**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KỸ THUẬT PHẦN MỀM**

**HỨA PHƯỚC TRƯỜNG**

**KHÓA LUẬN TỐT NGHIỆP**

**HỆ THỐNG LẮNG NGHE MẠNG XÃ HỘI - ZIE**

**KỸ SƯ NGÀNH KỸ THUẬT PHẦN MỀM**

**TP. HỒ CHÍ MINH, 2013**

**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA KỸ THUẬT PHẦN MỀM**

**HỨA PHƯỚC TRƯỜNG – 09520392**

**KHÓA LUẬN TỐT NGHIỆP**

**HỆ THỐNG LẮNG NGHE MẠNG XÃ HỘI - ZIE**

**KỸ SƯ KỸ THUẬT PHẦN MỀM**

**GIẢNG VIÊN HƯỚNG DẪN**

**THẠC SĨ NGUYỄN ĐĂNG KHOA**

**TP. HỒ CHÍ MINH, 2013**

DANH SÁCH HỘI ĐỒNG BẢO VỆ KHÓA LUẬN

Hội đồng chấm khóa luận tốt nghiệp, thành lập theo Quyết định số …………………… ngày ………………….. của Hiệu trưởng Trường Đại học Công nghệ Thông tin.

* 1. …………………………………………. – Chủ tịch.
  2. …………………………………………. – Thư ký.
  3. …………………………………………. – Ủy viên.
  4. …………………………………………. – Ủy viên.

**MỤC LỤC**

[TÓM TẮT LUẬN VĂN 1](#_Toc378095702)

[MỞ ĐẦU 2](#_Toc378095703)

[Chương 1. TỔNG QUAN 3](#_Toc378095704)

[1.1. Mạng xã hội 3](#_Toc378095705)

[1.2. Sức mạnh của mạng xã hội 3](#_Toc378095706)

[1.3. Mạng xã hội tại Việt Nam 5](#_Toc378095707)

[1.4. Bối cảnh nghiên cứu 6](#_Toc378095708)

[1.5. Động lực nghiên cứu 6](#_Toc378095709)

[1.6. Mục tiêu đề tài 7](#_Toc378095710)

[Chương 2. NHỮNG SẢN PHẨM TƯƠNG TỰ 8](#_Toc378095711)

[2.1. NM InCite 8](#_Toc378095712)

[2.2. Ubervu 9](#_Toc378095713)

[2.3. Sysomos 10](#_Toc378095714)

[2.4. Noti5 11](#_Toc378095715)

[Chương 3. PHÂN TÍCH CHỨC NĂNG 12](#_Toc378095716)

[3.1. Tổng quan chức năng 12](#_Toc378095717)

[3.1.1. Hệ thống thu thập 12](#_Toc378095718)

[3.1.2. Hệ thống báo cáo 12](#_Toc378095719)

[3.2. Sơ đồ Use case 13](#_Toc378095720)

[3.2.1. Use case Crawler 13](#_Toc378095721)

[3.2.2. Use case Client 14](#_Toc378095722)

[3.2.3. Use case Save post 15](#_Toc378095723)

[3.2.4. Use case Push url 17](#_Toc378095724)

[3.2.5. Use case Pop url 18](#_Toc378095725)

[3.2.6. Use case Posts 19](#_Toc378095726)

[3.2.7. Use case Trend 20](#_Toc378095727)

[3.2.8. Use case Sources 21](#_Toc378095728)

[3.2.9. Use case Threads 22](#_Toc378095729)

[3.2.10. Use case SMT (Social Media Type) 23](#_Toc378095730)

[3.3. Sơ đồ tuần tự 24](#_Toc378095731)

[3.3.1. Sơ đồ tuần tự Use case Pop url 24](#_Toc378095732)

[3.3.2. Sơ đồ tuần tự Use case Push url 25](#_Toc378095733)

[3.3.3. Sơ đồ tuần tự Use case Save post 25](#_Toc378095734)

[3.3.4. Sơ đồ tuần tự Use case Posts 26](#_Toc378095735)

[3.3.5. Sơ đồ tuần tự Use case Trend 26](#_Toc378095736)

[3.3.6. Sơ đồ tuần tự Use case Sources 27](#_Toc378095737)

