BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGUYỄN HOÀNG MINH HUY

PHÁT TRIỂN HỆ THỐNG QUẢNG CÁO THÔNG MINH TRÊN MẠNG XÃ HỘI

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công nghệ thông tin

Mã số ngành: 60480201

TP. HỒ CHÍ MINH, tháng 10 năm 2015

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

NGUYỄN HOÀNG MINH HUY

PHÁT TRIỂN HỆ THỐNG QUẢNG CÁO THÔNG MINH TRÊN MẠNG XÃ HỘI

LUẬN VĂN THẠC SĨ

Chuyên ngành : Công nghệ thông tin

Mã số ngành: 60480201

CÁN BỘ HƯỚNG DẪN KHOA HỌC: PGS.TS QUẢN THÀNH THƠ

TP. HỒ CHÍ MINH, tháng 10 năm 2015

CÔNG TRÌNH ĐƯỢC HOÀN THÀNH TẠI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ TP. HCM

Cán bộ hướng dẫn khoa học : PGS.TS Quản Thành Thơ

(Ghi rõ họ, tên, học hàm, học vị và chữ ký)

Luận văn Thạc sĩ được bảo vệ tại Trường Đại học Công nghệ TP. HCM ngày 17 tháng 10 năm 2015

Thành phần Hội đồng đánh giá Luận văn Thạc sĩ gồm:

(Ghi rõ họ, tên, học hàm, học vị của Hội đồng chấm bảo vệ Luận văn Thạc sĩ)

TT	Họ và tên	Chức danh Hội đồng
1	PGS.TSKH. Nguyễn Xuân Huy	Chủ tịch
2	TS. Lư Nhật Vinh	Phản biện 1
3	TS. Nguyễn Thị Thúy Loan	Phản biện 2
4	TS. Trần Đức Khánh	Ủy viên
5	TS. Võ Đình Bảy	Ủy viên, Thư ký

Xác nhận của Chủ tịch Hội đồng đánh giá Luận sau khi Luận văn đã được sửa chữa (nếu có).

Chủ tịch Hội đồng đánh giá LV

TRƯỜNG ĐH CÔNG NGHỆ TP. HCM
PHÒNG QLKH – ĐTSĐH

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc

TP. HCM, ngày 17 tháng 10 năm 2015

NHIỆM VỤ LUẬN VĂN THẠC SĨ

Họ tên học viên: NGUYỄN HOÀNG MINH HUY Giới tính: Nam

Ngày, tháng, năm sinh: 27/03/1985.....Nơi sinh: TP.HCM.....

Chuyên ngành: Công nghệ thông tin.....MSHV: 1341860041.....

I- Tên đề tài:

PHÁT TRIỂN HỆ THỐNG QUẢNG CÁO THÔNG MINH TRÊN MẠNG XÃ HỘI

II- Nhiệm vụ và nội dung:

Nghiên cứu giải thuật TF-IDF và ứng dụng xây dựng hệ thống quảng cáo trên mạng xã hội.

III- Ngày giao nhiệm vụ: 03/04/2015

IV- Ngày hoàn thành nhiệm vụ: 10/09/2015

V- Cán bộ hướng dẫn: PGS.TS. Quản Thành Thơ

CÁN BỘ HƯỚNG DẪN
(Họ tên và chữ ký)

KHOA QUẢN LÝ CHUYÊN NGÀNH
(Họ tên và chữ ký)

PGS.TS. Quản Thành Thơ

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả đánh giá, nhận xét và các đề xuất cải tiến mới nêu trong Luận văn là trung thực và chưa từng được ai công bố trong bất kỳ công trình nào khác.

Tôi xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện Luận văn này cũng như các trích dẫn hay tài liệu học thuật tham khảo đã được cảm ơn đến tác giả hay ghi rõ ràng nguồn gốc thông tin trích dẫn trong Luận văn.

.

Học viên thực hiện Luận văn

LỜI CÁM ƠN

Trước tiên, Em xin bày tỏ lòng biết ơn sâu sắc nhất đến thầy PGS.TS Quản Thành Thơ, người đã dìu dắt, định hướng cho em ngay từ bước đầu tiên em làm quen với con đường nghiên cứu khoa học. Thầy không những tận tình hướng dẫn, truyền đạt cho em những kiến thức chuyên môn quý báu, tạo mọi điều kiện thuận lợi cho em trong suốt quá trình làm luận văn.

Em xin gửi lời chân thành đến toàn thể quý thầy cô Khoa Công Nghệ Thông Tin, những người đã trang bị cho em kiến thức nền tảng, tận tình chỉ bảo, dìu dắt em trong suốt những năm học qua.

Để hoàn thành luận văn này, em không quên gửi lời cảm ơn chân thành đến bạn Nguyễn Công Đính cùng toàn thể các bạn trong lớp cao học 13SCT21 trường Đại học Công Nghệ TPHCM đã tạo mọi điều kiện giúp đỡ để em có thể hoàn thành luận văn này.

Với lòng biết ơn sâu sắc xin được gửi đến gia đình tôi, là nguồn động viên to lớn, là chỗ dựa vững chắc cho tôi, luôn bên tôi lúc tôi thành công cũng như tôi thất bại.

Lời cảm ơn sau cùng xin dành cho bạn bè tôi, những người đã gần gũi chia sẽ giúp đỡ và động viên tôi vượt qua khó khăn để hoàn thành luận văn này.

NGUYỄN HOÀNG MINH HUY

TÓM TẮT

Với sự phát triển ngày càng mạnh mẽ của mạng xã hội, đang dần dần thay đổi xu hướng của người tiêu dùng và cách truyền thông của các doanh nghiệp.Truyền thông truyền thống đang được dần thay thế bằng truyền thông trực tuyến (e-marketting) với trợ giúp của sự phát triển Internet. Một trong những hình thức truyền thông trực tuyến là tiếp thị trên mạng xã hội (social-media marketting).

Với sự phát triển nhanh chóng của mạng xã hội , thì tiếp thị trên mạng xã hội là hướng đi mới đầy tiềm năng cho lĩnh vực kinh tế nói chung và Công nghệ thông tin nói riêng. Làm thế nào để một doanh nghiệp có thể chọn mạng xã hội nào để quảng cáo hiệu quả thương hiệu của mình với chi phí thấp? Đó là câu hỏi đặt ra không những cho lĩnh vực Marketting mà còn cho lĩnh vực Công nghệ thông tin.

Đề tài "Phát triển hệ thống quảng cáo thông minh trên các mạng xã hội" sẽ hiện thực một mô hình tiếp thị trên mạng xã hội.

Kết quả thực nghiệm cho thấy đã bước đầu nghiên cứu xây dựng thành công hệ thống phân tích dữ liệu (thường là các mẫu quảng cáo) sau đó khi hệ thống phân tích xong, sẽ đề xuất một số pages và groups cho người dùng có thể đăng quảng cáo và tìm hiểu thông tin mình cần, ngoài ra hệ thống còn cho người dùng biết ước lượng phần trăm tỉ lệ chính xác hệ thống phân tích được.

.

ABSTRACT

Together with rapid growth of social networking sites, people are changing their ways of buying things, and companies are changing their ways of marketing too. Traditional marketing has been replaced by e-marketing thanks to the Internet. And one of the online marketing types is social-media marketing.

The dramatic development of social networking sites has opened an untapped potential for economy and internet technology. How a company can choose an effective and cost-saving social networking site? It is an issue not only for marketing but also for internet technology.

Project of "Smart advertisement system on social networking sites" will create a new marketing model on social networking sites.

Practical results show that the project can research and build a data analysis system (of advertisements). After analyzing, the tool will suggest some pages or group-pages to the user, and then they can post their advertisement and search for their desired information. In addition, the tool also gives user an estimated percentage rate of accuracy.

MỤC LỤC

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH

CHƯƠNG 1
MỞ ĐẦU

1.1/ Giới thiệu đề tài

Cùng với quá trình toàn cầu hóa và sự phát triển của công nghệ thông tin, mạng internet trên thế giới và Việt Nam ngày càng phát triển mạnh mẽ. Bên cạnh đó, thói quen trong mua sắm và tìm kiếm thông tin của người tiêu dùng cũng thay đổi nhiều theo xu hướng đó. Họ hoàn toàn có thể thông qua Internet để thực hiện những điều đó nhằm tiết kiệm được tối đa thời gian và chi phí.

Diễn biến ngoạn mục này đã mở ra kỷ nguyên mới trong truyền thông hiện đại và hiệu quả cho các doanh nghiệp đó chính là truyền thông kĩ thuật số với sự trợ giúp của sự phát triển Internet.

Một trong những hình thức truyền thông kĩ thuật số là tiếp thị liên kết trên mạng xã hội và quảng cáo thông minh trên mạng xã hội:

Tiếp thị liên kết (affiliate marketing) :là phương thức tiếp thị dựa trên nền tảng Internet trong đó một website sẽ quảng bá sản phẩm hoặc dịch vụ cho nhiều website khác mà được hưởng hoa hồng từ phương thức quảng bá này thông qua lượng truy cập, doanh số bán hàng hoặc khi mẫu đăng ký được hoàn tất… Tiếp thị liên kết khác với phương thức quảng cáo truyền thống nhờ việc thanh toán chỉ dựa trên hiệu quả của quảng cáo mà không phụ thuộc vào thời gian và tần suất quảng cáo.

Tiếp thị liên kết khác với phương thức quảng cáo truyền thống nhờ việc thanh toán chỉ dựa trên hiệu quả của quảng cáo (khi có đơn hàng hoặc có hành động của khách hàng như hoàn thành mẫu đăng kí, tải, trả lời khảo sát...) mà không phụ thuộc vào thời gian và tần suất quảng cáo.

Quảng cáo thông minh trên mạng xã hội: là hình thức cao cấp hơn của tiếp thị liên kết, là sự xây dựng hệ thống phân tích một số mẫu quảng cáo của người dùng thuộc nhiều loại khác nhau, mỗi quảng cáo này sau khi được hệ thống phân tích xong sẽ đưa ra một số kết quả đề nghị người dùng nên đăng quảng cáo này trên những

papes hoặc group (thuộc các trang mạng xã hội) nào và ngoài ra hệ thống còn cho người dùng biết ước lượng phần trăm tỉ lệ chính xác hệ thống phân tích được.

So với quảng cáo truyền thông bình thường thì quảng cáo thông minh trên mạng xã hội có những lợi thế như :

- Tiết kiệm tối đa khoản chi phí dành cho quảng cáo nhưng quảng cáo lại đạt hiệu quả cao vì hệ thống phân tích được mẫu quảng cáo nên đăng trên những papes và group nào.
- Ước lượng % tỉ lệ chính xác của độ tương thích mẫu quảng cáo mà hệ thống phân tích được
- Thừa hưởng được những ưu điểm của mạng xã hội:

➢ Tính lan truyền: Mạng xã hội là công cụ lan truyền nhanh chóng và hiệu quả nhất hiện nay, nó tăng theo cấp số nhân, chỉ bằng một cái click chuột "like" hay "fan page", thì những thông tin sẽ lan truyền đến bạn bè, hội nhóm trên mạng xã hội.

➢ Tính thân thiện: Mạng xã hội có những giao diện hết sức thân thiện với người dùng, thích hợp với tầng lớp tri thức, học sinh, sinh viên, giáo viên… Hơn nữa người quản lý có thể dễ dàng chỉnh sửa các thông tin hay cập nhật những thông tin mới dưới dạng những ghi chú (Notes) hoặc sự kiện (Events) với những thao tác khá đơn giản, hay có thể đăng tải để chia sẻ bất kỳ hình ảnh đẹp về hoạt động công tác thông tin thư viện lên mạng xã hội một cách nhanh chóng, hiệu quả tạo ra hứng thú đối với bạn đọc ngay tức thì.

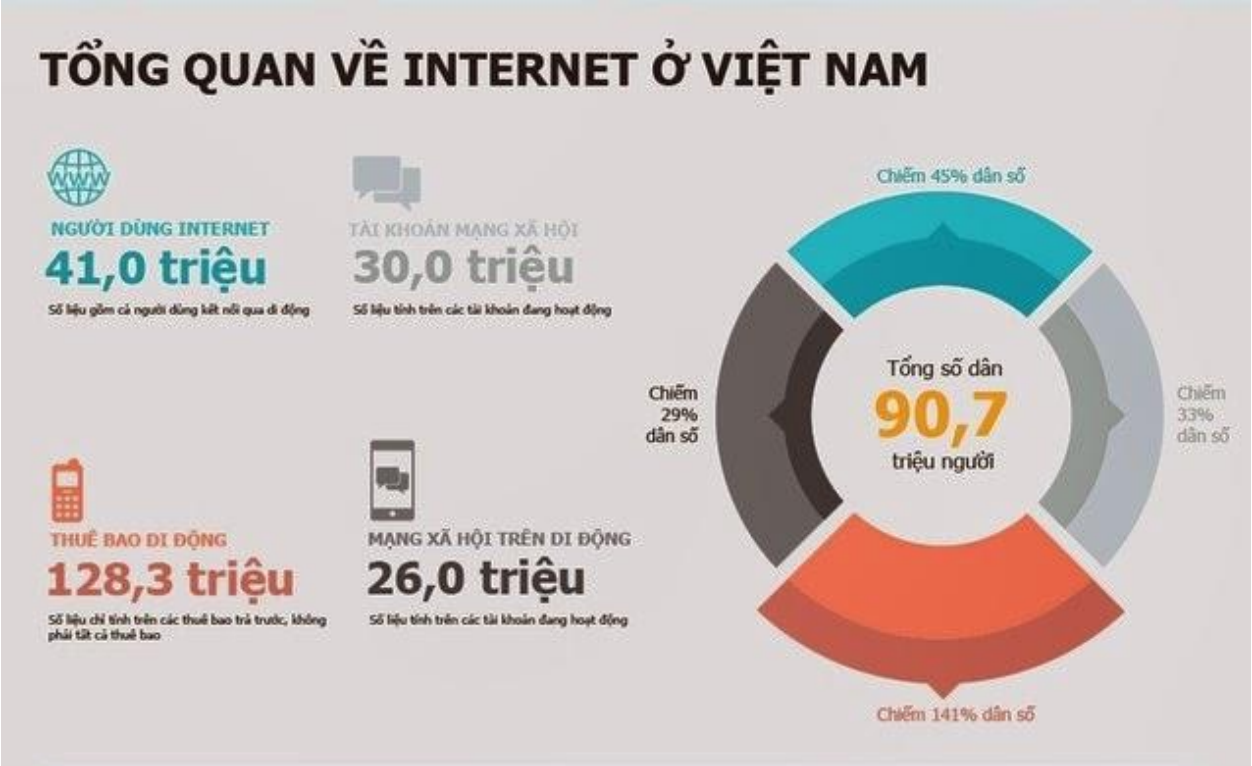
➢ Tính liên kết: Cung cấp các đường dẫn (link) đến các trang khác của các thư viện khác hay các trang cơ sở dữ liệu mà bạn đọc đang tìm kiếm.

1.2/ Tính cấp thiết của đề tài

Trước bối cảnh người dùng internet với mạng xã hội trên toàn cầu và Việt Nam phát triển một cách quá nhanh chóng [7,21]:

Hình 1.1: Bối cảnh dữ liệu toàn cầu - Nguồn:We Are Social

Tại Việt Nam, theo số liệu thống kê mới nhất (tháng 03/2015) [21] có hơn 45% dân số (41 triệu/90,7 triệu) sử dụng Internet và khoảng 33% dân số (30 triệu/90,7 triệu) sử dụng mạng xã hội. Do đó mạng xã hội là một thị trường rất tốt để kinh doanh nhiều loại sản phẩm và dịch vụ khác nhau.



Hình 1.2: Internet và mạng xã hội ở Việt Nam - Nguồn:We Are Social

Quảng cáo trên mạng xã hội sẽ tiếp cận được một khối lượng thị trường khổng lồ đặc biệt là nhóm đối tượng trẻ, thông qua tương tác và chia sẻ trên mạng xã hội sẽ giúp chi phí quảng cáo được tối ưu hóa một cách đáng kể.

Với sự phát triển vượt trội của Internet nói chung và mạng xã hội nói riêng , quảng cáo thông minh trên mạng xã hội là một lĩnh vực đầy tiềm năng nhưng vẫn rất mới ở Việt Nam. Để khai thác tiềm năng đó, chúng tôi thực hiện đề tài "PHÁT TRIỂN HỆ THỐNG QUẢNG CÁO THÔNG MINH TRÊN MẠNG XÃ HỘI".

1.3/ Mục tiêu của đề tài

Xây dựng thành công hệ thống cơ sở dữ liệu phục vụ cho việc phân tích mẫu quảng cáo.

Xây dựng thành công hệ thống phân tích dữ liệu (thường là mẫu quảng cáo).

Xây dựng thành công hệ thống quảng cáo thông minh trên mạng xã hội (Information retrieval social media).

Ngoài ra, để hệ thống chạy nhanh hơn và hiệu quả hơn, sẽ có sự phân tích xu hướng của người sử dụng trên mạng xã hội, đưa ra gợi ý về những sản phẩm, dịch vụ phù hợp, gần gũi và cập nhật theo thời gian cho người sử dụng.

Ứng dụng thực tiễn: Bước đầu xây dựng mô hình hệ thống quảng cáo thông minh và thử nghiệm trên thực tế để giúp cho các doang nghiệp nói riêng và nền kinh tế nói chung có trong tay một công cụ quảng cáo hiệu quả cao với chi phí thấp, từ đó mang lại lợi ích kinh tế vượt bậc.

CHƯƠNG 2

TỔNG QUAN VỀ MẠNG XÃ HỘI VÀ TỔNG QUAN VỀ PHÁT TRIỂN HỆ THỐNG QUẢNG CÁO THÔNG MINH TRÊN MẠNG XÃ HỘI

2.1/ Mạng xã hội (Social Network)

2.1.1/ Định nghĩa

Mạng xã hội là dịch vụ online, được phát triển trên nền tảng web, cung cấp cho user tương tác trên mạng Internet, tạo ra môi trường liên kết, tương tác như chia sẻ sở thích, ý kiến, hoạt động, giới thiệu bản thân, hoặc những hoạt động khác. Mạng xã hội bao gồm nhiều user (có thể là một cá nhân, một nhóm hay một tổ chức), các mối quan hệ, liên kết của user và các sản phẩm dịch vụ khác.

