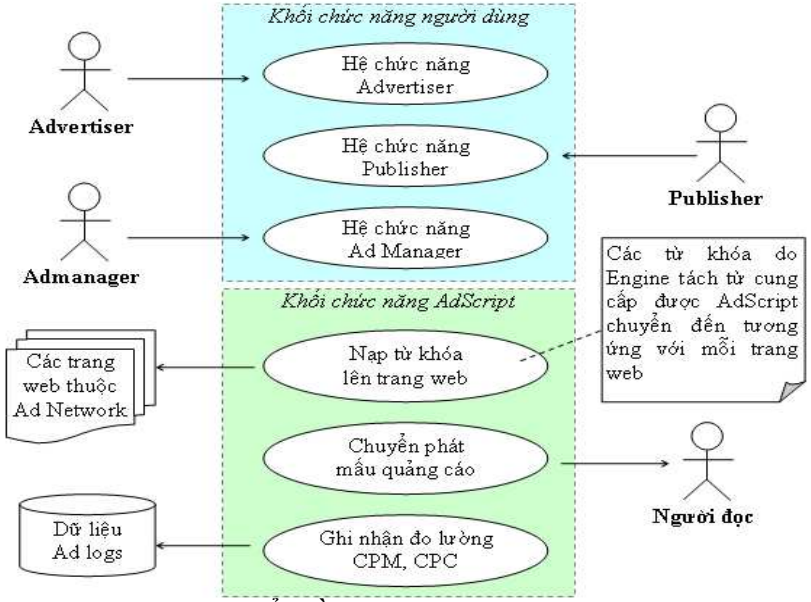
3.2.3.3. Chức năng của AdManager

Biểu đồ ca sử dụng của Ad Manager

Biểu đồ ca sử dụng Ad Manager

3.2.4. Portal AdServer

Portal AdServer là website bao gồm hai thành phần chính: thành phần giao diện tiện ích người dùng và thành phần chuyển phát quảng cáo AdScript.



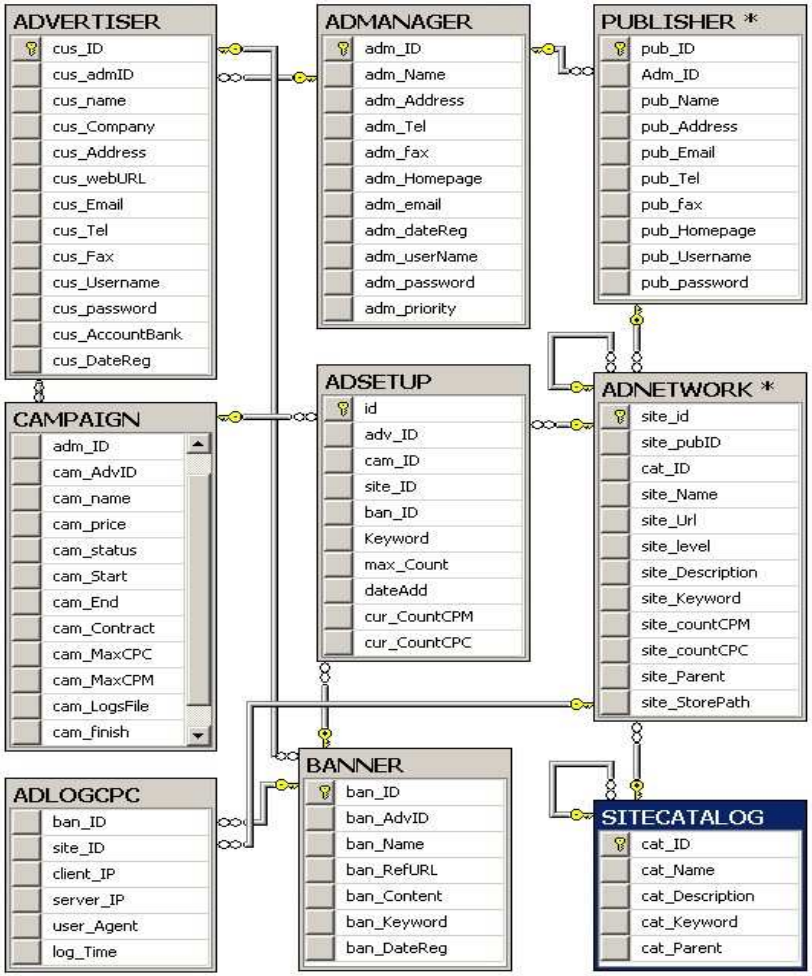
Hình 3.12. Biểu đồ ca sử dụng Portal AdServer