[3.3.7. Sơ đồ tuần tự Use case Threads 27](#_Toc378095738)

[3.3.8. Sơ đồ tuần tự Use case SMT 28](#_Toc378095739)

[3.4. Sơ đồ hoạt động 29](#_Toc378095740)

[3.4.1. Zie Stack 29](#_Toc378095741)

[3.4.2. Zie Data 30](#_Toc378095742)

[3.4.3. Sơ đồ hoạt động Zie Reporter 31](#_Toc378095743)

[3.4.4. Sơ đồ hoạt động Zie Crawler 32](#_Toc378095744)

[Chương 4. HƯỚNG TIẾP CẬN 33](#_Toc378095745)

[4.1. Một hệ thống duy nhất vừa đảm nhiệm lưu trữ, vừa đảm nhiệm truy xuất 33](#_Toc378095746)

[4.2. Hai hệ thống, một đảm nhiệm lưu trữ, một đảm nhiệm truy xuất 34](#_Toc378095747)

[4.3. Đánh giá liên kết dựa trên liên kết đến chính nó 35](#_Toc378095748)

[4.4. Đánh giá liên kết dựa trên độ mới của nội dung 36](#_Toc378095749)

[4.5. Thu thập dữ liệu theo chiều sâu 36](#_Toc378095750)

[4.6. Thu thập dữ liệu theo chiều rộng 37](#_Toc378095751)

[4.7. Kết luận 38](#_Toc378095752)

[Chương 5. NHỮNG VẤN ĐỀ GẶP PHẢI KHI HIỆN THỰC HỆ THỐNG ZIE 39](#_Toc378095753)

[5.1.1. Lấy dữ liệu từ Source 39](#_Toc378095754)

[5.1.2. Nhận diện Post 40](#_Toc378095755)

[5.1.3. Nhận diện ngôn ngữ 41](#_Toc378095756)

[5.1.4. Tính tương thích của hệ thống 43](#_Toc378095757)

[5.1.5. Xử lý dữ liệu lớn 44](#_Toc378095758)

[5.1.6. Nhiều Source dẫn đến nhiều Crawler 45](#_Toc378095759)

[Chương 6. THIẾT KẾ HỆ THỐNG ZIE 47](#_Toc378095760)

[6.1. Zie 47](#_Toc378095761)

[6.1.1. Zie Stack 47](#_Toc378095762)

[6.1.2. Zie Crawler 49](#_Toc378095763)

[6.1.3. Zie Data 50](#_Toc378095764)

[6.1.4. Zie Reporter 51](#_Toc378095765)

[6.1.5. Tổng quan về Zie 51](#_Toc378095766)

[6.2. Crawler 52](#_Toc378095767)

[6.3. Front-end 53](#_Toc378095768)

[6.4. Thiết kế dữ liệu 53](#_Toc378095769)

[6.4.1. Table authors 53](#_Toc378095770)

[6.4.2. Table contents 53](#_Toc378095771)

[6.4.3. Table languages 54](#_Toc378095772)

[6.4.4. Table posts 55](#_Toc378095773)

[6.4.5. Table sources 55](#_Toc378095774)

[6.4.6. Table stacks 57](#_Toc378095775)

[6.4.7. Table threads 58](#_Toc378095776)

[6.4.8. Table urls 58](#_Toc378095777)

[6.5. Kỹ thuật phân chia dữ liệu 59](#_Toc378095778)

[6.6. Sơ đồ quan hệ 60](#_Toc378095779)

[Chương 7. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 61](#_Toc378095780)

[7.1. Kết luận 61](#_Toc378095781)

[7.2. Hướng phát triển 63](#_Toc378095782)

[Chương 8. CÀI ĐẶT MINH HỌA 64](#_Toc378095783)

[8.1. Zie Stack 64](#_Toc378095784)

[8.2. Zie Data 66](#_Toc378095785)

[8.3. Zie Crawler 67](#_Toc378095786)

[TÀI LIỆU THAM KHẢO 69](#_Toc378095787)

[PHỤ LỤC 1: POSTGRESQL 70](#_Toc378095788)

[PHỤ LỤC 2: COMPACT LANGUAGE DETECTOR 71](#_Toc378095789)

[PHỤ LỤC 3: PQ 71](#_Toc378095790)