Mạng xã hội ra đời đã đáp ứng được như cầu đa dạng của con người dễ dàng, nhanh chóng như được tương tác, chia sẻ, trao đổi, kết nối với cộng đồng, cập nhật thông tin …Mạng xã hội là dịch vụ mới ra đời nhưng đã thu hút được số lượng lớn người sử dụng trong một thời gian ngắn so với các phương tiện truyền thông khác.

2.1.2/ Phân loại mạng xã hội

Không phải tự nhiên thời của các mạng xã hội lại lên ngôi. Đây có thể được xem là xu thế tất yếu, khi khoa học kỹ thuật hiện đại phát triển đến một mức độ nhất định, và khi con người có nhu cầu tăng cường sợi dây liên kết với thế giới mà không bị mất quá nhiều thời gian như trước kia.

Với sự ra đời của các mạng xã hội, trong các hình thức giao tiếp của con người mặc nhiên nảy sinh một dạng mới, đó là liên kết trên thế giới ảo. Nói rằng ảo, bởi lẽ chúng ta bắt được liên lạc với nhau qua những sợi dây vô hình khi cùng dùng chung một mạng xã hội nào đó. Tuy nhiên, điều ảo này lại liên kết trực tiếp cuộc sống thực. Trên thực tế, nó đã giúp chúng ta nói với nhau những điều mà khi đối diện có thể khó mở lời, nó lại giúp nhiều người chia sẻ được với nhau những khoảnh khắc yêu thương

mà không một cách truyền thống nào trước kia làm được, hơn thế, nó khiến chúng ta cảm thấy gần nhau hơn bao giờ hết, dù khoảng cách địa lý có là bao xa.

Với tất cả những ưu điểm đó, mạng xã hội xứng đáng trở thành một phần khó thiếu của cuộc sống hiện đại. Và vì nó đã hữu dụng đến vậy, bạn đừng để mình đứng ngoài cuộc chơi. Hãy thử đếm xem mình đã có bao nhiêu tài khoản trên các mạng xã hội, những mạng nào bạn dùng thường xuyên nhất để đo độ cập nhật với thế giới số đang thay đổi chóng mặt từng giờ.

2.1.2.1/ Facebook

Tất nhiên, nhắc đến mạng xã hội, địa chỉ đầu tiên mà mọi người nhớ tới là Facebook. Chỉ tính đến tháng 10/2012, Facebook đã có 1 tỉ người sử dụng, như vậy nếu tính là một quốc gia thì Facebook chẳng hề thua kém về dân số so với Trung Quốc hay Ấn Độ. Sức hấp dẫn của Facebook lớn đến mức giờ đây người ta có thể sử dụng mạng xã hội này để làm ăn và thậm chí kiếm được rất nhiều tiền.

Sở hữu một tài khoản trên Facebook, bạn sẽ nhanh chóng có được một cộng đồng chung thông qua các mối liên kết bạn bè hoặc các fanpage cùng một mối quan tâm. Từ đây, bạn sẽ cũng sẽ trở thành một trung tâm được xoay quanh bởi các vì sao là bạn bè, người thân và họ sẽ bày tỏ sự quan tâm của mình bằng cách để lại những lần nhấn like hoặc comment trên mỗi tấm hình, mỗi câu chia sẻ tâm trạng của bạn. Cũng nhờ thế, facebook đã trở thành một phần của đời sống hiện đại của chúng ta, bởi nó không chỉ giúp xóa tan khoảng cách, xây dựng một cộng đồng có mối liên kết mật thiết mà nó còn giúp ta thỏa mãn cái tôi cá nhân và bớt cảm thấy cô đơn hơn khi thường xuyên nhận được sự ủng hộ sau mỗi lần sẻ chia.

2.1.2.2/ Youtube

Những ai yêu thích các clip trực tuyến sẽ không bao giờ bỏ lỡ cơ hội trở thành cư dân Youtube. Tính đến tháng 5/2014, Youtube đã đạt mốc 2 tỉ lượt xem trên một video (MV Gangnam Style), và con số này chắc chắn không dừng ở đó. Với một tài khoản trên Youtube, bạn sẽ có một kênh video của riêng mình để có thể đăng những

đoạn clip yêu thích, cùng chia sẻ chúng với bạn bè, người thân, thậm chí những người lạ. Ngoài ra, bạn có thể sưu tập những clip mình muốn lưu giữ, sắp xếp chúng theo thứ tự mà bạn mong muốn, theo dõi các kênh mình quan tâm, bình luận trên các video yêu thích từ đó có thể kết thêm bạn bè…

Có quá nhiều lợi ích từ Youtube tới mức nhiều người sử dụng đây không chỉ là một trang mạng xã hội mà còn là một công cụ tìm kiếm bất cứ thứ gì liên quan đến phim ảnh, ca nhạc, thời sự…. Và nếu có sở thích, bạn có thể đầu tư vào kênh youtube của riêng mình, biến nó trở thành một kênh thông tin hình ảnh hữu ích, từ đó người theo dõi bạn tăng cao, cùng với đó, những lợi nhuận khác cũng sẽ đồng thanh gõ cửa.

2.1.2.3/ Instagram

Cũng là một mạng xã hội liên kết bạn bè, tuy nhiên khi sử dụng Instagram, bạn sẽ có cảm giác mọi thứ diễn ra nhanh chóng, như những lát cắt gọn gàng và chính xác từ cuộc sống xung quanh. Không nặng về các chia sẻ liên quan đến câu chữ, Instagram chủ yếu mang đến cho bạn những tấm hình chớp nhoáng, nhưng chân thật và nhanh nhạy nhất.

Ngoài ra, vì chú trọng về hình ảnh, Instagram rất biết chiều chuộng người dùng bằng một ứng dụng chỉnh ảnh với rất nhiều lựa chọn hoàn hảo. Có rất nhiều người dùng đã đem lòng yêu Instagram bởi chế độ chỉnh ảnh quá tuyệt vời này, để rồi sau đó họ sử dụng kết quả đó để đăng lên các mạng xã hội khác như Facebook, Twitter, Tumblr hay Flickr. Bởi vậy, Instagram rất được ưa chuộng, không chỉ bởi những tính năng ưu việt của mình mà còn bởi nó là một mạng xã hội trung gian hữu ích.

2.1.2.4/ Tumblr

Không ồn ào như Facebook, cũng không có được các chế độ chỉnh ảnh trong mơ như Instagram, thế nhưng Tumblr vẫn được yêu thích vì thể hiện được tính cá

nhân rất cao. Người dùng trên Tumblr thực ra không đặt nặng các mối quan hệ bởi trên thực tế họ chỉ follow nhau và chế độ comment cũng không quá phát triển. Nhưng cũng chính bởi thế, khi đến với Tumblr, bạn cảm thấy thế giới như của riêng mình, bạn có thể thỏa sức thể hiện cái tôi mà ít có cảm giác bị đánh giá và phán xét.

Giao diện Tumblr được xây dựng theo một khung dọc và khá hẹp chứ không thông thoáng như Facebook, tuy nhiên không vì thế mà nó làm cho người dùng cảm thấy gò bó. Trái lại, bạn càng cảm thấy ấm cúng hơn với cảm giác góc nhỏ của riêng mình. Những chia sẻ trên Tumblr vì thế thường được đánh giá là có chiều sâu và mang màu sắc cá nhân rõ rệt, từ hình ảnh cho đến câu chữ.

2.1.2.5/ Twitter

Thực ra, trên thế giới, Twitter là một mạng xã hội rất được ưa chuộng, tuy nhiên nó chưa thực sự phổ biến tại Việt Nam. Điểm ưu việt của Twitter đó là ở độ nhanh nhạy và khả năng lan truyền rộng rãi. Những đoạn cập nhật tình hình thời sự, nhất là trong các thời điểm nhạy cảm, trên Twitter được đánh giá là nhanh nhạy hơn cả các loại báo chí chính thống.

Một điểm đặc biệt nữa của Twitter là nó giới hạn ký tự trong mỗi tweet, vì thế nó thể hiện rõ tính ngắn gọn, nhanh chóng và góp phần thúc đẩy các dịch vụ thu gọn địa chỉ website như bit.ly, tr.im, tinyurl. Ngoài ra, Twtter còn là một mạng nhắn tin thú vị, nhanh chóng bằng cách follow người khác và để người khác follow, cộng thêm cả khả năng tương tác với các Twitter khác trên thiết bị di động. Hãy yên tâm rằng dù đang ở bất cứ đâu bạn cũng có thể giữ liên lạc với một nhóm về các chủ đề hot hoặc thông báo để khách hàng có thể cập nhật những thông tin mới và kịp thời nhất.

2.1.2.6/ Flickr

Những nhiếp ảnh chuyên nghiệp chắc hẳn sẽ không thể không có một kho hình riêng ở Flickr, đơn giản vì sự phổ biến và tính năng thân thiên của nó. Các blogger đều dành cho Flickr mối quan tâm lớn, bởi ở đó họ vừa lưu giữ được những hình ảnh

đẹp, vừa có thể lựa chọn phương án giữ bản quyền hay không, lại vừa được chia sẻ niềm đam mê nhiếp ảnh với những ai quan tâm.

Sự tương tác trên Flickr được thể hiện qua các comment và like, ngoài ra bạn còn có thể sắp xếp hình ảnh vào các nhóm mục liên quan để nâng cao khả năng tìm kiếm. Nếu yêu các tấm hình và muốn chia sẻ tình yêu này thật rộng rãi, bạn hãy mở một tài khoản trên Flickr.

2.1.2.7/ Pinterest

Nếu là người làm thiết kế, hẳn bạn chẳng thể bỏ qua Pinterest, và thậm chí không làm ở lĩnh vực này, bạn cũng sẽ bị mê hoặc. Tất cả mọi ý tưởng từ nội thất, thời trang, đồ họa, nhiếp ảnh đều có thể được tìm thấy ở đây. Tuy nhiên, đây vốn là một mạng xã hội, vì thế nó sẽ giúp bạn tương tác với bạn bè bằng nút "pin" để kéo họ cùng tham gia vào những ý tưởng hay, những tấm hình đẹp. Ngược lại, bạn bè của bạn cũng có thể bình luận hoặc chia sẻ lại bức ảnh ấy để thể hiện sự quan tâm.

Hãy tham gia Pinterest để theo kịp các xu hướng mới nhất, chia sẻ những tấm hình đặc sắc và quan trọng hơn là thấy mình không bao giờ cạn nguồn ý tưởng. Ngôi nhà của các ý tưởng luôn chào đón bạn và những nhóm cùng sở thích, vì thế đừng bỏ lỡ cơ hội trải nghiệm từ hôm nay.

2.1.2.8/ LinkedIn

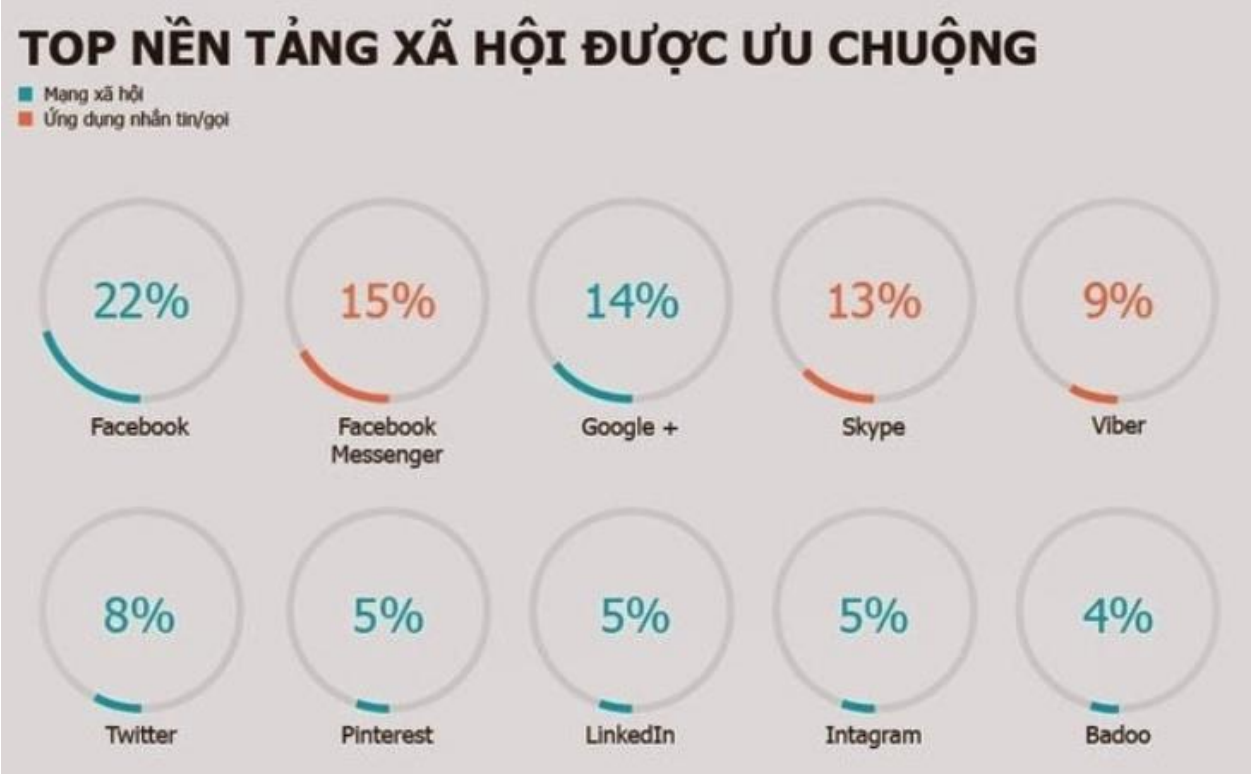
Nếu là một người đang trong độ tuổi đi làm, hẳn bạn sẽ quan tâm đến LinkedIn. Đây là một mạng xã hội với những người dùng chủ yếu là các thành viên chuyên nghiệp như doanh nghiệp hoặc cá nhân có nhu cầu kết nối tìm việc, tuyển dụng hoặc quảng cáo. Cũng giống như Facebook, khi tham gia LinkedIn, bạn sẽ phải xây dựng một hồ sơ cá nhân, nhưng vì nó nhắm đến việc tuyển dụng chuyên nghiệp nên hồ sơ này cần có sự chau chuốt hơn rất nhiều so với trên Facebook.

Bằng các kết nối bạn bè trên LinkedIn, mạng lưới của bạn sẽ tăng lên rất lớn, và càng lớn thì cơ hội công việc của bạn sẽ càng cao. Thêm một tính năng ưu việt nữa là nhờ LinkedIn, bạn có thể follow những công ty mình quan tâm để cập nhật tin tức, tìm hiểm về các nhà tuyển dụng tương lai để hiểu thêm cơ hội của mình hoặc tìm kiếm bất cứ thứ gì bạn quan tâm xem có mối liên hệ nào chung không để kiếm cách tiếp cận. Nói chung, với LinkedIn, cơ hội công việc của bạn sẽ trở nên phong phú và ngày càng hấp dẫn hơn nếu bạn biết tận dụng tối đa những tính năng ưu việt của mạng xã hội này.

2.1.2.9/ Lief

Nếu bạn muốn sử dụng mạng xã hội để tăng mối liên kết với người thân nhưng không muốn nó quá rộng và đại trà như Facebook, hãy nghĩ đến Lief. Đây là một mạng xã hội mới ra đời, nhằm mang đến cho bạn góc riêng tư, nơi bạn là chính mình, với ít bạn bè hơn nhưng những người được xếp vào Friend list sẽ thật sự là những người thân thiết nhất, như cha mẹ, anh chị, vợ chồng, họ hàng và những người bạn tâm giao…

Không cần phải sống cùng một xã hội thu nhỏ với các mối quan hệ chằng chịt như trên Facebook, không bị giới hạn 140 ký tự như Twitter, cũng không khó comment như Tumblr, Lief đủ thân thiện và đủ riêng tư để bạn thiết lập một mạng xã hội gia đình mật thiết để không bỏ lỡ bất cứ phút giây ý nghĩa nào của người thân. Hiện nay, bạn có thể bắt đầu tạo tài khoản, mời người thân tham gia trực tiếp trên website http://lief.com/lief để cùng trải nghiệm những giây phút ý nghĩa và tuyệt vời như thế.