Sơ đồ hoạt động chức năng nạp từ khóa lên trang web

Sơ đồ hoạt động chuyển phát mẩu quảng cáo

3.3. Xây dựng và triển khai

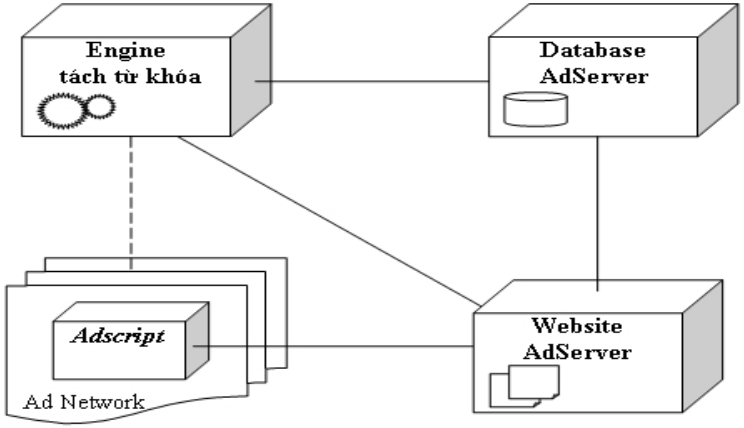
3.3.1. Thiết kế cơ sở dữ liệu



Hình 3.15. Biểu đồ quan hệ thực thể hệ thống quảng cáo trực tuyến

3.3.2. Công cụ và môi trường lập trình

3.3.3. Sơ đồ triển khai hệ thống



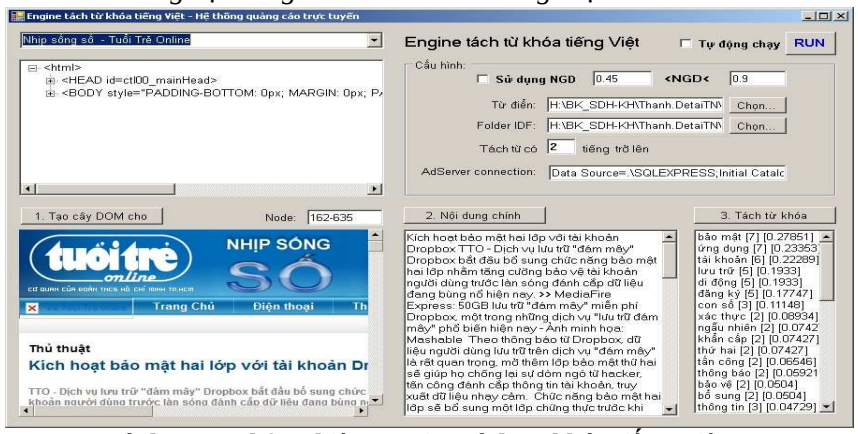
Hình 3.16. Sơ đồ triển khai hệ thống quảng cáo trực tuyến

3.4. Thử nghiệm và đánh giá kết quả

3.4.1. Thử nghiệm

Môi trường và dữ liệu thử nghiệm ứng dụng

3.4.1.1. Thử nghiệm Engine tách từ khóa tiếng Việt



Hình 3.17. Thử nghiệm Engine tách từ khóa tiếng Việt

3.4.1.2. Thử nghiệm triển khai quảng cáo trên Portal AdServer

Cổng truy nhập hệ thống

Ad Manager quản lý các Publisher

Publisher thiết lập mạng quảng cáo

Publisher cài đặt Ad Script quảng cáo vào website

Advertiser quản lý đợt quảng cáo

Advertiser thiết lập phát quảng cáo lên mạng quảng cáo

3.4.1.3. Thử nghiệm phát quảng cáo trực tuyến qua Ad Script

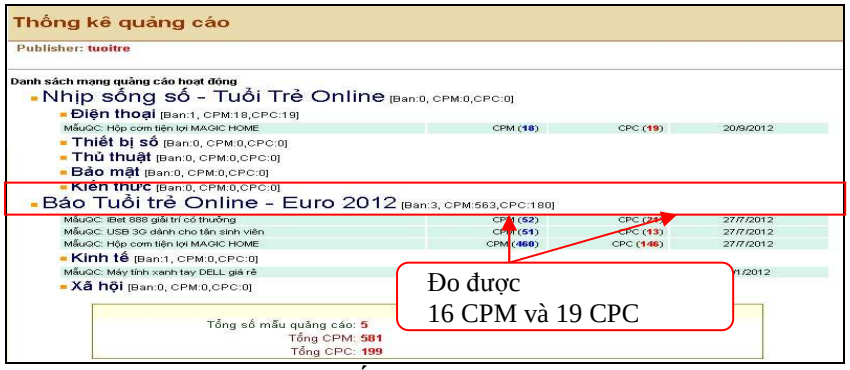
Các từ khóa được thiết lập quảng cáo được phát chính xác vào phần văn bản (text) nội dung chính của trang web. Khi di chuột qua từ khóa có đánh dấu, mẩu quảng cáo sẽ xuất hiện.

Hình 3.25. Quảng cáo từ khóa tiếng Việt xuất hiện trên báo điện tử

Thử nghiệm quảng cáo trên các thiết bị có màn hình truy cập hạn chế như Tablet PC, SmartPhone với các từ khóa "đông đảo", "thông minh".

3.4.1.4. Đo lường hiệu quả quảng cáo đã thực hiện

Công cụ đo lường hiệu quả quảng cáo trực tuyến Ad Manager, Publisher và Advertiser.