[PHỤ LỤC 4: SPHINX CLIENT 71](#_Toc378095791)

[PHỤ LỤC 5: ICU 72](#_Toc378095792)

[PHỤ LỤC 6: CHUẨN GIAO TIẾP TRONG ZIE 72](#_Toc378095793)

**DANH MỤC HÌNH ẢNH**

[Hình 1‑1 - Tốc độ phát triển của mạng xã hội 2012 3](#_Toc378125321)

[Hình 1‑2 - Mạng xã hội phát tán theo cấp số nhân (Trích từ slide giới thiệu sản phẩm của WebM) 4](#_Toc378125322)

[Hình 1‑3 - Những hoạt động thông thường của người dùng mạng xã hội (trích từ báo cáo thường niên của WebM Vietnam 2013) 5](#_Toc378125323)

[Hình 1‑4 - Sức lan tỏa cửa mạng xã hội 6](file:///C:\Users\Truong\Desktop\KLTN\report.docx#_Toc378125324)

[Hình 3‑1 – Use case Crawler 13](#_Toc378125325)

[Hình 3‑2 - Use case Client 14](#_Toc378125326)

[Hình 3‑3 - Sơ đồ tuần tự Use case Pop url 24](#_Toc378125327)

[Hình 3‑4 - Sơ đồ tuần tự Use case Push url 25](#_Toc378125328)

[Hình 3‑5 - Sơ đồ tuần tự Use case Save Post 25](#_Toc378125329)

[Hình 3‑6 - Sơ đồ tuần tự Use case Post 26](#_Toc378125330)

[Hình 3‑7 - Sơ đồ tuần tự Use case Trend 26](#_Toc378125331)

[Hình 3‑8 - Sơ đồ tuần tự Use case Sources 27](#_Toc378125332)

[Hình 3‑9 - Sơ đồ tuần tự Use case Threads 27](#_Toc378125333)

[Hình 3‑10 - Sơ đồ tuần tự Use case SMT 28](#_Toc378125334)

[Hình 3‑11 - Sơ đồ hoạt động Zie Stack 29](#_Toc378125335)

[Hình 3‑12 - Sơ đồ hoạt động Zie Data 30](#_Toc378125336)

[Hình 3‑13 - Sơ đồ hoạt động Zie Reporter 31](#_Toc378125337)

[Hình 3‑14 - Sơ đồ hoạt động Zie Crawler 32](#_Toc378125338)

[Hình 4‑1 - Sơ đồ mô tả hai hệ thống tách rời (Sphinx + Postgres) 35](#_Toc378125339)

[Hình 5‑1 - Chuẩn canonical meta tag 41](#_Toc378125340)

[Hình 5‑2 - Người dùng viết lẫn lộn các ngôn ngữ 42](#_Toc378125341)

[Hình 5‑3 - Một phần bảng mã character entity từ w3.org 43](#_Toc378125342)

[Hình 5‑4 - Một tường lửa đơn giản để nhận diện người dùng 46](#_Toc378125343)

[Hình 6‑1 - Mô hình đơn giản Zie 47](#_Toc378125344)

[Hình 6‑2 - Hướng di chuyển của người dùng qua liên kết 48](#_Toc378125345)

[Hình 6‑3 - Quy trình hoạt động của Zie Stack 49](#_Toc378125346)

[Hình 6‑4 - Sơ đồ quan hệ đơn giản giữa các thành phần trong Zie 51](#_Toc378125347)

[Hình 6‑5 – Một Crawler được viết bằng API của Zie 52](#_Toc378125348)

[Hình 6‑6 - Sơ đồ quan hệ 60](#_Toc378125349)

[Hình 7‑1 - Mức độ quan tâm của người dùng với 3 sản phẩm Iphone 5S, Galaxy S4 và LG G2. 62](#_Toc378125350)

[Hình 7‑2 - Mức độ quan tâm của người dùng internet đến phim tết 2013 62](#_Toc378125351)

[Hình 8‑1 - Khởi động Stack 64](#_Toc378125352)

[Hình 8‑2 - Pop Zie Stack 65](#_Toc378125353)

[Hình 8‑3 - Cơ chế Log trong Zie Stack 65](#_Toc378125354)