Hình 2.1: Top nền tảng xã hội được ưu chuộng

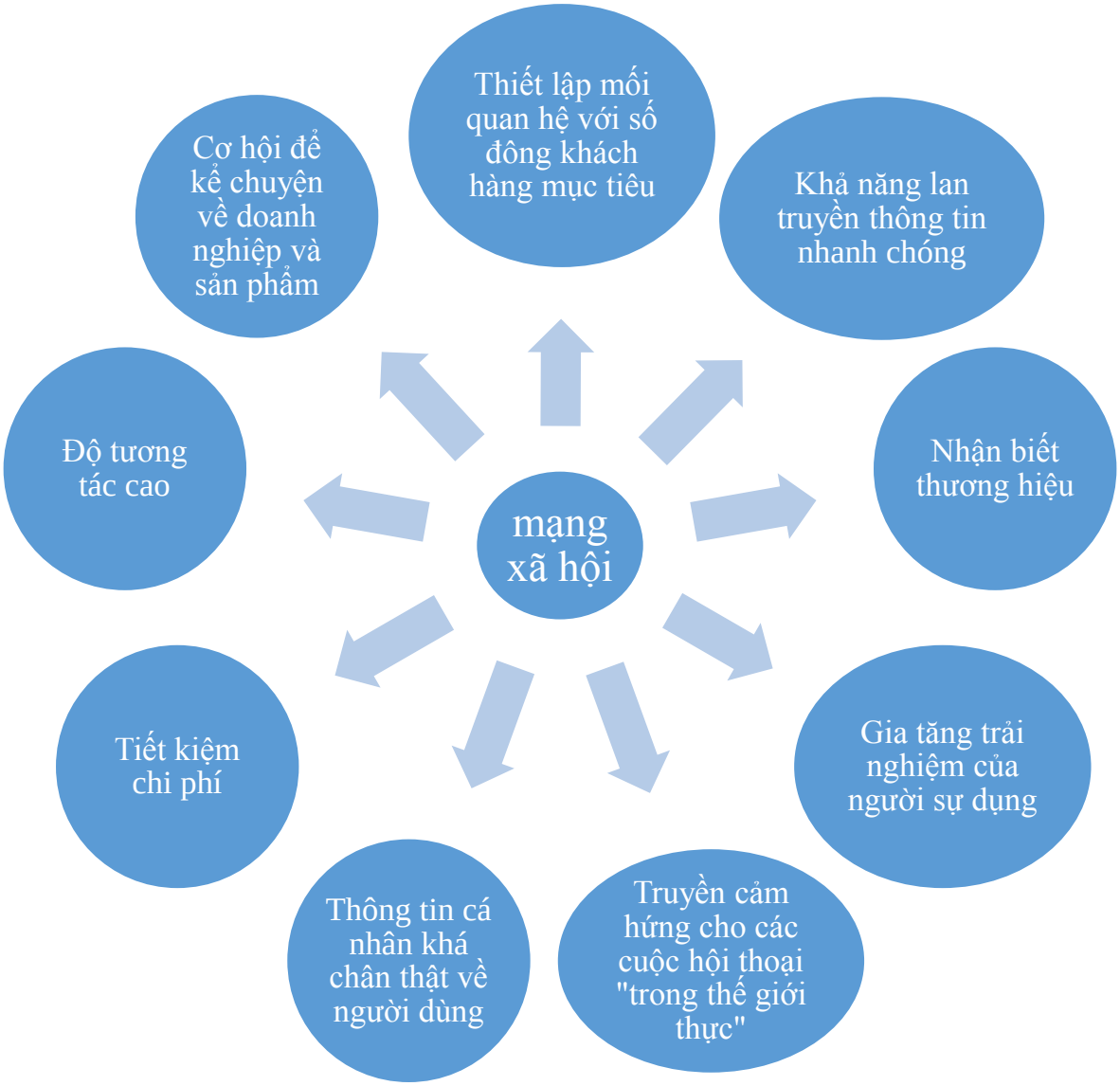
Ở Việt Nam, các mạng xã hội cũng phát triển như ZingMe, Yume, Tamtay…

2.2/ Quảng cáo trên mạng xã hội

2.2.1/ Định nghĩa

Với sự phát triển nhanh chóng và mạnh mẽ của mạng xã hội, đã tác động đến nhiều mặt trong xã hội hiện nay, tạo xu hướng mới cho người tiêu dùng

Điểm nổi bật của mạng xã hội mà ai cũng nhận thấy đó là tính kết nối và chia sẻ rất mạnh mẽ. Nó phá vỡ những ngăn cách về địa lý, ngôn ngữ, giới tính lẫn quốc gia.Những gì bạn làm, bạn nghĩ, cả thế giới có thể chia sẻ chỉ trong tích tắc.



Hình 2.2: Những ưu điểm của mạng xã hội

Cũng vì thế, từ một công cụ giao tiếp, mạng xã hội đã trở thành phương tiện truyền thông mới hiệu quả, đây chính là cách nhìn nhận mới cho doanh nghiệp tạo nên một cách truyền thông tích cực hơn đó là trên mạng xã hội.. 58% các công ty tiếp thị quảng cáo trên thế giới chính thức đưa truyền thông xã hội vào các chiến dịch kinh doanh của mình [7]

Với truyền thông truyền thống như là quảng cáo trên tivi, đặt biển quảng cáo ngoài trời, đặt banner trên các trang web điện tử… có những hạn chế như là chỉ tương

tác có một chiều, hiệu quả quảng cáo không cao, không được lan truyền rộng, chi phí lại cao...

Truyền thông trên mạng xã hội ra đời đã khắc phục được nhược điểm trên. Với những tính năng trên mạng xã hội, việc truyền thông sẽ có sự tương tác từ 2 phía, có thể nhận được sự phản hồi của khách hàng một cách nhanh chóng, hiệu quả, dựa trên các mối quan hệ, liên kết thì thông tin sẽ được lan truyền nhanh và rộng rãi hơn, sẽ giảm được chi phí cho việc truyền thông.

Tóm lại ưu điểm nổi trội của quảng cáo trên mạng xã hội:

Độ tương tác cao: Là một ưu điểm nổi trội của quảng cáo qua mạng xã hội . Doanh nghiệp có thể nhanh chóng tiếp cận ý kiến phản hồi từ khách hàng để cùng nhau trao đổi, chia sẻ vấn đề cùng họ, thực hiện các cuộc thăm dò hoặc giải đáp các thắc mắc khó khăn của họ, tạo ra một mối quan hệ thân thiện, sự tin tưởng của khách hàng dành cho doanh nghiệp. Thiết lập mối quan hệ với số đông khách hàng mục tiêu.

Tiết kiệm chi phí: Với việc bỏ ra chi phí nhỏ, Doanh nghiệp hoàn toàn có được kết quả tích cực. Xu hướng quảng cáo qua mạng xã hội được sử dụng rộng rãi trên thế giới.

Tính lan truyền: Khả năng lan truyền thông tin nhanh chóng và phải nói là 1 cách chóng mặt**.** Một khi các thông tin về sản phẩm của doanh nghiệp được đưa lên các trang web xã hội , các thông tin này được lan truyền từ người này sang người khác trong thời gian rất ngắn.

Tính cộng đồng: Khác với các kênh quảng cáo truyền thông khác là sản phẩm hoặc dịch vụ của doanh nghiệp chỉ đến với khách hàng theo hướng một chiều từ bạn, với mạng xã hội, doanh nghiệp có thể xây dựng cộng đồng mang tính tương hỗ giữa sản phẩm – khách hàng; khách hàng – sản phẩm – khách hàng. Sự phản hồi trực tiếp từ khách hàng sẽ giúp doanh nghiệp cải thiện sản phẩm và dịch vụ tốt hơn.

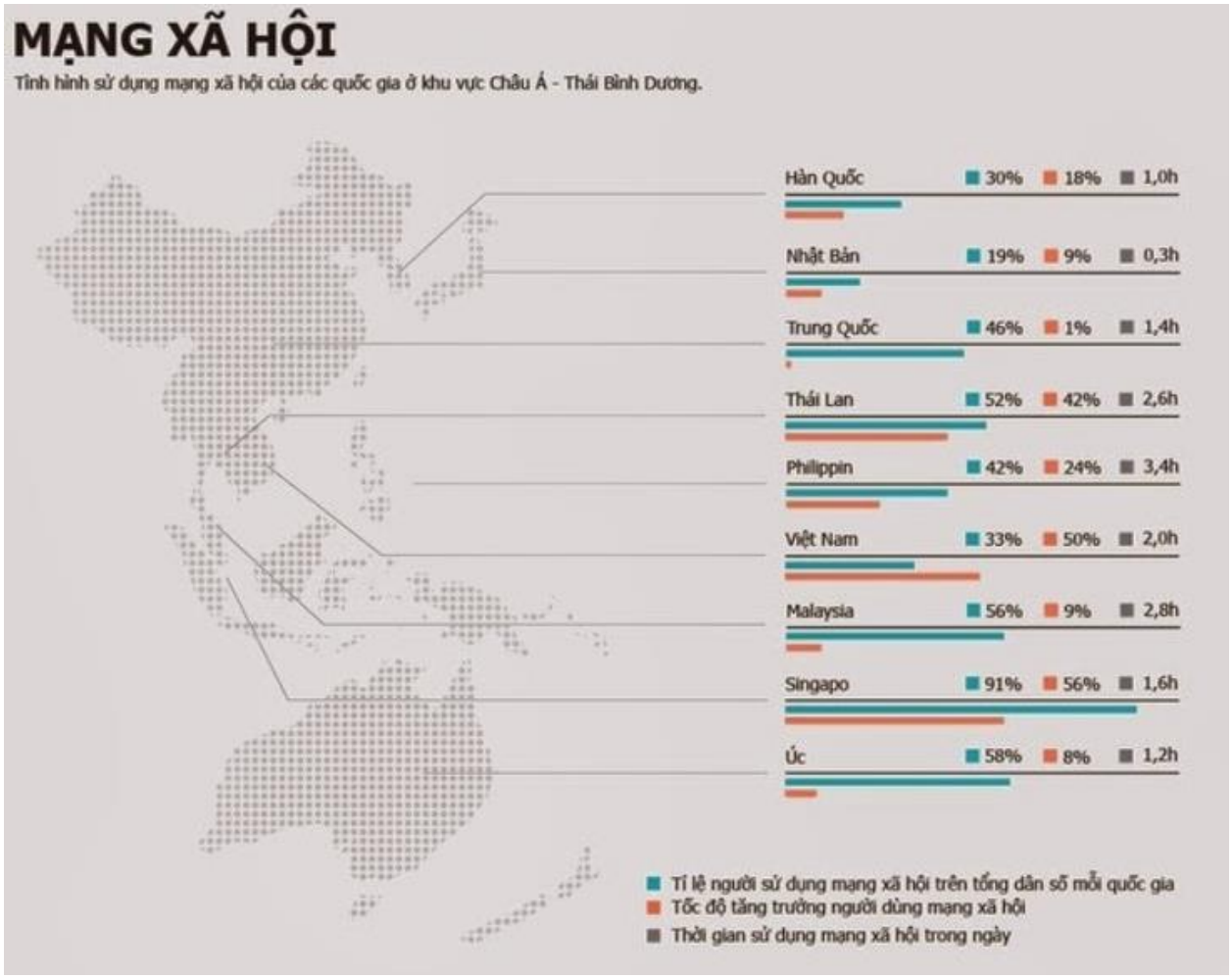
Đo lường, đánh giá kết quả: Với mỗi hình thức khác nhau của Social media, sẽ có những thông số cần thiết để làm cơ sở đánh giá hiệu quả, kết hợp việc sử dụng

phương thức đo lường tương ứng. Xử lý dữ liệu, đưa ra đánh giá, nhận định của chiến dịch.

2.2.2/ Tiềm năng quảng cáo trên mạng xã hội

Theo thống kê của Simon Kemp[21], tính đến tháng 1/2015 trên thế giới đã có 2078 tỷ người dùng mạng xã hội, tăng hơn năm 2014 là 222 triệu người.

Với những ưu điểm vượt trội của quảng cáo trên mạng xã hội kết hợp với xu hướng người dùng mạng xã hội ngày càng tăng vượt bậc như hiện nay thì quảng cáo trên mạng xã hội là một mảng kinh tế sẽ thu về nhiều lợi nhuận với tiềm năng to lớn.



Hình 2.3 : Tình hình sử dụng mạng xã hội của các quốc gia khu vực Châu Á – Thái Bình Dương [nguồn We are Social]

2.2.3. Các cách thức quảng cáo trên mạng xã hội

2.2.3.1/ Quảng cáo tìm kiếm (Search Marketing)

Theo thói quen, người dùng Internet khi muốn mua một sản phẩm, dịch vụ nào đó thường tra cứu trên Google, Yahoo, Bing... Nhà quảng cáo sẽ thông qua các đại lý hoặc trực tiếp trả tiền cho các công cụ quảng cáo để sản phẩm dịch vụ của họ được hiện lên ở các vị trí ưu tiên. Họ cũng có thể lựa chọn nhóm người xem quảng cáo theo vị trí địa lý, độ tuổi và giới tính hoặc theo một số tiêu chí đặc biệt khác.

Nhờ vậy, doanh nghiệp có thể hướng tới đúng nhóm khách hàng mục tiêu, tăng hiệu quả tiếp thị, đồng thời có thể theo dõi, thống kê mức độ hiệu quả của mỗi từ khóa để kiểm soát cả chiến dịch và tạo dựng thương hiệu tốt hơn.

2.2.3.2/ Quảng cáo theo mạng lưới trên Internet (Ad-network)

Thay vì gõ cửa từng đại lý (agency) hoặc phòng quảng cáo của mỗi tờ báo, giờ đây, nhà quảng cáo có thể thông qua mạng quảng cáo trực tuyến - phương tiện hiệu quả để xây dựng các chiến dịch quảng cáo. Mạng quảng cáo trực tuyến là hệ thống trung gian kết nối bên bán và bên mua quảng cáo trực tuyến, hỗ trợ người mua quảng cáo tìm thấy những vùng và website bán quảng cáo phù hợp với chiến dịch truyền thông của mình từ hàng nghìn website. Hình thức này hiện được nhiều công ty đánh giá cao vì nó giúp tiết kiệm thời gian và chí phí nhân. Tại Việt Nam, Innity, Vietad, Ambient là những mạng lớn có thể đáp ứng hầu hết các nhu cầu của nhà quảng cáo.

2.2.3.3/ Quảng cáo trên mạng xã hội (Social Media Marketing)

Với sự phát triển của hàng loạt mạng xã hội như Facebook, Twitter, Go, Yume..., người làm tiếp thị có thêm lựa chọn để tiếp cận cộng đồng. Khi sử dụng hình thức này, doanh nghiệp thường quảng bá dưới dạng hình ảnh, video clip có khả năng phát tán và thu hút bình luận (comment). Tính tương tác chính là ưu điểm nổi trội của loại hình này so với các kiểu marketing truyền thống. Theo Tim O'Reilly thuộc công ty O'Reilly Media, social media (truyền thông xã hội) "không phải để nói về bạn, về sản phẩm hay câu chuyện của bạn. Nó phải tạo ra những giá trị cho cộng đồng mà có

bạn trong đó. Càng nhiều giá trị bạn mang lại cho cộng đồng thì sẽ càng nhiều lợi ích cộng đồng mang đến cho bạn".

2.2.3.4/ Marketing tin đồn (Buzz Marketing)

Viral Marketing (phát tán kiểu virus), Buzz Marketing (marketing tin đồn) hay Words Of Mouth Marketing (marketing truyền miệng) được thực hiện thông qua blog, mạng xã hội, chat room, diễn đàn… bắt đầu từ giả thuyết người này sẽ kể cho người kia nghe về sản phẩm hoặc dich vụ họ thấy hài lòng. Viral Marketing là chiến thuật nhằm khuyến khích khách hàng lan truyền nội dung tiếp thị đến những người khác, để sản phẩm và dịch vụ được hàng ngàn, hàng triệu người biết đến. Trường hợp của Susan Boyle, thí sinh của chương trình Britain's Got Talent, là một ví dụ. Cô nổi tiếng toàn cầu chỉ sau một đêm khi đoạn video của cô trên YouTube nhờ được hàng triệu người chia sẻ.

Chi phí thực hiện Buzz/Viral Marketing không nhiều và hiệu quả truyền thông lại rất cao nhưng cũng là một phương thức marketing tiềm ẩn rủi ro nếu sản phẩm và dịch vụ của doanh nghiệp chưa thực sự tốt như cách mà họ quảng cáo.

2.2.3.5/ E-mail marketing

E-mail đang dần thay thế cách gửi thư qua bưu điện và doanh nghiệp có thể nhanh chóng gửi thông tin tiếp thị tới hàng loạt địa chỉ e-mail với chi phí rẻ. Một hình thức khác mà doanh nghiệp có thể áp dụng là khuyến khích đăng ký nhận bản tin điện tử (eNewsletters) nhằm tạo sự chủ động tiếp nhận thông tin cho khách hàng, từ đó tạo tâm lý thoải mái, thiện cảm với thông tin doanh nghiệp đem đến.

Tại Việt Nam, nhiều doanh nghiệp đã sử dụng quả hình thức e-mail marketing để tiếp cận khách hàng như Vietnamworks với bản tin việc làm, Jetstar với bản tin khuyến mãi giá vé máy bay, Nhommua hay Muachung với các e-mail thông tin về mặt hàng giảm giá.

2.2.4/ Ba hình thức quảng cáo cụ thể trên Facebook

Có 3 hình thức chính: Facebook Ads – Sponsored Stories – Post Engagement, mỗi hình thức phù hợp với một mục tiêu marketing nhất định và mang lại hiệu quả khác nhau.

Điểm giống nhau ở cả 3 hình thức:

- Chi phí tính theo click (CPC) hoặc theo lượt hiển thị (CPM).
- Lựa chọn tiếp cận người xem theo: Độ tuổi, Giới tính, Địa lý, Ngôn ngữ, Sở thích, Tình trạng hôn nhân, Học vấn…

2.2.4.1/ Facebook Ads

Hình 2.4: Mô tả vị trí quảng cáo của Facebook Ads

- Mẫu quảng cáo truyền thống trên Facebook.
- Mẫu quảng cáo bao gồm 1 hình ảnh 100×72 px, dòng tiêu đề tối đa 25 ký tự, đoạn mô tả tối đa 90 ký tự.
- Chỉ hiển thị bên phải trang Facebook.
- Link trực tiếp về website của khách hàng.

2.2.4.2/ Sponsored Stories

Hình 2.5: Mô tả vị trí quảng cáo của Sponsored Stories

Quảng cáo Facebook dạng Sponsored Stories của Vietjet Air do Infolinks thực hiện.

- Mẫu quảng cáo gồm 1 hình ảnh + text được lấy tự động từ Fanpage, hoặc 1 post trong Fanpage
- Có thể hiển thị trong News Feed
- Link về Fanpage hoặc 1 post trong Fanpage
- Tiếp cận người xem theo: Fan của Fanpage, Friends của Fan của Fanpage
- Người xem có thể trực tiếp Like Fanpage, Share – Comment ngay trên mẫu quảng cáo

2.2.4.3/ Post Engagement hay Promoted Post

Hình 2.6: Facebook Post Engagement / Promoted Post

Mẫu quảng cáo gồm 1 hình ảnh + text được lấy tự động từ 1 post trong Fanpage

Chỉ hiển thị trong News Feed Link về 1 post trong Fanpage

Tiếp cận người xem theo: Fan của Fanpage, Friends của Fan của Fanpage

Người xem có thể trực tiếp Like fanpage, Share – Comment post ngay khi trên mẫu quảng cáo

Dựa vào mô hình AISAS do Densu đề xuất về hành vi của người dùng, nhà kinh doanh sẽ thấy được giai đoạn nào lựa chọn hình thức nào là phù hợp.

Hình 2.7: Mô hình AISAS

Attention – Gây chú ý:

Gây chú ý với 1 sản phẩm mới thông qua Facebook là một lựa chọn hết sức phù hợp, Facebook có gần 10 triệu người dùng tại Việt Nam (Tháng 11/2012) chiếm 30% lượng người dùng internet, độ tuổi tập trung từ 18 – 30, và tăng thêm khoảng 500.000 người dùng mới mỗi tháng. Quảng cáo trên Facebook dễ dàng và hết sức nhanh chóng tiếp cận toàn bộ hay một phần lượng khách hàng rộng lớn này.