Hình 3.27. Thử nghiệm thống kê đo lường hiệu quả quảng cáo

3.4.2. Phân tích số liệu thống kê thử nghiệm hệ thống

Bảng 3.1. Kết quả thử nghiệm hệ thống

STT	Nội dung	Kết quả
1	Thời gian xử lý tách nội dung chính trang web	0.2 giây /1 trang
2	Thời gian tách từ khóa tiếng Việt với từ điển tiếng Việt 30.000 từ	6 giây / 1 trang
3	Thời gian xử lý tách từ khóa tiếng Việt sử dụng phương pháp kết hợp từ điển và thống kê qua Interrnet với độ đo NGD	58 giây /1 trang
5	Khả năng mở rộng dịch vụ cung cấp quảng cáo trực tuyến đa người dùng (nhiều Ad Manager, Advertiser, Publisher)	Không hạn chế
6	Khả năng mở rộng mạng quảng cáo và kho dữ liệu trang web của mạng quảng cáo	Tùy thuộc vào khả năng lưu trữ
7	Số lượng mẫu quảng cáo Advertiser có thể tạo	Không hạn chế
8	Khả năng mô tả nội dung mẫu quảng cáo trên Portal AdServer	Còn hạn chế
9	Tốc độ chuyển phát trung bình từ khóa quảng cáo với số lượng từ khóa tiếng trung bình 5 từ khóa	0,9 giây / toàn trang web
10	Thời gian trung bình phản hồi và ghi các độ đo hiệu quả quảng cáo	1,7 giây/mỗi lần nhấp chuột ở từ khóa
11	Ảnh hưởng tốc độ, mã nguồn trình bày trang web của mạng quảng cáo	Không ảnh hưởng

3.4.3. Đánh giá kết quả

- Kết quả thử nghiệm phát quảng cáo và hiển thị quảng cáo đúng vào nội dung văn bản chính trên trang web ở các trình duyệt web trên máy tính và thiết bị cầm tay: điện thoại smartphone, máy tính bảng, Internet TV.
- Phân hệ Engine tách từ khóa tiếng Việt tách chính xác phần nội dung chính và từ khóa cho hệ thống quảng cáo trực tuyến.
- Xây dựng cổng thông tin quản lý nghiệp vụ quảng cáo trực tuyến Portal AdServer trực quan và thuận lợi như việc thiết lập mạng quảng cáo, đăng quảng cáo và thống kê.
- Hệ thống xây dựng là sự kết hợp quy trình xử lý thông tin nhuần nhuyễn từ mạng quảng cáo, Engine tách từ khóa tiếng Việt, quản lý và thực hiện chuyển phát, đo lường quảng cáo.
- Có tiềm năng phát triển trong tương lai cũng như mở rộng áp dụng sang một số lĩnh vực liên quan đến dịch vụ từ khóa trực tuyến.

3.5. Kết chương

Trong chương này, luận văn tiến hành phân tích và thiết kế một số chức năng chính của hệ thống quảng cáo trực tuyến với từ khóa tiếng Việt. Phân tích các ca sử dụng, các biểu đồ mô tả hoạt động từ đăng mẩu quảng cáo đến nạp từ khóa lên các trang web, phát mẩu quảng cáo đến người đọc. Cuối chương là lập trình, xây dựng và triển khai hệ thống với phần thử nghiệm và đánh giá kết quả thực hiện.

KẾT LUẬN

1. Kết quả đạt được

Đề tài luận văn đã đạt được những yêu cầu đã đặt ra về mặt lý thuyết cũng như ứng dụng trong thực tiễn.

Về mặt lý thuyết, đề tài đã nghiên cứu và thử nghiệm về lĩnh vực khai phá nội dung web. Thực hiện xử lý ngôn ngữ, tách từ và xác định từ khóa tiếng Việt. Đề tài đã đề xuất các phương pháp mới dựa trên những nghiên cứu trước đây nhằm vận dụng giải quyết bài toán đặt ra.

Về mặt thực tiễn, đề tài đã xây hệ thống quản lý quảng cáo trực tuyến với từ khóa tiếng Việt, tạo ra một sản phẩm cung cấp dịch vụ quảng cáo trên Internet với kỹ thuật mới, đáp ứng xu thế phát triển của thị trường quảng cáo trực tuyến ở Việt nam còn nhiều tiềm năng.

2. Hạn chế

Độ chính xác tách từ tiếng Việt ở phân hệ Engine tách từ khóa vẫn còn phụ thuộc vào sự phong phú của dữ liệu từ điển và tốc độ truyền tải trên Internet. Các Ad Script chuyển phát quảng cáo chưa hoạt động tốt với tất cả các trình duyệt web ở tất cả các thiết bị.

3. Hướng phát triển

Cần được cập nhật công nghệ khắc phục những hạn chế nêu trên.

Phát triển Engine có phân tích, tổng hợp các chủ đề thông tin theo cấu trúc website trên mạng quảng cáo giúp hệ thống phát nội dung quảng cáo tự động theo suy diễn, tăng hiệu quả quảng cáo.

Phát triển khả năng phân phối quảng cáo trên nội dung chính của trang web một cách hợp lý, phù hợp địa phương, thời gian, nhu cầu khai thác thông tin của người đọc.