[Hình 8‑4 - Khởi động Zie Data 66](#_Toc378125355)

[Hình 8‑5 - Lệnh save của Zie Data 67](#_Toc378125356)

[Hình 8‑6 - Zie Crawler thông báo không tìm thấy Stack và Data 68](#_Toc378125357)

[Hình 8‑7 - Crawler chạy thành công 68](#_Toc378125358)

# 

**DANH MỤC BẢNG**

[Bảng 8‑2 – Cấu trúc table authors 53](#_Toc378095693)

[Bảng 8‑3 - Cấu trúc table contents 54](#_Toc378095694)

[Bảng 8‑4 - Cấu trúc table languages 54](#_Toc378095695)

[Bảng 8‑5 - Cấu trúc table posts 55](#_Toc378095696)

[Bảng 8‑6 - Cấu trúc table sources 56](#_Toc378095697)

[Bảng 8‑7 - Cấu trúc table stacks 57](#_Toc378095698)

[Bảng 8‑8 - Cấu trúc table threads 58](#_Toc378095699)

[Bảng 8‑9 - Cấu trúc table urls 59](#_Toc378095700)

# TÓM TẮT LUẬN VĂN

Với sự phát triển vượt bậc của mạng xã hội trong những năm gần đây, 81% người dùng Internet sử dụng mạng xã hội để liên lạc, kết nối với bạn bè. Mạng xã hội dần trở thành một món ăn tinh thần không thể thiếu đối với người dùng internet. Bên cạnh đó, mạng xã hội còn là một kênh chứa đựng rất nhiều những thông tin, ý kiến, nhận xét về một thương hiệu, một sản phẩm hay dịch vụ nào đó. Theo thống kê năm 2013 của WebM có 26% người dùng sử dụng mạng xã hội đề cập đến một thương hiệu, sản phẩm hay một dịch vụ. 70% người dùng tin tưởng vào những nhận xét, giới thiệu từ internet chỉ đứng sau duy nhất con số 92% người dùng tin vào những lời nhận xét từ người quen, bạn bè. Trong khi đó, con số tương ứng cho các quảng cáo trên TV là 40%. Từ những số liệu hấp dẫn trên, các chiến dịch quảng bá sản phẩm, theo dõi ý kiến người dùng trên mạng xã hội trở thành một vấn đề rất đáng quan tâm đối với các doanh nghiệp lớn và thậm chí những doanh nghiệp nhỏ đang chọn internet làm điểm đến của họ. Trong bài báo cáo này, em xin đề xuất một hệ thống lắng nghe mạng xã hội, thu thập các bài viết, ý kiến của người dùng sao cho đầy đủ nhất, giá trị nhất và đưa ra những dữ liệu hữu ích cho các nhà nghiên cứu.

# MỞ ĐẦU

Thị trường luôn thay đổi và doanh nghiệp phải luôn theo sát những thay đổi đó để có những chiến lược, chiến thuật phù hợp, nhanh nhạy. Phân tích thị trường là một trong những lý do đầu tiên và quan trọng nhất để lập kế hoạch kinh doanh. Bất luận việc một doanh nghiệp mới bắt đầu hoạt động kinh doanh hay xem xét lại hoạt động kinh doanh hiện tại đều cần phải có phân tích mới về thị trường.



Thị trường doanh nghiệp cần tìm kiếm là thị trường tiềm năng, chứ không phải là thị trường hiện tại. Thị trường mục tiêu của một doanh nghiệp luôn lớn hơn rất nhiều so với số người mà các nhân viên trong công ty tiếp cận được. Đó là những người mà một ngày nào đó doanh nghiệp có thể tiếp cận hoặc họ tự tìm đến với doanh nghiệp. Chính vì vậy, yêu cầu thực tiễn đã đặt ra một bài toán: có cách nào hỗ trợ doanh nghiệp phân tích, theo dõi thị trường bằng sức mạnh của công nghệ.