Gây chú ý với 1 Fanpage mới là khá khó khăn. Từ việc xây dựng nội dung thật tốt và đều đặn, hay tổ chức các cuộc thi thì giải pháp nhanh và phù hợp nhất là làm quảng cáo Facebook Ads để gây sự chú ý cho đúng khách hàng tiềm năng (độ tuổi, giới tính, khu vực…).

Interest – Thích:

Khi đã đạt được mục tiêu gây sự chú ý hay tăng độ nhận diện thương hiệu cho sản phẩm mới hoặc đạt được mục tiêu về số lượng fan cho Fanpage, nhà kinh doanh có thể gây sự thích thú bằng cách sử dụng Sponsored Stories kết hợp với Promoted Post và cần phải target đúng đối tượng khách hàng tiềm năng chính là những người đã và có khả năng sẽ thích sản phẩm của nhà kinh doanh, chỉ cần chọn target audience là Fans + Friends of Fans + tùy chọn theo sở thích cho các mẫu quảng cáo của nhà kinh doanh.

Search / Action – Tìm kiếm / Mua hàng:

Sau khi khách hàng thích thú sản phẩm của bạn thông qua việc nhìn thấy và click vào các mẫu quảng cáo họ có thể sử dụng các công cụ tìm kiếm (95% là Googe) để tìm hiểu thêm và mua sản phẩm (Action). Việc của nhà kinh doanh là đón sẵn ở đó bằng cách sử dụng Quảng cáo Google hay làm SEO!

Hoặc khi click vào mẫu quảng cáo trên Facebook khách hàng sẽ được dẫn ngay tới website hay Fanpage của nhà kinh doanh. Việc còn lại của nhà kinh doanh là thuyết phục họ Action mà thôi!

Share – Chia sẻ:

Khách hàng sau khi mua cảm thấy hài lòng và thích thú, họ sẽ quay lại Fanpage của nhà kinh doanh để cảm ơn, khen ngợi, bấm Like… Vậy làm sao để cho càng

nhiều người khác biết về những lời khen quý giá đó? Sponsored Stories là câu trả lời phù hợp nhất.

Hãy đặt câu hỏi "Nếu bạn thấy ai đó Like một post của Fanpage nào đó, liệu bạn có tò mò để click vào xem thử hay không? Nhất là khi người đó lại là bạn của bạn" –> Sponsored Stories là một hình thức cực kỳ hiệu quả để tạo hiệu ứng truyền miệng và gây ảnh hưởng đến đối tượng khách hàng mục tiêu của nhà kinh doanh.

2.3/ Tổng quan về phát triển hệ thống quảng cáo thông minh trên mạng xã hội

2.3.1/ Quảng cáo thông minh trên mạng xã hội

Các thông tin trên mạng xã hội lan truyền một cách nhanh chóng, dựa vào các đặt tính này khi người dùng hoặc doanh nghiệp muốn tìm kiếm thông tin hay đăng quảng cáo thì có thể trực tiếp tìm kiếm thông tin về những pages, diễn đàn hay groups thuộc các lĩnh vực để đăng quảng cáo của mình cũng như tìm kiếm thông tìn cần thiết cho công việc.

Dựa vào ưu thế này chúng tôi đề xuất một hệ thống phân tích dữ liệu (thường là các mẫu quảng cáo) sau đó khi hệ thống phân tích xong, sẽ đề xuất một số pages và groups cho người dùng có thể đăng quảng cáo và tìm hiểu thông tin mình cần, ngoài ra hệ thống còn cho người dùng biết ước lượng phần trăm tỉ lệ chính xác hệ thống phân tích được – Đó chính là *hệ thống quảng cáo thông minh trên mạng xã hội*

Do đó một thách thức rõ ràng cho hướng nghiên cứu là vấn đề tìm kiếm thông tin, chủ yếu dựa vào các hoạt động cá nhân và doanh nghiệp.

Các phương pháp dùng để lấy thông tin một cách hiệu quả cho phương tiện truyền thông xã hội, đã trở thành một hướng nghiên cứu quan trọng để thu thập thông tin từ mạng xã hội.

2.3.2/ Hệ thống quảng cáo thông minh trên mạng xã hội:

Hệ thống Information Retrieval Social Media bao gồm các chức năng:

- Hệ thống sẽ crawl dữ liệu từ các trang mạng xã hội để làm dữ liệu chính.

- Lọc dữ liệu và phân loại các dữ liệu đã được crawl về, sau đó biểu diễn chúng thành các vector có thể tính toán được.
- Hệ thống sẽ tính toán dựa vào dữ liệu đầu vào (biểu diễn thành các vector) với các tập dữ liệu đã có sẵn và đưa ra kết quả mong muốn.

Các dữ liệu mà hệ thống crawl về (hiện tại hệ thống crawl dữ liệu từ Facebook và Youtube) sẽ được lọc và biểu diễn thành các vector có thể tính toán được, sau đó đưa vào lưu trữ dưới dạng Database (DB - hệ thống hiện tại đang dùng MySQL). Trong quá trình crawl dữ liệu hệ thống dùng giải thuật TF -IDF (Term Frequency–Inverse Do cument Frequency) để phân tích dữ liệu đưa ra các từ từ khóa theo từng categories, các từ khóa đặt trưng này dùng để phân loại dữ liệu. Trong quá trình này để cho dữ liệu được chính xác hơn và ít bị ảnh hưởng, sẽ có một bước lọc ra những từ khóa nhiễu và loại bỏ chúng, bước này được thực hiện có sự can thiệp của con người.

Tiếp theo hệ thống xây dựng từng module dùng để tính toán khoảng cách (Module Distance Calculation) dựa trên dữ liệu đầu vào và dữ liệu lưu trữ trên DB.

Dựa vào khoảng cách này hệ thống sẽ biết được các pages hay groups nào sẽ nổi lên, phù hợp với dữ liệu đầu vào.

Tiếp theo sau phần phân tích dữ liệu đầu vào và xác định các pages và groups nổi lên, hệ thống sẽ tính toán độ chính xác dựa trên dữ liệu có được, tùy vào khả năng chi phí cho số lượng pages và groups mà người dùng muốn hệ thống đưa ra. Hệ thống sẽ ước lượng độ chính xác trên dữ liệu cung cấp cho người dùng.

Cuối cùng hệ thống sẽ đề xuất một số pages hoặc group cho người dùng, tùy thuộc nhu cầu sử dụng.

2.3.3/ Khai phá dữ liệu để xây dựng hệ thống quảng cáo thông minh trên mạng xã hội

Các mạng truyền thông xã hội đã và đang tạo ra vô số dữ liệu từ người dùng, làm thế nào để các doanh nghiệp có thể chuyển dữ liệu bình luận thô trong các mạng xã hội như Twitter, Facebook, các blog và các diễn đàn thành những hiểu biết kinh

doanh? Câu trả lời nằm ở việc áp dụng công nghệ ngữ nghĩa và khai phá văn bản cho các nguồn dữ liệu không có cấu trúc này.

Khai phá văn bản đề cập đến các kỹ thuật được sử dụng trong việc trích ra thông tin từ các nguồn văn bản viết khác nhau. Điều này rất quan trọng. Người ta đã ước tính rằng 80% thông tin liên quan đến kinh doanh nằm trong dữ liệu văn bản không có cấu trúc và nửa cấu trúc. Nói cách khác, nếu thiếu ứng dụng cho việc phân tích văn bản để tìm ra nội dung phong phú của dữ liệu được biểu diễn trong 80% đó, thì đã lãng phí tất cả dữ liệu hành vi người tiêu dùng và thông tin kinh doanh nhúng trong đó.

Thuật ngữ khai phá văn bản, còn được gọi là khai thác dữ liệu văn bản hoặc khám phá tri thức từ văn bản cơ sở dữ liệu [3,4], thường được coi là phân tích văn bản, có nhiều mục đích thực tế, chẳng hạn như các ứng dụng lọc thư rác, trích ra thông tin từ các đề xuất và các khuyến nghị trên các trang web thương mại điện tử, lắng nghe xã hội và khai phá dư luận từ các blog và các trang web phê bình, nâng cao dịch vụ khách hàng và hỗ trợ thư điện tử (email), xử lý tự động các tài liệu kinh doanh, khám phá điện tử (e-discovery) trong lĩnh vực pháp lý, đo lường sở thích của người tiêu dùng, phân tích tổn thất, phát hiện gian lận, tội phạm mạng và các ứng dụng an ninh quốc gia.

Khai phá văn bản tương tự như khai phá dữ liệu ở chỗ nó được nhằm vào việc xác định các mẫu dữ liệu đáng chú ý. Mặc dù việc khai phá văn bản thủ công (cần nhiều người làm) đã nổi lên trong những năm 1980. Lĩnh vực khai phá văn bản đã trở nên quan trọng trong những năm gần đây để tinh chỉnh các thuật toán kết quả của công cụ tìm kiếm và chọn lọc thông qua các nguồn dữ liệu để khám phá các thông tin chưa biết. Tất cả các kỹ thuật như máy học, thống kê, ngôn ngữ học máy tính và khai phá dữ liệu đều được sử dụng trong quá trình này. Mục tiêu của việc khám phá tri thức từ văn bản, ví dụ, là để phát hiện ra các mối quan hệ ngữ nghĩa nằm bên dưới văn bản cũng như nội dung và bối cảnh ngụ ý với NLP (Natural Language Processing - Xử lý ngôn ngữ tự nhiên). Các quá trình này đều nhằm vào việc sử dụng NLP để

sao chép lại, rồi điều chỉnh quy mô cho hợp với cùng kiểu phân biệt ngôn ngữ, nhận dạng mẫu và hiểu kết quả, diễn ra khi con người đọc và xử lý văn bản.

Các phương pháp khác nhau tồn tại trong lĩnh vực khai phá văn bản. Dưới đây giới thiệu một danh sách các bước tuần tự và phổ biến liên quan đến việc khai phá văn bản.

Bước đầu tiên trong bất kỳ nỗ lực khai phá văn bản nào là xác định các nguồn dựa trên-văn bản cần được phân tích và thu thập tư liệu này thông qua việc lấy ra thông tin hoặc chọn kho văn bản chuyên đề (corpus) gồm một tập hợp các tệp văn bản và nội dung đang quan tâm.

Sau đó triển khai NLP mở rộng, gọi ra "thành phần gắn thẻ tiếng nói" và sắp xếp thứ tự văn bản để phân tích cú pháp (đó là, *biểu tượng hóa* (tokenizing) văn bản) và áp dụng *nhận dạng thực thể có tên* (Named Entity Recognition) (đó là, nhận biết việc nêu ra các nhãn hàng, các tên người, các địa điểm, các chữ viết tắt phổ biến và v.v).

Một bước *Lọc các từ phổ biến* (Filter Stopwords) hay dùng liên quan đến việc loại bỏ các từ phổ biến để tinh lọc nội dung của chủ đề mong muốn. *Các thực thể đã xác định mẫu* (Pattern Identified Entities) nhận biết các địa chỉ email và các số điện thoại và *Tài liệu cùng tham khảo* (Coreference) xác định các cụm danh từ và các đối tượng liên quan trong văn bản

Tiếp theo là *Trích ra mối quan hệ, sự vật và sự kiện (Relationship, Fact and Event Extraction). Các N-Gram* thường được sinh ra để tạo các điều kiện dưới dạng một loạt từ liên tiếp.

Cuối cùng, một cách tiếp cận được các công cụ lắng nghe và phân loại môi trường truyền thông xã hội hiện nay sử dụng rộng rãi là *phân tích tâm lý tiêu dùng*, để trích ra thông tin về thái độ theo đối tượng hoặc chủ đề nào đó. Thông thường, các chức năng lập bản đồ và vẽ đồ thị khác cung cấp hiển thị trực quan để kiểm tra chính xác hơn nữa.

2.3.3.1/ Các công cụ khai phá văn bản

Có một số tùy chọn nguồn mở và thương mại cho phần mềm và các ứng dụng khai phá văn bản.

IBM cung cấp một loạt các giải pháp khai phá văn bản rộng lớn và mạnh mẽ. Một sản phẩm mạnh, sử dụng các khả năng Big Data của IBM® InfoSphere® BigInsights™, cung cấp một mô đun phân tích văn bản bổ sung, thực hiện trích ra phân tích văn bản từ cụm BigInsights InfoSphere.

Các sản phẩm IBM SPSS® trải rộng theo quy mô và phạm vi. Một công cụ, hoạt động tốt để tìm kiếm một tài liệu và gán nó cho một chủ đề hay chuyên đề là IBM SPSS Modeler (Trình mô hình hóa SPSS của IBM), cung cấp một giao diện đồ họa để thực hiện phân loại và phân tích tài liệu văn bản tổng quát.

Một sản phẩm khác là IBM SPSS Text Analytics for Surveys (Phân tích văn bản SPSS của IBM dành cho khảo sát điều tra) sử dụng NLP để phân tích các câu hỏi khảo sát mở trong một tài liệu.

IBM SPSS Modeler Premium chạy trên cùng một công cụ như SPSS Text Analytics dành cho khảo sát, nhưng có khả năng mở rộng quy mô cao để xử lý toàn bộ kho dữ liệu gồm nhiều loại tài liệu (PDF, các trang web, các blog, email, các nguồn cấp dữ liệu Twitter và nhiều hơn nữa) trong một nhánh công việc, để tạo điều kiện thuận lợi cho việc tích hợp giữa dữ liệu có cấu trúc và không có cấu trúc. Một nút mã nguồn tùy chỉnh liên quan dành cho Facebook mở rộng các khả năng của SPSS Modeler Premium để đọc dữ liệu trực tiếp từ một trang Facebook và tích hợp nó với một nguồn cấp dữ liệu Twitter trong SPSS Modeler để có được phối cảnh nhiều kênh truyền thông xã hội.

Trong số các công cụ khai phá văn bản nguồn mở, RapidMiner và R dường như là hai công cụ phổ biến nhất. R có một cơ sở người dùng rộng hơn; một ngôn ngữ lập trình yêu cầu có mã nguồn trong đó, nó có một lựa chọn lớn về các thuật toán. Tuy nhiên, khả năng điều chỉnh quy mô là một vấn đề với R nên nó không phải là lý tưởng cho các tập dữ liệu lớn (big data) nếu không có các cách giải quyết. RapidMiner có một cơ sở người dùng nhỏ hơn, nhưng nó không đòi hỏi mã nguồn và có một giao

diện người dùng (UI) mạnh mẽ. Nó cũng có khả năng điều chỉnh quy mô cao và có thể xử lý các cụm và lập trình trong cơ sở dữ liệu. IBM cung cấp một mô đun Jaql R có tích hợp dự án R trong các truy vấn, còn về phần mình dự án R lại cho phép các tác vụ MapReduce chạy tính toán R song song.

2.3.3.2/ Các kho dữ liệu của môi trường truyền thông xã hội và Big Data

Khi bắt đầu áp dụng khai phá văn bản, có những thách thức đặc biệt riêng của dữ liệu của môi trường truyền thông xã hội. Dữ liệu, do các trang web mạng xã hội, các blog và các diễn đàn tạo ra, rơi vào thể loại của những thứ thường được gọi là big data. Dữ liệu thường không có cấu trúc và nửa cấu trúc, tạo ra rất nhiều petabyte dữ liệu hàng ngày xung quanh các nhãn hàng lớn và các cơ sở dữ liệu quan hệ truyền thống không thể mở rộng quy mô có hiệu quả để hỗ trợ phân tích thời gian thực dựa trên dữ liệu đó. Vì thế rất cần các giải pháp cơ sở dữ liệu NoSQL và big data.

Dữ liệu của môi trường truyền thông xã hội, nếu không được thu thập và lưu trữ thích hợp theo định kỳ đều đặn, về cơ bản dễ mất đi. Hầu hết các công cụ nguồn mở lắng nghe mạng xã hội chỉ lưu lịch sử bình luận của môi trường truyền thông xã hội trong một vài ngày. Chỉ có Twitter mới đây đã thông báo rằng toàn bộ lịch sử của dữ liệu sẽ có sẵn, nhưng nó sẽ được giới hạn với các bình luận do chủ tài khoản đăng lên. Dữ liệu này có sẵn từ một số các nhà cung cấp dữ liệu xã hội lớn hơn đã nói ở trên, chẳng hạn như Gnip và DataSift và thông qua rất nhiều giao diện lập trình ứng dụng (các API) và các giao diện lập trình ứng dụng dựa trên cuộc gọi thông qua các công cụ khác. Tuy nhiên, ở nơi dữ liệu có sẵn (đối với Twitter), nó vẫn rất đắt với tất cả mọi người, trừ những doanh nghiệp lớn nhất.

Mỗi trang web của môi trường truyền thông xã hội xử lý vấn đề này một cách khác nhau. Mỗi trang có thể sử dụng các yêu cầu tìm kiếm và có các đáp ứng theo định dạng JavaScript Object Notation (JSON), có dữ liệu chưa được phân tích cú pháp để đưa ngay vào một cơ sở dữ liệu MySQL hoặc cơ sở dữ liệu NoSQL, tùy thuộc vào khối lượng và tính chất của dữ liệu.

CHƯƠNG 3
CƠ SỞ LÝ THUYẾT

3.1/ Các đề tài nghiên cứu trên thế giới

David R. H. Miller (2007) [16] đã đưa ra phương pháp và hệ thống IR cải thiện thực hiện tìm kiếm thông tin bằng cách sử dụng xác suất. Khi thực hiện truy vấn thông tin, hệ thống IR cải thiện tận dụng cả hai khả năng có thể xảy ra: một tài liệu là truy vấn độc lập có liên quan cũng như những khả năng mà các truy vấn đã được tạo ra bởi một tài liệu cụ thể có liên quan. Bằng cách sử dụng các khả năng này, hệ thống IR cải thiện lấy tài liệu một cách chính xác hơn so với các hệ thống thông thường dựa trên phương pháp cũ.

Tie-Yan Liu (2009) [18] Learning-to-rank cho Information Retrieval (IR) là một nhiệm vụ tự động xây dựng một mô hình xếp hạng sử dụng training data, như là, mô hình có thể sắp xếp các đối tượng mới theo bằng cấp liên quan, sở thích, hoặc tầm quan trọng. Công nghệ IR có thể có khả năng nâng cao bằng cách sử dụng các kỹ thuật học tập-to-rank để giải quyết các vấn đề. Cụ thể, các thuật toán Learning-to-rank được xem xét và phân loại thành ba phương pháp: phương pháp tiếp cận theo từng điểm, phương pháp tiếp cận theo từng cặp, và phương pháp tiếp cận theo từng danh sách. Những lợi thế và bất lợi với mỗi phương pháp được phân tích và dường như cho thấy rằng cách tiếp cận listwise là một trong những hiệu quả nhất trong số tất cả các phương pháp tiếp cận. Sau đó, một lý thuyết thống kê xếp hạng được giới thiệu, có thể mô tả các thuật toán học khác nhau của Learning-to-rank, và được sử dụng để phân tích khả năng khái quát hóa cấp truy vấn của nó. Cuối của nghiên cứu, họ cung cấp một bản tóm tắt và thảo luận về công việc tiềm năng trong tương lai của việc Learning-to-rank.

Xuerui Wang (2007) [19] Hầu hết các mô hình chủ đề, chẳng hạn như phân bổ Dirichlet tiềm ẩn, dựa trên giả định túi-of-từ. Tuy nhiên, thứ tự từ và cụm từ thường quan trọng để nắm bắt ý nghĩa của văn bản trong nhiều nhiệm vụ khai phá văn bản. Bài viết này trình bày chuyên đề n-gram, một mô hình chủ đề mà phát hiện ra các chủ

đề cũng như các cụm từ chuyên đề. Các mô hình xác suất tạo ra từ theo thứ tự văn bản của mình bằng cách, mỗi từ, đầu tiên lấy mẫu một chủ đề, sau đó lấy mẫu tình trạng của nó như là một unigram hoặc Bigram, và sau đó lấy mẫu từ từ một unigram hoặc Bigram phân phối theo từng chủ đề. Như vậy mô hình của chúng tôi có thể mô hình "nhà trắng" là một cụm từ ý nghĩa đặc biệt trong chủ đề 'chính trị', nhưng không phải trong 'bất động sản' chủ đề. Bigrams tiếp tạo thành cụm từ dài hơn. Chúng tôi trình bày các thí nghiệm cho thấy các cụm từ có ý nghĩa và chủ đề của phiên dịch được nhiều hơn từ các dữ liệu NIPS và cải thiện hiệu suất thu hồi thông tin về một bộ sưu tập TREC.

3.2/ Kỹ thuật trích xuất thông tin từ văn bản

3.2.1/ Khái niệm

Khai thác văn bản (Text Mining) là một nhánh của Data Mining nhằm tìm kiếm và trích xuất thông tin nằm trong văn bản. Hiện nay, với sự tăng trưởng nhanh chóng của dữ liệu văn bản, Text Mining ngày càng có nhiều ứng dụng trong thực tế như:

- Lọc thư rác
- Đối chiếu lý lịch cá nhân
- Phân tích tình cảm (Sentiment)
- Phân loại tài liệu

3.2.2/ Nội dung

Hiện nay, cơ sở dữ liệu văn bản (Text Database) đang phát triển một cách nhanh chóng và thu hút nhiều sự quan tâm của giới nghiên cứu b ởi sự gia tăng nhanh chóng số lượng thông tin ở dạng số hóa.Ví dụ như các loại tài liệu điện tử, email, thư điện tử, và các trang web. Có thể thấy hầu hết thông tin của các chính phủ, các ngành công nghiệp, kinh doanh, trường học đều được số hóa và lưu trữ ở dưới dạng cơ sở dữ liệu để phục vụ việc tính toán.

Dữ liệu lưu trữ trong cơ sở dữ liệu văn bản là dữ liệu bán cấu trúc (Semistructrured Data), điều đó co nghĩa chúng không hoàn toàn phi cấu trúc, nhưng

thực chất cũng không hoàn toàn có cấu trúc. Ví dụ: một tài liệu có thể chứa một vài trường có cấu trúc chẳng hạn như tiêu đề, tên tác giả, ngày xuất bản, phân loại. Nh ưng cũng có thể chứa một lượng lớn những thành phần văn bản phi cấu trúc chẳng hạn như phần tóm tắt hay nội dung của tài liệu.

Do đó vấn đề đặt ra là làm sao để có thể tìm kiếm và khai thác tri thức từ nguồn dữ liệu này. Các kỹ thuật để giải quyết vấn đề này đượ c gọi là kỹ thuật Text Mining hay thường đượ c gọi là khai phá dữ liệu văn bản.

Khai phá văn bản chia thành các vấn đề nhỏ hơn bao gồm phân loại văn bản (Text Categorization), gom cụm văn bản (Text Clustering), rút trích thực thể (Entity Extrac tion), phân tích quan điểm (Sentiment Analysis), tóm tắt tài liệu (Do cument Summarization), và mô hình hóa quan hệ giữa các thực thể (Entity Relation Mo deling). Trong hệ thống này chúng ta c ó sử dụng lý thuyết về phân loại văn bản (Text Categorization) để phân loại dữ liệu khi crawl về, sau đó chia thành các domains để quản lý.

3.3/ Vector Space Model

Phương pháp này mô hình hóa một tài liệu thành một vector. Để làm được điều này, trước tiên cần phải đánh trọng số cho các từ trong tài liệu. Cách tiếp cận đơn giãn nhất là gán trọng số bằng số lần xuất hiện của từ trong tài liệu (từ xuất hiện càng nhiều, trọng số càng lớn, mức độ quan trọng càng cao). Trọng số này được gọi là tần suất xuất hiện của từ **t** trong văn bản **d** và được ký hiệu là $\mathbf{tf_{t,d}}$.

Cách tiếp cận này gặp phải vấn đề khi ta xét một tập các tài liệu, mà trong đó từ "quan trọng" xuất hiện trong hầu hết các văn bản. Ví dụ: Khi xét một tập tài liệu nói về ngành công nghiệp ô tô. Từ "ô tô" xuất hiện gần như là tất cả các văn bản với tần suất khá cao. Nhưng xét trong phạm vi cả tập tài liệu thì từ "ô tô" gần như là không quan trọng. Do đó, để hạn chế tầm ảnh hưởng của các từ như thế này, ta xét tiếp hệ số:

$$Idf_t = \log(N/df)$$

Trong đó:

- N: là tổng số tài liệu trong tập văn bản đang xét
- df: là tổng số các tài liệu trong tập văn bản đang xét có chưa từ t

Trọng số của từ t lúc này sẽ được cho bởi giá trị:

$$\text{tf-idf}_t = \text{tf}_{t,d} \text{ x } \text{idf}_t$$

Các từ xuất hiện trong các tài liệu sẽ được thu thập lại thành không gian n chiều.

Mỗi văn bản sẽ được xây dựng thành các vector n chiều với các thành phần là trọng số các từ có trong văn bản.

Khi đó, việc tính độ tương tự giữa hai văn bản sẽ quy về tính góc giữa hai vector.

$$\text{Sim(d1,d2)} = \frac{V(d1).V(d2)}{|V(d1)||V(d2)|}$$

Khi nhận được một câu query với các từ khóa tìm kiếm, thông tin rút trích được từ hệ thống không phải là một tài liệu cụ thể nào mà sẽ là một tập các tài liệu có nội dung gần với yêu cầu tìm kiếm. Câu query sẽ được xử lý như một tài liệu rất ngắn và sẽ được sinh một vector tương ứng cho câu query này. Để tìm các tài liệu phù hợp với câu query này, cần phải tính độ tương tự giữa câu query với các tài liệu có trong tập hợp văn bản. Khi đó, độ tương tự giữa câu query **q** và tài liệu **d** sẽ được cho bởi công thức:

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

3.4/ Công cụ thu thập dữ liệu trên môi trường Internet(Crawler)

3.4.1/ Botnet

Botnet là thuật ngữ chỉ các chương trình hoạt động một cách tự động trên môi trường internet. Một ví dụ điển hình về botnet là các search engine. Ngoài chức năng tìm kiếm botnet còn có thể là các chương trình hoạt động tự động và mô phỏng hành động của con người (bot game).

3.4.2/ Các thành phần của một cỗ máy tìmn kiếm tự động

Hầu hết các hệ thống tìm kiếm được chia làm 3 phần cơ bản:

Crawler: Là thành phần đóng vai trò duyệt và tìm kiếm thông tin trên internet. Khi crawler viếng thăm một trang web, ngoài việc rút trích dữ liệu, crawler còn tiến hành cập nhật danh sách các url mới dựa vào các siêu liên kết trong trang web mà nó đang viếng thăm.

Dữ liệu thu thập được từ crawler sẽ là đầu vào cho quá trình xử lý của thành phần thứ 2: Lập chỉ mục (index). Quá trình lập chỉ mục sẽ tổ chức dữ liệu thành những kho dữ liệu đáp ứng quá trình tìm kiếm nhanh và chính xác. Quá trình lập chỉ mục là quá trình tiên quyết cho phép dữ liệu xuất hiện trên các máy tìm kiếm.

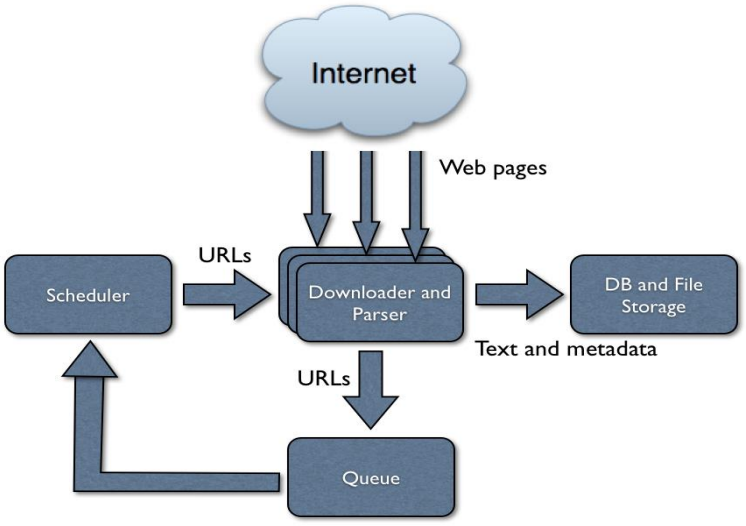
Khi dữ liệu đã được lập chỉ mục, thành phần thứ ba của một cỗ máy tìm kiếm là một phần mềm có chức năng tìm kiếm thông tin từ dữ liệu trên và cung cấp kết quả cho người dùng.

3.4.3/ Major Search Engines: Các cỗ máy tìm kiếm chính

Cùng có những thành phần cơ bản được mô tả ở trên và cùng hoạt động trên nguồn dữ liệu internet, các máy tìm kiếm khác nhau lại cho những kết quả khác nhau khi cùng nhận một bộ từ khóa. Sở dĩ có tình trạng này là do các máy tìm kiếm khác nhau được trang bị những thuật toán cũng như chiến lược tìm kiếm khác nhau

3.4.4/ Cấu trúc cơ bản và hoạt động của một crawler điển hình

Cấu trúc cơ bản:



Hình 3.1 : Cấu trúc cơ bản của một Crawler điển hình

Hoạt động: Một chương trình crawler gồm có một danh sách các URL chưa được viếng thăm gọi là frontier. Danh sách các URL ban đầu đã được cung cấp bởi người dùng hoặc các chương trình khác.

Quá trình hoạt động của crawler được mô tả như sau:

- Lấy ra URL cần được index tiếp theo từ frontier
- Nạp trang web tương ứng với URL
- Duyệt trang web vừa tải về để lấy ra các từ URL và các thông tin mà ứng dụng cần
- Thêm các trang URL chưa được thăm vào frontier. Trước khi các URL được thêm vào frontier.

Hoạt động của crawler sẽ kết thúc khi đã đạt được lượng dữ liệu cần thiết hoặc danh sách các URL trong frontier là rỗng.

3.5/ Giải thuật TF-IDF (TERM FREQUENCY – INVERSE DOCUMENT)

Định nghĩa 1: TF-IDF là mức độ quan trọng của một từ nằm trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

TF: Tần số xuất hiện của một từ trong một văn bản:

$$TF(t,d) = \frac{f(t,d)}{max\{f(w,d) : w \in d\}}$$

Thương của số lần xuất hiện một từ trong văn bản: là số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản đó (giá trị sẽ thuộc khoảng [0, 1]).

$f(t,d)$: số lần xuất hiện từ t (term) trong văn bản d (document).

$max\{f(w,d) : w \in d\}$: số lần xuất hiện nhiều nhất của một từ bất kỳ trong văn bản.

IDF: Nghịch đảo tần số của 1 từ trong tập văn bản. Mục đích tính IDF để giảm giá trị của những từ phổ biến. Mỗi từ chỉ có 1 giá trị IDF duy nhất trong tập văn bản.

$$IDF(t, D) = \log \frac{|D|}{|\{d \in D: t \in d\}|}$$

$|D|$: tổng số văn bản trong tập D.

$|\{d \in D: t \in d\}|$: số văn bản chứa từ nhất định, với điều kiện số từ đó phải $\neq 0$ (nghĩa là $TF(t, d) \neq 0$). Nếu từ đó không xuất hiện ở bất cứ một văn bản nào trong tập thì mẫu số sẽ bằng $0 \Rightarrow$ phép chia cho 0 không hợp lệ, chính vì vậy trong nghiên cứu thường mẫu sẽ thay bằng biểu thức $1 + |\{d \in D: t \in d\}|$.

Cơ số logarit trong công thức: cơ số không thay đổi giá trị của một từ mà chỉ thu hẹp hoảng giá trị của từ đó. Vì thay đổi cơ số sẽ dẫn đến việc giá trị của các từ thay đổi bởi 1 số nhất định và tỷ lệ giữa các trọng lượng với nhau sẽ không thay đổi. (nói cách khác, hay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF).

Tuy nhiên việc thay đổi khoảng giá trị sẽ giúp tỷ lệ giữa IDF và TF tương đồng để dùng cho công thức $TF - IDF$ như bên dưới.

TF-IDF: Những từ có giá trị $TF - IDF$ cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

Độ đo Recall, Presicion và F-measure

Giả sử một hệ thống trích xuất văn bản trả về một tập tài liệu dựa vào truy vấn đầu vào. Câu hỏi đặt ra là làm cách nào chúng ta đánh giá được độ chính xác hoặc độ đúng đắn của hệ thống.

Gọi tập các tài liệu có liên quan đến câu truy vấn là *{Relevant}* và tập các tài liệu được trích xuất trả về là *{Retrieval}*.

Tập các tài liệu vừa có liên quan vừa được trích xuất trả về sẽ được ký hiệu là $\{Relevant\} \cap \{Retrieval\}$

Có hai độ đo cơ bản cho việc đánh giá chất lượng trích xuất văn bản này gồm:

Recall: là tỷ lệ giữa các tài liệu có liên quan đến tài liệu truy vấn và trên thực tế được trích xuất trả về.

$$Recall = \frac{|\{Relevant\} \cap \{Retrieval\}|}{|\{Relevant\}|}$$

Precision: là tỷ lệ giữa các tài liệu được trả về thực sự có liên quan đến tài liệu truy vấn.

$$Precision = \frac{|\{Relevant\} \cap \{Retrieval\}|}{|\{Retrieval\}|}$$

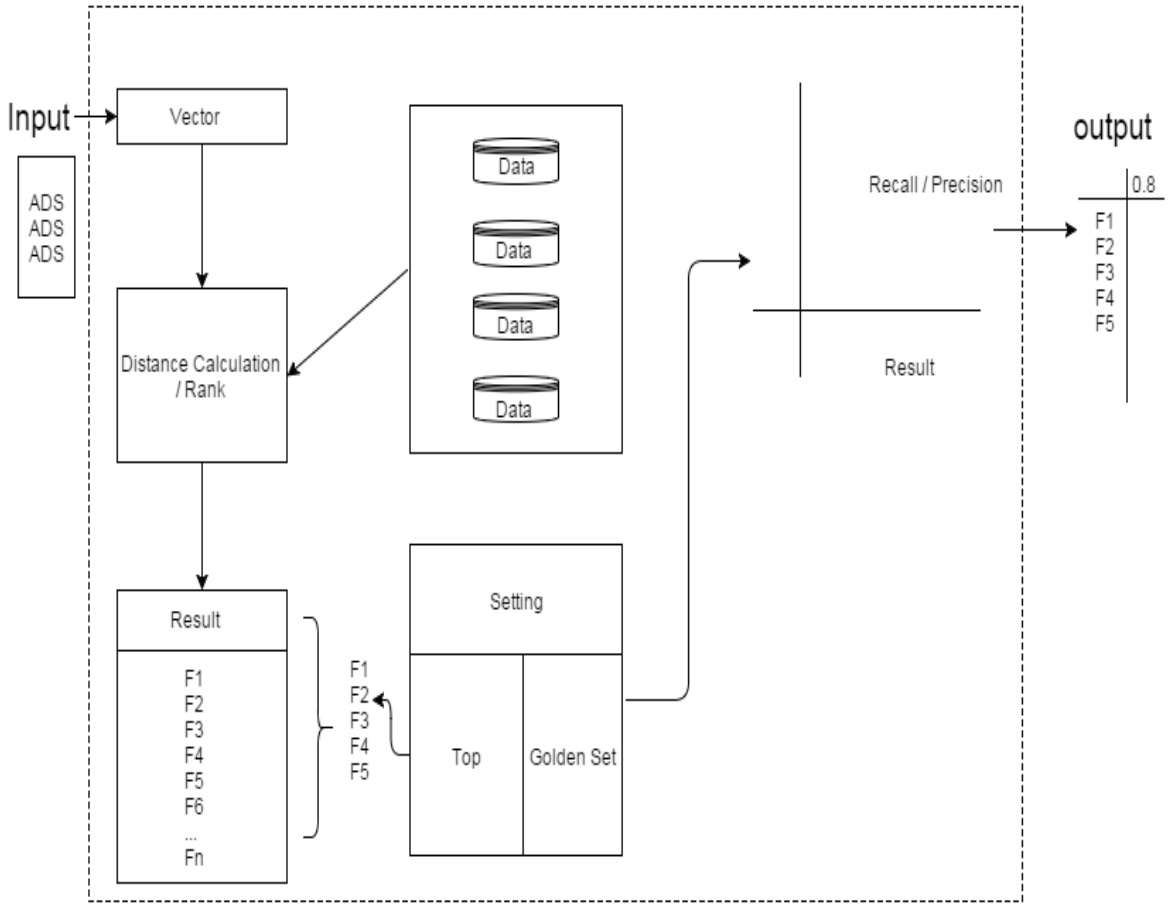
Một hệ thống trích xuất thông tin (IR - Information Retriecal) cần phải cân đối giữa recall và precision, bởi vậy một độ đo khác cũng thường được sử dụng đó là F − measure được xây dựng dựa trên recall và precision:

$$F - measure = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

Precision, Recall và F-measure là các độ đo cơ bản của một tập các tài liệu được trích xuất. Trên thực tế, đôi khi ta không thể sử dụng trực tiếp các độ đo này để so sánh hai danh sách có sắp xếp các tài liệu trả về, bởi chúng không hề quan tâm đến thứ tự nội tại của các tài liệu.