# TỔNG QUAN

## Mạng xã hội

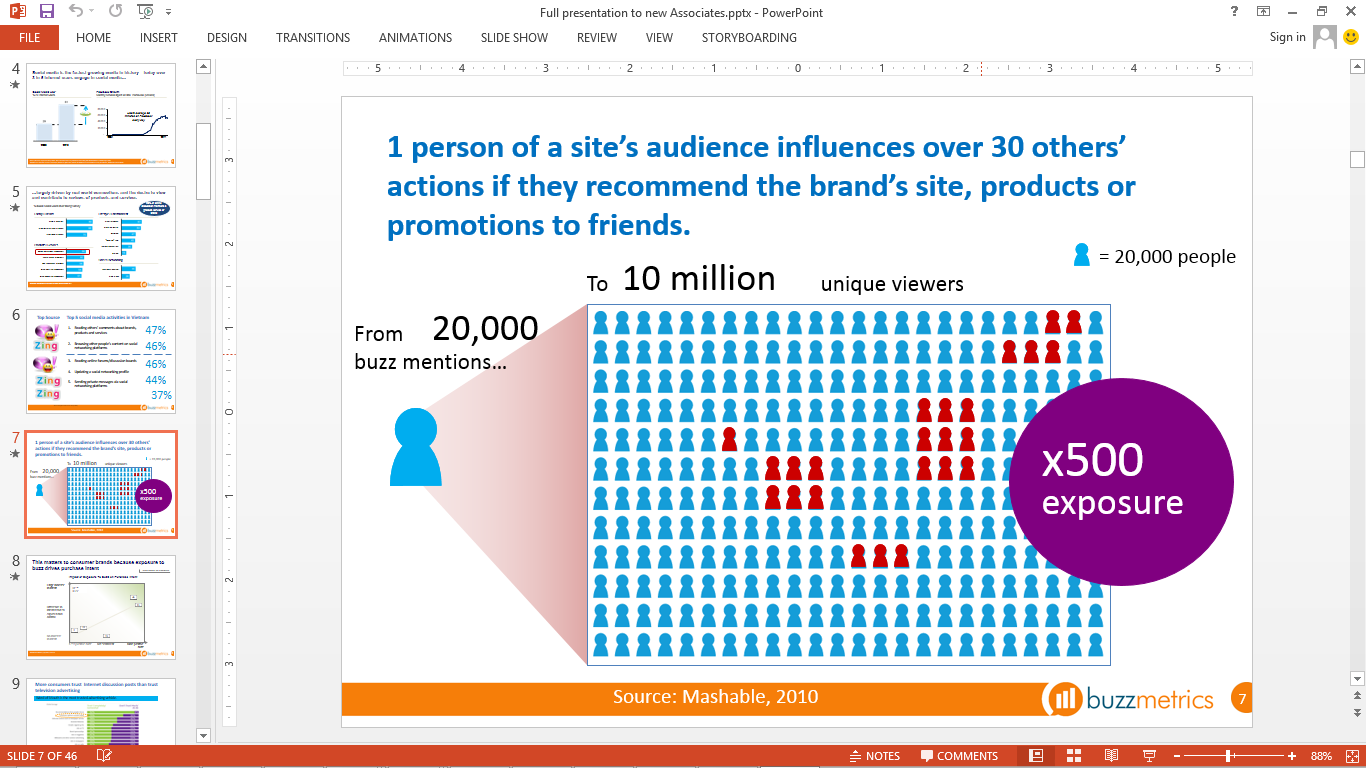
Mạng xã hội, hay gọi là mạng xã hội ảo, là dịch vụ nối kết các thành viên cùng sở thích trên Internet lại với nhau với nhiều mục đích khác nhau mà không phân biệt không gian và thời gian.

Mạng xã hội xuất hiện lần đầu tiên năm 1995 với sự ra đời của trang Classmate với mục đích kết nối bạn học. Đến nay, mạng xã hội liên tiếp phát triển với tốc độ chóng mặt và mang lại rất nhiều lợi ích cho người dùng internet.