CHƯƠNG 4

HỆ THỐNG ĐỀ NGHỊ



Hình 4.1: Hệ thống Information Retrieval Social Media

4.1/ Các thành phần trong hệ thống Information Retrieval Social Media

Hệ thống xây dựng bao gồm các khối cơ bản sau:

- **Data:** Dùng lưu trữ dữ liệu sau khi đã được crawl và phân loại đưa về dạng vector có thể tính toán được.

Ví dụ: Dữ liệu sẽ được chuyển vector có thể tính toán được

Bảng 4.1: Tập traning set

OID	Name	Iphone	Nokia	Samsung	DT	Điện	Thoại	Xe	Máy	Cũ
1	Mua Bán điện thoại Iphone, Samsung, Nokia	1	1	1	0	1	1	0	0	0
2	Mua bán điện thoại iphone	1	0	0	0	1	1	0	0	0
3	Mua bán điện thoại TP.HCM	0	0	0	0	1	1	0	0	0
4	Mua bán xe máy cũ	0	0	0	0	0	0	1	1	1
5	Mua bán điện thoại samsung cũ	0	0	1	0	1	1	0	0	1
6	Bán DT NOKIA độc	0	1	0	1	0	0	0	0	0
7	Bán điện thoại Samsung giá rẻ	0	0	1	0	1	1	0	0	0
8	Thu mua điện thoại iphone, nokia, samsung cũ	1	1	1	0	1	1	0	0	1
9	Thu mua điện thoại cũ	0	0	0	0	1	1	0	0	1
10	Mua bán xe máy Việt Nam	0	0	0	0	0	0	1	1	0

- **Distance Calculation/Rank:** Dùng để tính toán khoảng cách của dữ liệu đầu vào với dữ liệu trong DB.

Ví dụ:

Dữ liệu input là: Bán điện thoại iphone cũ

Biểu diễn cho vector dữ liệu đầu vào

Bảng 4.2: Biểu diễn cho vector đầu vào

OID	Name	Iphone	Nokia	Samsung	DT	Điện	Thoại	Xe	Máy	Cũ
1	Bán điện thoại iphone cũ	1	0	0	0	1	1	0	0	1

Bảng 4.3: Bảng tính toán khoảng cách

OID	NAME	Rank
1	Mua Bán điện thoại Iphone, Samsung, Nokia	1.732050808
2	Mua bán điện thoại iphone	1
3	Mua bán điện thoại TP.HCM	1.414213562
4	Mua bán xe máy cũ	2.236067977
5	Mua bán điện thoại samsung cũ	1.414213562
6	Bán DT NOKIA độc	2.449489743
7	Bán điện thoại Samsung giá rẻ	1.732050808
8	Thu mua điện thoại iphone, nokia, samsung cũ	1.414213562
9	Thu mua điện thoại cũ	1
10	Mua bán xe máy Việt Nam	2.449489743

Bảng 4.4: Bảng sắp xếp lại sau khi tính toán

OID	NAME	Rank
2	Mua bán điện thoại iphone	1
9	Thu mua điện thoại cũ	1
3	Mua bán điện thoại TP.HCM	1.414214
5	Mua bán điện thoại samsung cũ	1.414214
8	Thu mua điện thoại iphone,nokia, samsung cũ	1.414214
1	Mua Bán điện thoại Iphone, Samsung, Nokia	1.732051
7	Bán điện thoại Samsung giá rẻ	1.732051
4	Mua bán xe máy cũ	2.236068
6	Bán DT NOKIA độc	2.44949
10	Mua bán xe máy Việt Nam	2.44949

- **Setting:** Khối này sau khi tính khoảng cách giữa dữ liệu đầu vào và dữ liệu trong DB và được sắp xếp theo thứ tự tăng dần khoảng cách.

Top: số lượng dữ liệu sau khi tính toán và được sắp xếp giảm dần khoảng cách.

Golden Set: tập dữ liệu người dùng mong muốn có được. Đây cũng là chi phí mà người dùng mong muốn hệ thống xuất ra (bao nhiêu pages hoặc group) và cũng là kết quả dùng để tính toán độ chính xác và độ bao phủ.

Ví dụ:

Sau khi đã sắp xếp xong nếu ta chọn là 5 thì hệ thống đưa ra 5 kết quả có thứ hạng cao nhất

Bảng 4.5: Kết quả sắp xếp sau khi tính toán

OID	NAME	Rank	EXPECTED (Person)
2	Mua bán điện thoại iphone	1	1
9	Thu mua điện thoại cũ	1	1
3	Mua bán điện thoại TP.HCM	1.414214	1
5	Mua bán điện thoại samsung cũ	1.414214	0
8	Thu mua điện thoại iphone,nokia, samsung cũ	1.414214	1

- **Recall/Precision:** Dùng để tính độ chính xác và độ bao phủ từ kết quả hệ thống đưa ra. Sau đó dựa trên kết quả recall và precision hệ thống sẽ ước lượng độ chính xác *Fmeasure* của tập dữ liệu đề xuất cho người dùng.

Theo ví dụ trên:

Score: Kết quả máy đưa ra là OID{F2, F9, F3, F5, F8}

Expected: Người dùng mong muốn {F1, F2, F3, F8, F9}

Theo kết quả trên ta dùng để tính Recall (Độ bao phủ), Precision (Độ chính xác) theo công thức:

$$Recall = \frac{Score \cap Expected}{Expected}$$

$$Recall = \frac{4}{5} = 0.8$$

$$Precision = \frac{Score \cap Expected}{Score}$$

$$Precision = \frac{4}{5} = 0.8$$

Dựa vào giá trị của Recall (Độ bao phủ), Precision (Độ chính xác) ta sẽ tính được F-measure (Độ đo) theo công thức

$$F - measure = 2\frac{Recall * Precision}{Recall + Precision}$$

$$F - measure = 2\frac{0.8 * 0.8}{0.8 + 0.8} = 0.8$$

4.2/ Thiết kế dữ liệu và sử dụng ngôn ngữ lập trình trong hệ thống

4.2.1/ Thiết kế dữ liệu

Dữ liệu của hệ thống: dữ liệu được lưu trữ dưới DB theo cấu trúc riêng để dể dàng cho việc tìm kiếm và lưu trữ thông tin. Dữ liệu càng lớn, càng yêu cầu về vấn dề thiết kế DB sao cho phù hợp cho việc tìm kiếm và lưu trữ. Hiện tại hệ thống đang dùng Mysql để lưu trữ dữ liệu.

Trong phần này các phương pháp trích xuất văn bản cũng được áp dụng, có thể chia các phương pháp ra thành: lựa chọn tài liệu (Document Selection) và sắp xếp tài liệu (Document Ranking).

Phương pháp lựa chọn tài liệu: câu truy vấn được xem như một ràng buộc cụ thể cho việc lựa chọn các tài liệu có liên quan. Một ví dụ điển hình cho phương pháp này đó là mô hình trích xuất boolean (Boolean retrieval model), trong đó mỗi tài liệu được biểu diễn bởi một tập các từ khóa và người sử dụng sẽ cung cấp một biểu thức boolean các từ khóa, chẳng hạn như:

- *mô tô AND của hàng xe máy*
- *trà OR coffee*

Hệ thống trích xuất sẽ nhận một truy vấn dạng boolean như vậy và trả về các tài liệu thỏa mãn biểu thức. Khó khăn đối với phương pháp này đó là việc mô tả thông tin mà người sử dụng cần bằng biểu thức boolean, bởi vậy nó chỉ thường hoạt động tốt khi người sử dụng hiểu rõ về tập tài liệu cũng như có khả năng trình bày rõ ràng câu truy vấn.

Phương pháp sắp xếp tài liệu: Phương pháp sắp xếp tài liệu sử dụng truy vấn để sắp xếp các tài liệu theo thứ tự liên quan. Thực tế cho thấy phương pháp này thích hợp cho việc trích xuất văn bản hơn so với phương pháp lựa chọn tài liệu. Hầu hết các hệ thống trích xuất văn bản (IR) hiện đại đều sử dụng cách này để trả về một danh

sách có sắp xếp các tài liệu tùy theo câu truy vấn của người sử dụng. Những kỹ thuật được dùng trong những phương pháp dạng này cũng rất đa dạng, bao gồm:

- Đại số học
- Logic học
- Thống kê

Vấn đề chính của hướng tiếp cận này đó là làm cách nào để xấp xỉ độ đo liên quan (Degree of Relevant) của một tài liệu dựa vào các từ có sẵn trong tài liệu cũng như trong toàn bộ Dataset. Trong phạm vi của hệ thống này, chúng ta chỉ xem xét một trong những hướng tiếp cận phổ biến nhất hiện nay, đó là mô hình không gian vector (Vector Space Model - VSM).

Ý tưởng chính của VSM được sử dụng trong hệ thống như sau: chúng ta biểu diễn tất cả các tài liệu trong Dataset và câu truy vấn thành các vector trong không gian nhiều chiều tương ứng với tất cả các từ khóa, sau đó sử dụng một độ đo tương tự (Similarity Measure) thích hợp nào đó để tính toán độ tương tự giữa vector truy vấn với các vector tài liệu. Giá trị độ tương tự sẽ được dùng để sắp xếp các tài liệu trả về.

4.2.2/ Mô hình hóa tài liệu

4.2.2.1/ Token hóa

Bước đầu tiên của việc trích xuất văn bản và thiết kế dữ liệu đó là định nghĩa các từ khóa đại diện cho các tài liệu, bước tiền xử lý này thường được gọi là token hóa (tokenization). Để tránh việc xử lý các từ nhiễu, chúngta sẽ áp dụng một danh sách nhiễu cho tập các tài liệu trong Dataset. Danh sách nhiễu là tập các từ được cho rằng không liên quan đến nội dung của tài liệu.

Ví dụ: một, vài, của, cho, với, các ... là các từ nhiễu.

Mặc dù chúng có thể xuất hiện rất thường xuyên trong tài liệu. Ngoài ra, ta có thể thấy rằng một nhóm các từ có thể chia sẻ chung một từ gốc. Do vậy bước tiếp theo chúng ta sẽ định ra các nhóm từ mà trong đó các từ chỉ có sự khác biệt nhỏ về cú pháp.

Ví dụ: danh từ, động từ, tính từ ...

4.2.2.2/ Mô hình hoá tài liệu

Giả sử một Dataset gồm *d* tài liệu và *t* từ khóa, chúng ta có thể mô hình hóa mỗi tài liệu thành một vector *v*. Trong đó không gian *t* chiều <*t*. Tần số từ khóa được định nghĩa là số lần xuất hiện của từ *t* trong tài liệu *d*, được ký hiệu là **Freq(d,t)**. Tiếp theo ta xây dựng ma trận trọng số *TF*(*d, t*) phản ánh độ liên kết của từ t tương ứng với tài liệu d, trong đó:

- Bằng 0: nếu tài liệu đó không chứa từ khóa *t*.
- Khác 0 trong các trường hợp ngược lại.

Ví dụ: ta có thể đơn giản gán giá trị **TF(d,t) = 1** nếu từ *t* xuất hiện trong văn bản *d*, hoặc sử dụng chính giá trị **freq(d,t)**.

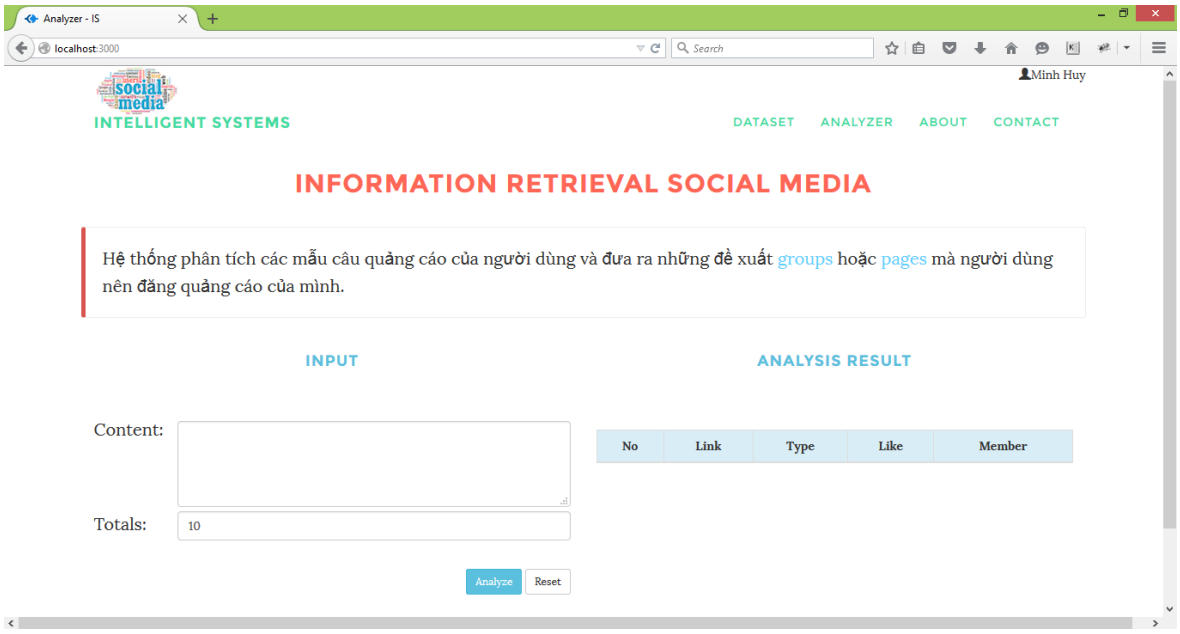
4.2.3/ Ngôn ngữ được sử dụng cho hệ thống

Ngôn ngữ lập trình được sử dụng cho hệ thống bao gồm:

- **Node JS**: ngôn ngữ này dùng cho cả phần backend và frontend (Ajax, jQuery).
- **Jade, CSS**: ngôn ngữ cho phần phát triển giao diện.

CHƯƠNG 5
THỰC NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Hệ thống Information Retrieval Social Media được xây dựng bằng ngôn ngữ Node JS với giao diện như sau:



Hình 5.1: Giao diện chương trình Information Retrieval Social Media

5.1/ Thực nghiệm

Demo với hệ thống Information Retrieval Social Media:

- Đầu tiên, tôi yêu cầu người dùng đưa ra dữ liệu đầu vào và dự kiến những page hoặc group sẽ đăng tin
- Sau đó tôi sẽ chạy demo hệ thống Information Retrieval Social Media với dữ liệu đầu vào người dùng cung cấp. Sau khi hệ thống phân tích sẽ đưa ra dữ liệu đầu ra.
- Tiếp theo tôi sẽ sử dụng các công thức Recall, Precision, F-measure để đo tính hiệu quả của hệ thống
- Hệ thống được chạy trên máy tính Dell N3010 bộ xử lý Core i5 2.2Ghz, 4G bộ nhớ chính, sử dụng hệ điều hành Windows 8.1.

Thực nghiệm 1:

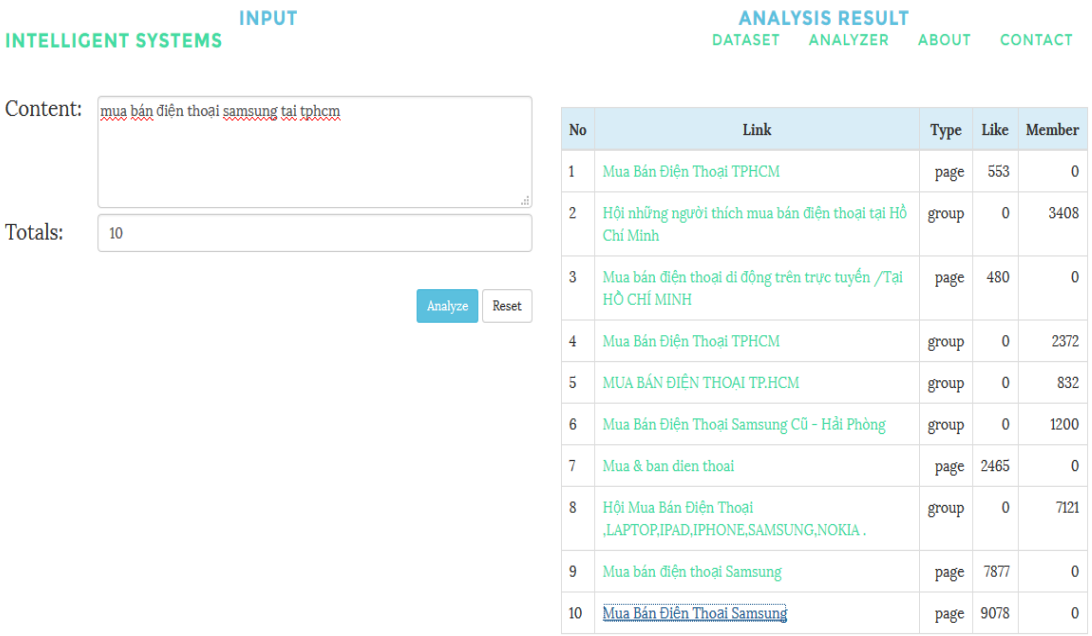
Giả sử: người dùng mong muốn quảng cáo "**Mua bán điện thoại samsung tại tphcm**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.1: Bảng người dùng dự kiến trong thực nghiệm 1

No	Link	Type	Like	Member
1	Mua bán điện thoại TPHCM	Page	553	0
2	Hội những người thích mua điện thoại tại TPHCM	Group	0	3408
3	Mua bán điện thoại di động trực tuyến tài TPHCM	Page	480	0
4	Mua bán điện thoại tại TP.HCM	Group	0	2372
5	Mua & ban dien thoai	Page	2465	0
6	Hội Mua Bán Điện Thoại ,LAPTOP,IPAD,IPHONE,SAMSUNG,NOKIA	Group	0	7121
7	Mua bán điện thoại Samsung	Page	7877	0
8	Mua bán điện thoại tại TPHCM	Group	0	832
9	Mua Bán Điện Thoại Samsung	Page	9078	0
10	Điện Thoại Samsung	Page	1696	0

Với dữ liệu đầu vào là: **Mua bán điện thoại samsung tại tphcm**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



No	Link	Type	Like	Member
1	Mua Bán Điện Thoại TPHCM	page	553	0
2	Hội những người thích mua bán điện thoại tại Hồ Chí Minh	group	0	3408
3	Mua bán điện thoại di động trên trực tuyến /Tại HỒ CHÍ MINH	page	480	0
4	Mua Bán Điện Thoại TPHCM	group	0	2372
5	MUA BÁN ĐIỆN THOẠI TP.HCM	group	0	832
6	Mua Bán Điện Thoại Samsung Cũ - Hải Phòng	group	0	1200
7	Mua & ban dien thoai	page	2465	0
8	Hội Mua Bán Điện Thoại ,LAPTOP,IPAD,IPHONE,SAMSUNG,NOKIA .	group	0	7121
9	Mua bán điện thoại Samsung	page	7877	0
10	Mua Bán Điện Thoại Samsung	page	9078	0

Hình 5.2: Kết quả hệ thống phân tích thí nghiệm 1

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là 9 tin.

$$\text{Precision} = \frac{9}{10} = 0.9$$

$$\text{Recall} = \frac{9}{10} = 0.9$$

$$\text{F} - \text{measure} = 2 * \frac{0.9 * 0.9}{0.9 + 0.9} = 0{,}9$$

Thí nghiệm 2:

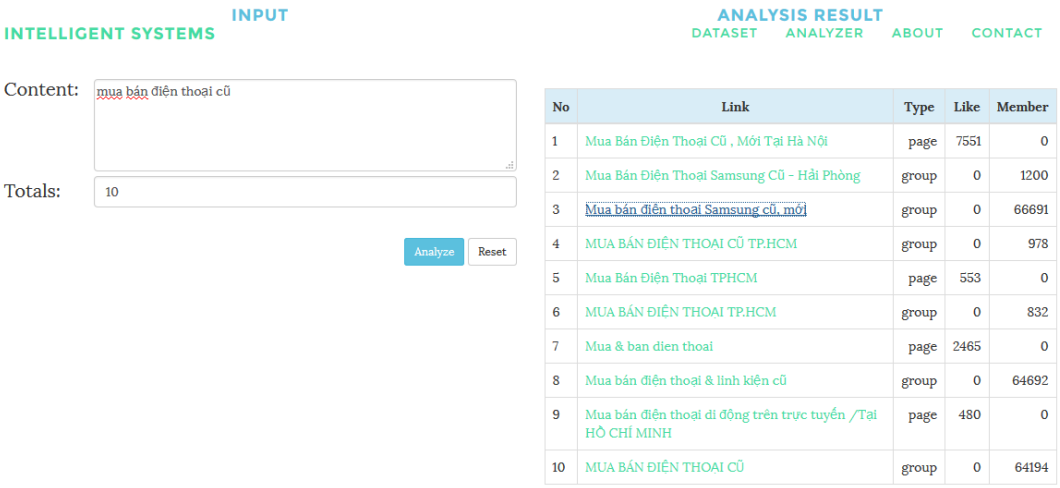
Giả sử: người dùng mong muốn quảng cáo "**Mua bán điện thoại cũ**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.2: Bảng người dùng dự kiến trong thực nghiệm 2

No	Link	Type	Like	Member
1	Mua bán điện thoại cũ mới tại Hà Nội	Page	7551	0
2	Mua bán điện thoại samsung cũ – Hài Phòng	Group	0	1200
3	Bán điện thoại samsung cũ mới	Group	0	908
4	Mua bán điện thoại cũ TP.HCM	Group	0	978
5	Mua bán điện thoại, linh kiện cũ	Group	0	2011
6	Mua bán điện thoại cũ	Group	0	1022
7	Mua bán điện thoại TPHCM	Page	553	0
8	Mua bán điện thoại Iphone cũ	Page	1200	0
9	Mua bán điện thoại trực tiếp	Page	480	0
10	Thu mua điện thoại iphone, samsung, LG	Page	1700	0

Với dữ liệu đầu vào là: **Mua bán điện thoại cũ**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



Hình 5.3: Kết quả hệ thống phân tích thí nghiệm 2

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là 8 tin.

$$\text{Precision} = \frac{8}{10} = 0.8$$

$$\text{Recall} = \frac{8}{10} = 0.8$$

$$\text{F} - \text{measure} = 2 * \frac{0.8 * 0.8}{0.8 + 0.8} = 0{,}8$$

Thí nghiệm 3:

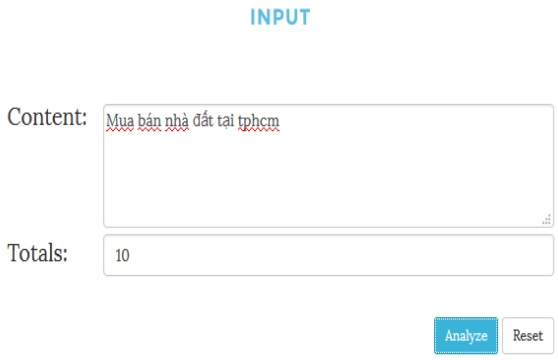
Giả sử: người dùng mong muốn quảng cáo "**Mua bán nhà đất tại tphcm**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.3: Bảng người dùng dự kiến trong thực nghiệm 3

No	Link	Type	Like	Member
1	Hội mua bán nhà đất TP HCM	Page	790	0
2	Cho thuê – Mua bán nhà đất PMH Quận 7 TPHCM	Group	0	1200
3	Mua nhà đất quận 7 TPHCM	Group	0	850
4	Mua bán đất HCM	Page	8873	0
5	Mua bán đất tại thành phố HCM	Page	8590	0
6	Mua bán đất giá rẻ TP HCM	Group	0	350
7	Dịch vụ công chứng mua bán nhà đất TPHCM	Page	970	0
8	Hội mua bán nhà đất quận 7	Group	0	1200
9	Mua bán đất căn hộ giá rẻ	Page	870	0
10	Mua bán nhà phố đất nền tại Hồ Chí Minh	Page	1028	0

Với dữ liệu đầu vào là: **Mua bán nhà đất tại tphcm**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



ANALYSIS RESULT

No	Link	Type	Like	Member
1	Dịch Vụ Công Chứng Mua Bán Nhà Đất TP HCM	page	970	0
2	Hội mua bán nhà đất thành phố Hồ Chí Minh	group	0	280
3	Cho Thuê - Mua Bán Nhà Đất PMH Q7 HCM	group	0	180
4	Mua Bán Nhà Đất Quận 7 - TP HCM	group	0	850
5	Mua bán nhà đất tại TPHCM	page	8590	0
6	Mua bán nhà đất HCM	page	8873	0
7	Mua Bán Nhà Đất Hà Nội	group	0	120
8	Hội mua bán nhà đất Hà Nội	group	0	2836
9	hội mua bán nhà đất q7	group	0	1200
10	Nhà Đất Thổ Cư Khu Vực Hà Nội	group	0	578

Hình 5.4: Kết quả hệ thống phân tích thí nghiệm 3

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là 7 tin.

$$\text{Precision} = \frac{7}{10} = 0.7$$

$$\text{Recall} = \frac{7}{10} = 0.7$$

$$\text{F} - \text{measure} = 2 * \frac{0.7 * 0.7}{0.7 + 0.7} = 0.7$$

Thí nghiệm 4:

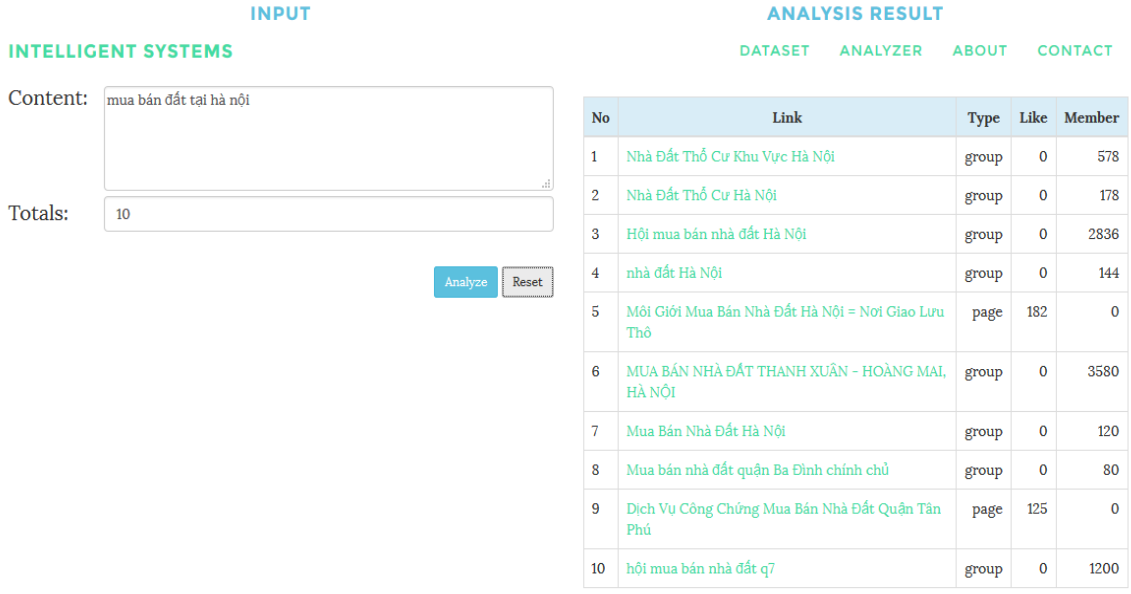
Giả sử: người dùng mong muốn quảng cáo "**Mua bán nhà đất tại hà nội**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.4: Bảng người dùng dự kiến trong thực nghiệm 4

No	Link	Type	Like	Member
1	Nhà đất thổ cư khu vực hà nội	Group	0	578
2	Nhà đất thổ cư Hà Nội	Group	0	178
3	Hội mua bán nhà đất Hà Nội	Group	0	2836
4	Nhà đất Hà Nội	Group	0	144
5	Môi giới mua bán nhà đất Hà Nội	Page	182	0
6	Mua bán nhà đất Thanh Xuân- Hoàng Mai Hà Nội	Group	0	3580
7	Mua bán Nhà Đất Hà Nội	Group	0	120
8	Mua bán nhà đất quận Ba Đình chính chủ	Page	125	0
9	Hội mua bán nhà đất , căn hộ, chung cư tại Hà Nội	Group	0	310
10	Mua bán chung cư giá rẻ tại Hà Nội	Page	1290	0

Với dữ liệu đầu vào là: **Mua bán nhà đất tại hà nội**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



Hình 5.5: Kết quả hệ thống phân tích thí nghiệm 4

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là 8 tin.

$$\text{Precision} = \frac{8}{10} = 0.8$$

$$\text{Recall} = \frac{8}{10} = 08$$

$$\text{F} - \text{measure} = 2 * \frac{0.8 * 0.8}{0.8 + 0.8} = 0{,}8$$

Thí nghiệm 5:

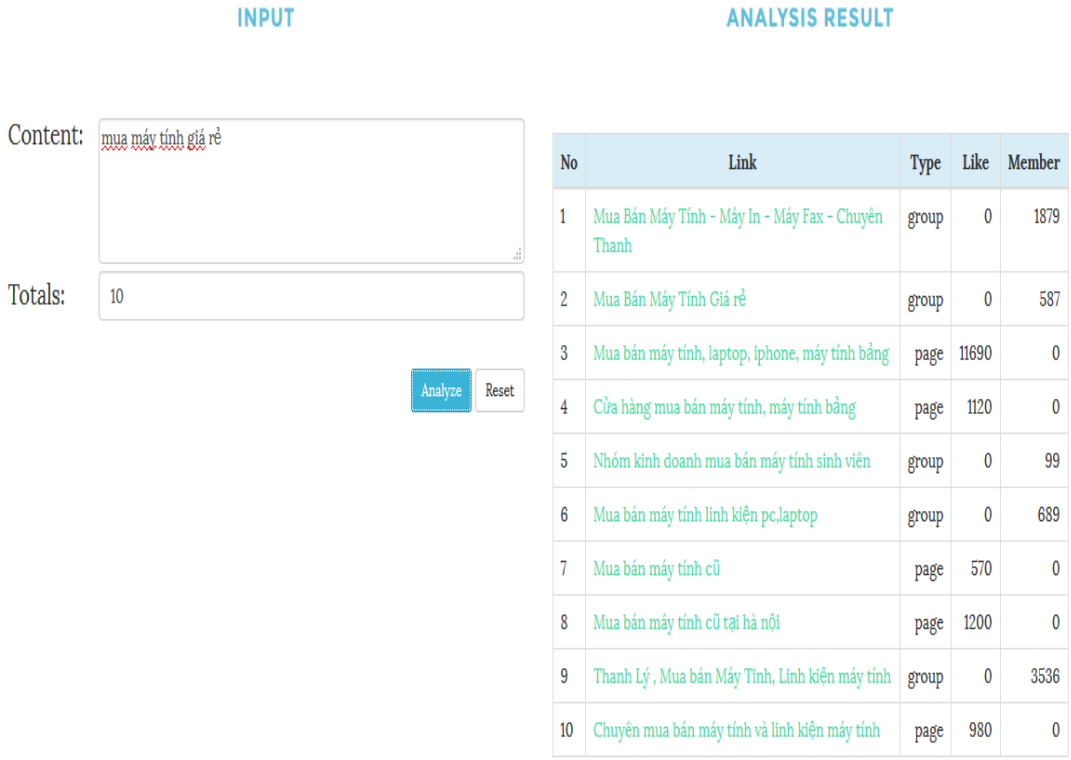
Giả sử: người dùng mong muốn quảng cáo "**Mua máy tính giá rẻ**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.5: Bảng người dùng dự kiến trong thực nghiệm 5

No	Link	Type	Like	Member
1	Mua bán máy tính – máy in – máy fax – chuyên nghiệp	Group	0	1879
2	Mua bán máy tính giá rẻ	Group	0	587
3	Cửa hàng mua bán máy tính, máy tính bảng	Page	1120	0
4	Nhóm kinh doanh mua bán máy tính sinh viên	Group	0	99
5	Mua bán máy tính cũ	Page	570	0
6	Thanh lý mua bán máy tính linh kiện máy tính	Group	0	3536
7	Chuyên mua bán máy tính và linh kiện máy tính	Page	980	
8	Bán laptop xách tay giá rẻ	Page	1542	0
9	Mua bán máy tính cũ tại Hà Nội	Page	1200	0
10	Mua bán máy tính, laptop, iphone, máy tính bảng	Page	11690	0

Với dữ liệu đầu vào là: **Mua máy tính giá rẻ**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



No	Link	Type	Like	Member
1	Mua Bán Máy Tính - Máy In - Máy Fax - Chuyên Thanh	group	0	1879
2	Mua Bán Máy Tính Giá rẻ	group	0	587
3	Mua bán máy tính, laptop, iphone, máy tính bảng	page	11690	0
4	Cửa hàng mua bán máy tính, máy tính bảng	page	1120	0
5	Nhóm kinh doanh mua bán máy tính sinh viên	group	0	99
6	Mua bán máy tính linh kiện pc,laptop	group	0	689
7	Mua bán máy tính cũ	page	570	0
8	Mua bán máy tính cũ tại hà nội	page	1200	0
9	Thanh Lý , Mua bán Máy Tính, Linh kiện máy tính	group	0	3536
10	Chuyên mua bán máy tính và linh kiện máy tính	page	980	0

Hình 5.6: Kết quả hệ thống phân tích thí nghiệm 5

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là 9 tin.

$$\text{Precision} = \frac{9}{10} = 0.9$$

$$\text{Recall} = \frac{9}{10} = 0.9$$

$$\text{F} - \text{measure} = 2 * \frac{0.9 * 0.9}{0.9 + 0.9} = 0{,}9$$

Thí nghiệm 6:

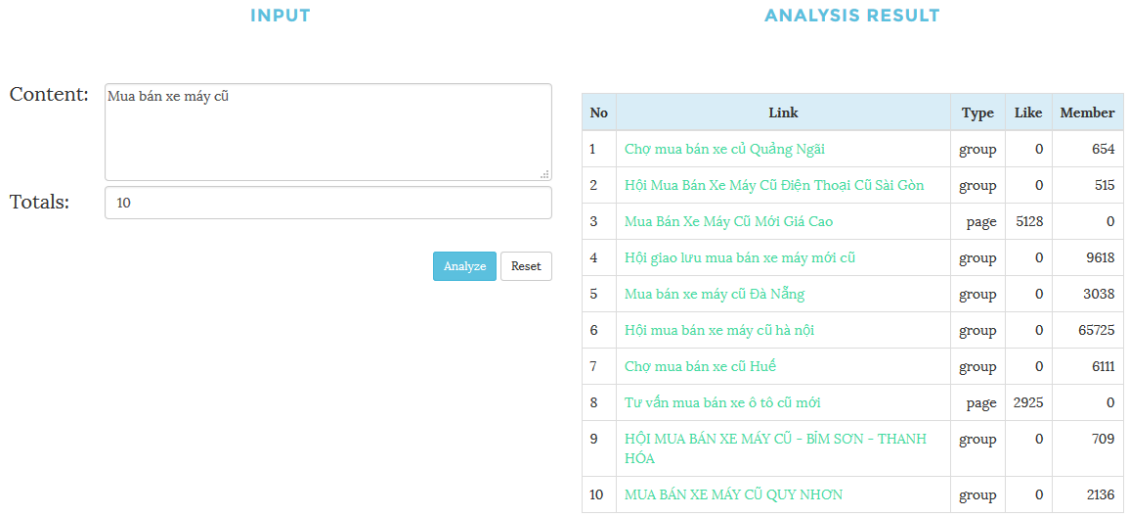
Giả sử: người dùng mong muốn quảng cáo "**Mua bán xe máy cũ**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.6: Bảng người dùng dự kiến trong thực nghiệm 6

No	Link	Type	Like	Member
1	Chợ mua bán xe củ Quảng Ngãi	Group	0	654
2	Hội Mua Bán Xe Máy Cũ Điện Thoại Cũ Sài Gòn	Group	0	515
3	Mua Bán Xe Máy Cũ Mới Giá Cao	Page	5128	0
4	Hội giao lưu mua bán xe máy mới cũ	Group	0	9618
5	Mua bán xe máy cũ Đà Nẵng	Group		3038
6	Hội mua bán xe máy cũ hà nội	Group	0	65725
7	Chợ mua bán xe cũ Huế	Group	0	6111
8	HỘI MUA BÁN XE MÁY CŨ - BỈM SƠN - THANH HÓA	Group	2925	0
9	MUA BÁN XE MÁY CŨ QUY NHƠN	Group	0	709
10	Hội mua bán xe máy củ ở thị xã long khánh đồng nai	Group	0	2136

Với dữ liệu đầu vào là: **Mua bán xe máy cũ**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



Hình 5.7: Kết quả hệ thống phân tích thí nghiệm 6

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là tin.

$$\text{Precision} = \frac{9}{10} = 0.9$$

$$\text{Recall} = \frac{9}{10} = 0.9$$

$$\text{F} - \text{measure} = 2 * \frac{0.9 * 0.9}{0.9 + 0.9} = 0{,}9 = 90\%$$

Thí nghiệm 7:

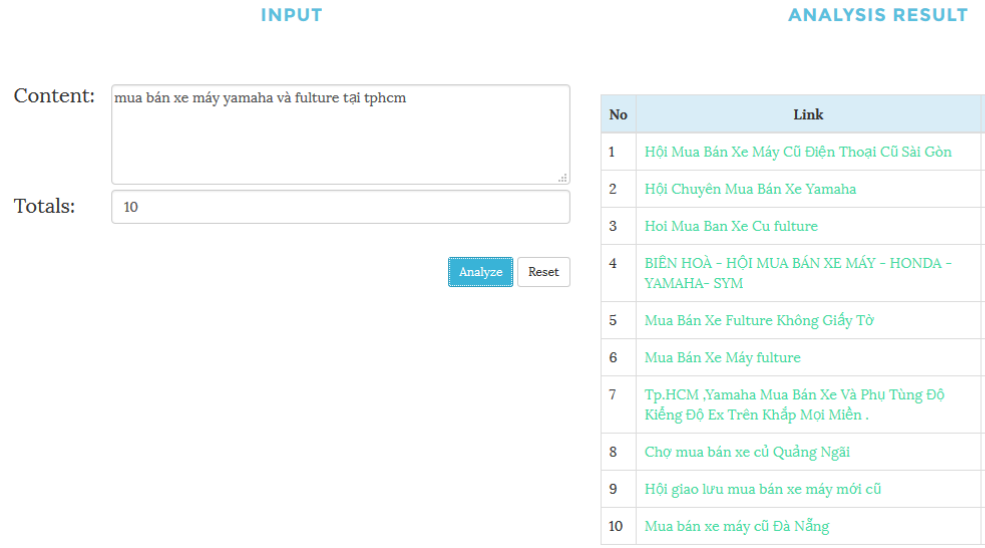
Giả sử: người dùng mong muốn quảng cáo "**mua bán xe máy yamaha và fulture tại tphcm**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.7: Bảng người dùng dự kiến trong thực nghiệm 7

No	Link	Type	Like	Member
1	Hội chung mua bán xe Yamaha	Group	0	5476
2	Hội mua bán xe cũ Fulture	Group	0	290
3	BIÊN HOÀ - HỘI MUA BÁN XE MÁY - HONDA - YAMAHA- SYM	Group	0	1620
4	Mua Bán Xe Fulture Không Giấy Tờ	Group	0	676
5	Mua Bán Xe Máy fulture	Page	1959	8990
6	Tp.HCM ,Yamaha Mua Bán Xe Và Phụ Tùng Độ Kiểng Độ Ex Trên Khắp Mọi Miền	Group	0	410
7	Hội giao lưu xe máy mới cũ	Group	0	9618
8	Hội mua bán xe máy cũ điện thoại cũ sài gòn	Group	0	515
9	Mua xe máy Fultture	Page	1200	0
10	Hội những nhiều chạy Yamaha	Group	0	1479

Với dữ liệu đầu vào là: **mua bán xe máy yamaha và fulture tại tphcm**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



No	Link	Type	Like	Member
1	Hội Mua Bán Xe Máy Cũ Điện Thoại Cũ Sài Gòn	group	0	515
2	Hội Chuyên Mua Bán Xe Yamaha	group	0	5476
3	Hoi Mua Ban Xe Cu fulture	group	0	290
4	BIÊN HOÀ - HỘI MUA BÁN XE MÁY - HONDA - YAMAHA- SYM	group	0	1620
5	Mua Bán Xe Fulture Không Giấy Tờ	group	0	676
6	Mua Bán Xe Máy fulture	page	1959	0
7	Tp.HCM ,Yamaha Mua Bán Xe Và Phụ Tùng Độ Kiểng Độ Ex Trên Khắp Mọi Miền .	group	0	410
8	Chợ mua bán xe củ Quảng Ngãi	group	0	654
9	Hội giao lưu mua bán xe máy mới cũ	group	0	9618
10	Mua bán xe máy cũ Đà Nẵng	group	0	3038

Hình 5.8: Kết quả hệ thống phân tích thí nghiệm 7

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là 8 tin.

$$\text{Precision} = \frac{8}{10} = 0.8$$

$$\text{Recall} = \frac{8}{10} = 0.8$$

$$\text{F} - \text{measure} = 2 * \frac{0.8 * 0.8}{0.8 + 0.8} = 0{,}8$$

Thí nghiệm 8:

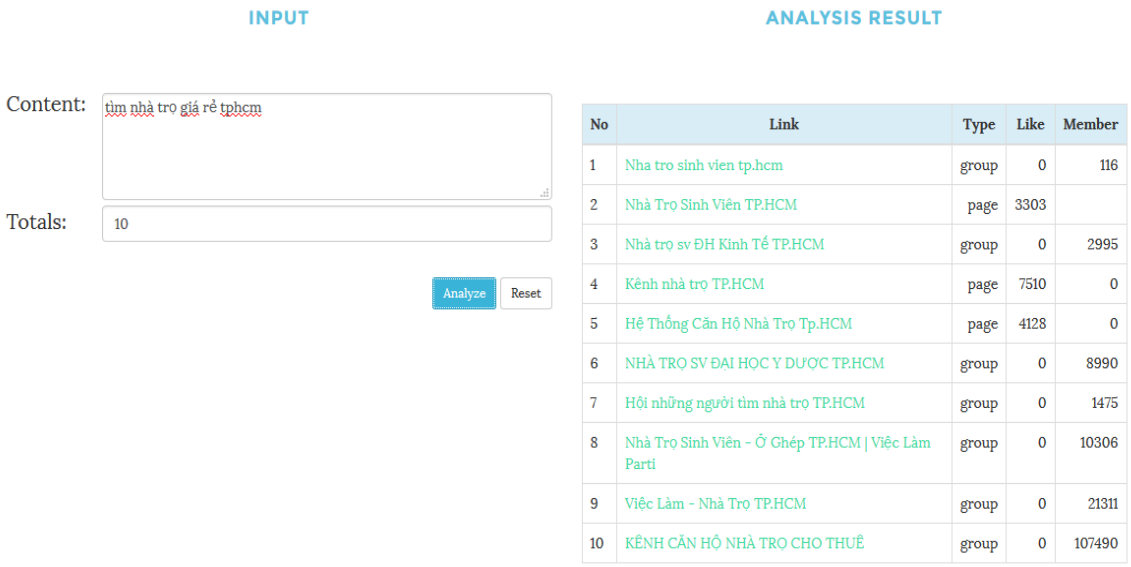
Giả sử: người dùng mong muốn quảng cáo "**Nhà trọ sinh viên giá rẻ tphcm**", người dùng dự kiến sẽ đăng trên page hoặc group của Facebook như bảng sau:

Bảng 5.8: Bảng người dùng dự kiến trong thực nghiệm 8

No	Link	Type	Like	Member
1	Nha tro sinh vien tp.hcm	Group	0	116
2	Nhà Trọ Sinh Viên - Ở Ghép TP.HCM Việc Làm Parti	Group	0	10306
3	Việc làm - Nhà trọ TP.HCm	Group	0	21311
4	Nhà trọ sinh viên TPHCM	Page	3303	0
5	NHÀ TRỌ SV ĐẠI HỌC Y DƯỢC TPHCM	Group	0	8990
6	Nhà trọ SV Đại học Kinh Tế TPHCM	Group	0	2995
7	Hội những người tìm nhà trọ	Group	0	1475
8	Kênh nhà trọ TP.HCM	Page	7510	0
9	Hệ thống Căn hộ Nhà trọ TPHCM	Page	4128	709
10	HỘI SINH VIÊN TÌM NHÀ TRỌ GIỚI THIỆU NHÀ TRỌ	Group	0	29011

Với dữ liệu đầu vào là: **Nhà trọ sinh viên giá rẻ tphcm**

Đưa câu truy vấn vào hệ thống phân tích được kết quả như sau:



Hình 5.9: Kết quả hệ thống phân tích thí nghiệm 8

Đánh giá hiệu quả của hệ thống với câu truy vấn trên:

Người dùng dự kiến sẽ đăng 10 tin trên page hoặc group kết quả so với hệ thống chạy giống với dự kiến người dùm là 9 tin.

$$\text{Precision} = \frac{9}{10} = 0.9$$

$$\text{Recall} = \frac{9}{10} = 0.9$$

$$\text{F} - \text{measure} = 2 * \frac{0.9 * 0.9}{0.9 + 0.9} = 0{,}9$$

5.2/ Đánh giá thí nghiệm

Thách thức và điểm quan trọng của hệ thống quảng cáo thông minh trên mạng xã hội là vấn đề tìm kiếm thông tin, xếp hạng các tài liệu theo thứ tự giảm dần mức độ liên quan với nhu cầu thông tin của người dùng, thường dưới dạng một truy vấn, và loại bỏ những tài liệu không liên quan. Nên khi xây dựng hệ thống này, nó hoạt động gần giống như hệ thống tìm kiếm thông tin.

Để đánh giá hiệu năng của một hệ thống tìm kiếm thông tin người ta sử dụng hai độ đo: độ chính xác (Precision và độ bao phủ (recall) của hệ thống trên một tập dữ liệu.

Với kết quả thu được sau 8 lần chạy thí nghiệm

Câu truy vấn dữ liệu dầu vào	Recall	Precision	F-measure
Mua bán điện thoại samsung tại tphcm	0.9	0.9	0.9
Mua bán điện thoại cũ	0.8	0.8	0.8
Mua bán nhà đất tại tphcm	0.7	0.7	0.7
Mua bán nhà đất tại hà nội	0.8	0.8	0.8
Mua máy tính giá rẻ	0.9	0.9	0.9
Mua bán xe máy cũ	0.9	0.9	0.9
Mua bán xe máy yamaha và fulture tại tphcm	0.9	0.9	0.9
Nhà trọ sinh viên giá rẻ tphcm	0.9	0.9	0.9

CHƯƠNG 6
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

6.1/ Kết luận

Đây là hệ thống phân tích các mẫu quảng cáo giúp người dùng dể dàng tìm kiếm thông tin về các link đến diễn đàn mạng xã hội dựa vào kết quả đưa ra từ hệ thống (gồm các link) từ đó người dùng có thể truy cập và đăng quảng cáo của mình.

Kết quả hiện thực mô hình các thí nghiệm cho thấy bước đầu xây dựng thành công hệ thống quảng cáo thông minh trên các mạng xã hội (Information Retrieval Social Media) bao gồm các chức năng:

- Hệ thống tập dữ liệu có sẵn từ các trang mạng xã hội để làm dữ liệu chính.
- Lọc dữ liệu và phân loại các dữ liệu, sau đó biểu diễn chúng thành các vector có thể tính toán được.
- Hệ thống sẽ tính toán dựa vào dữ liệu đầu vào (biểu diễn thành các vector) với các tập dữ liệu đã có sẵn và đưa ra kết quả mong muốn.

Nhìn chung, có thể thấy rằng các kỹ thuật phân tích tài liệu phức tạp hơn so với kỹ thuật Data Mining truyền thống bởi phải thực hiện trên dữ liệu văn bản vốn ở dạng phi cấu trúc và có tính mờ (Fuzzy). Tuy nhiên, thực tế cho thấy hiện nay người sử dụng vẫn ưa thích và dùng ngày càng nhiều các hệ thống lưu trữ và phân tích dữ liệu ở dạng văn bản. Từ đó ta có thể thấy hệ thống phân tích các mẫu quảng cáo cho người dùng rất khả thi và hệ thống có tính thương mại cao.

6.2/ Hướng nghiên cứu tiếp theo:

Trong tương lai hệ thống sẽ tiếp tục phát triển trên các trang mạng xã hội khác; bao gồm: Flicker, YouTube, Windows Live và đặc biệt hệ thống sẽ phát triển thêm các chức năng khác (phân loại theo từng domain hay categories) để giúp người dùng dể dàng tìm kiếm thông tin cũng như dữ liệu được tốt hơn.

Dù chưa làm thực nghiệm nhưng dựa trên cớ sở lý thuyết đã nghiên cứu được, tôi mạnh dạn đề xuất một số hướng nghiên cứu tiếp theo để có thể kết hợp với “ tên

đề tài" tạo nên một công cụ tối ưu giúp cho các doanh nghiệp phát triển việc quảng cáo thương hiệu của mình trên mạng xã hội.

Các đề xuất như sau:

Tiếp tục xây dựng hệ thống crwal dữ liệu từ các mạng xã hội như là Facebook, Twitter, Google + để có thể luôn cập nhật cơ sở dữ liệu mới nhất.

Phát triển hệ thống tiếp thị liên kết trên mạng xã hội bao gồm các hạng mục nghiên cứu:

- Xây dựng cơ sở dữ liệu gồm các từ khóa và đường link tới sản phẩm, dịch vụ tương ứng, nghiên cứu các thuật toán để hiệu suất tạo ra các từ khóa nhanh, chính xác, phù hợp với người dùng và thực tế, có thể áp dụng được.
- Với một lượng dữ liệu tổng hợp từ các nguồn đem lại có thể sẽ rất lớn và sẽ gây mất nhiều thời gian trong việc load trang website, trong đó, với tùy nội dung của các trang mạng xã hội khác nhau, việc chọn lọc đưa ra các từ khóa phù hợp với nội dung của trang mạng xã hội đó cũng cần được giải quyết, và sau đây là một số giải pháp
- Phân loại các từ khóa thành các lĩnh vực, các mạng xã hội có thể chọn lựa lĩnh vực phù hợp với nội dung của mạng xã hội đang hoạt động.
- Tạo một trang dữ liệu cache, sẽ chỉ lấy một số lượng các từ khóa nhất định từ trong tập cơ sở dữ liệu để tăng tốc độ load website.
- Sắp xếp cơ sở dữ liệu và đưa ra các mức độ ưu tiên của các từ khóa.

TÀI LIỆU THAM KHẢO

1. Ah-Hwee Tan, Text mining: The state of the art and the challenges, Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases 1999/4/26
2. Chris, "The long tail: Why the future of business is selling less of more," 2006.
3. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," Computational linguistics, vol. 22, no. 1, pp. 39–71, 1996.
4. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom, "Expertise identification using email communications," in Proceedings of the twelfth international conference on Information and knowledge management. ACM, 2003, pp.528–531
5. Cooper D., Schindler P., Business research methods, McGraw Hill (2006).
6. Steven White, All things marketing, Sosial media growth from 2006 to 2012 (2013)
7. Dr. M. Saravanakumar , Dr.T.SuganthaLakshmi. Social Media Marketing. Life Science Journal 2012;9(4)
8. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning user interaction models for predicting web search result preferences," in Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006, pp. 3–10.
9. J. Burstein and M. Wolska, "Toward evaluation of writing style: finding overly repetitive word use in student essays," in Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003, pp.35–42.
10. J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder, and M. D. Harris, "Automated scoring using a hybrid feature identification technique," in Proceedings of the 17th international conference on

Computational linguistics-Volume 1. Association for Computational Linguistics, 1998, pp. 206–210.

11. J. Kincaid, R. Fishburn, R. Rogers, and B. Chissom, “Derivation of new readability formulas for navy enlisted personnel (research branch report 8-75),” Memphis, TN: Naval Air Station, Millington, Tennessee, 1975.
12. K. Ali and M. Scarr, “Robust methodologies for modeling web click distributions,” in Proceedings of the 16th international conference on World Wide Web. ACM, 2007, pp. 511–520.
13. P. Jurczyk and E. Agichtein, “Discovering authorities in question answer communities by using link analysis,” in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM, 2007, pp. 919–922
14. Pawel Jurczyk, Eugene Agichtein “Hits on question answer portals: exploration of link analysis for author ranking,” in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007, pp. 845–846.
15. R. Feldman and I. Dagan. Kdt - knowledge discovery in texts. In Proc.of the First Int. Conf. on Knowledge Discovery (KDD), pages 112–117,1995
16. Richard M. Schwartz, David R. H. Miller, Timothy R. Leek, Information retrieval system, US7162468 B2 (2007/01/09)
17. T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, “Accurately interpreting clickthrough data as implicit feedback,” in Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005, pp. 154–161.
18. Tie-Yan Liu, Learning to Rank for Information Retrieval, Journal Foundations and Trends in Information Retrieval, Volume 3 Issue 3, March 2009 Pages 225-331

19. Xuerui Wang, Topical N-Grams: Phrase and Topic Discovery, with an Application to Information Retrieval, Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on Page(s): 697 – 702
20. Y. Attali and J. Burstein, “Automated essay scoring with e-rater R v. 2,” The Journal of Technology, Learning and Assessment, vol. 4, no. 3, 2006.
21. http://wearesocial.net/blog/2015/01/digital-social-mobile-worldwide-2015/