Hình ‑ - Tốc độ phát triển của mạng xã hội 2012

## Sức mạnh của mạng xã hội

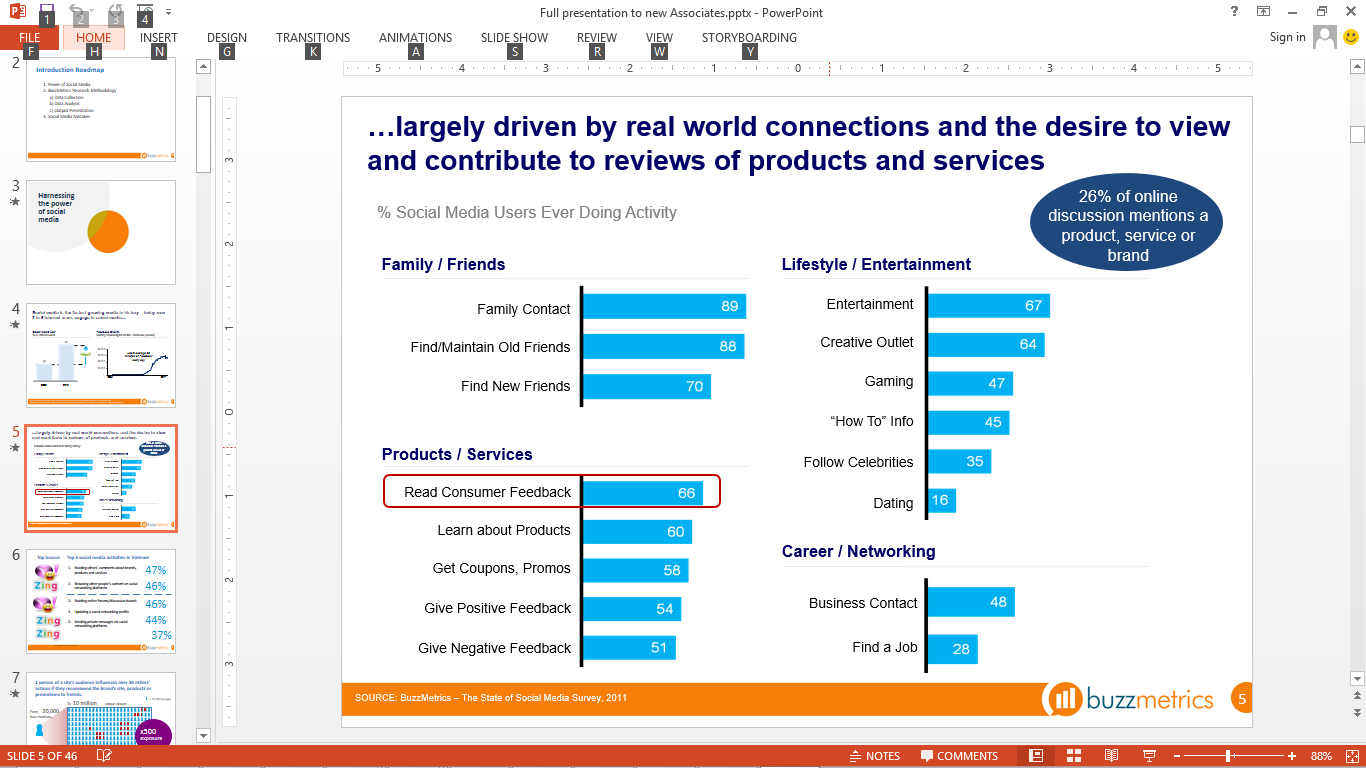
Cùng với sự phát triển của mạng xã hội, càng ngày mạng xã hội càng trở nên là một nơi lý tưởng để mọi người có thể kết nối với nhau, chia sẽ thông tin. Ngoài ra, bên cạnh những tính năng riêng tư, mạng xã hội còn là một công cụ mạnh mẽ phản ảnh thị trường, thậm chí là chính trị, mọi thứ trên mạng xã hội có khả năng lan truyền với tốc độ nhanh đến chóng mặt.



Hình ‑ - Mạng xã hội phát tán theo cấp số nhân  
(Trích từ slide giới thiệu sản phẩm của WebM)

Ngày nay, khi quảng cáo trên các kênh truyền thống không còn mang lại hiểu quả cao, hơn 70% người dùng TV nói họ sẽ bỏ đi hoặc chuyển kênh ngay lập tức khi nhìn thấy quảng cáo.

Theo chân sự phát triển của mạng xã hội, các doanh nghiệp ngày càng chú ý đến vị thế của họ trên internet và Zie là hệ thống giúp cho họ biết được đâu là điểm đứng của họ và đâu là hướng phát triển đúng đắn.



Hình ‑ - Những hoạt động thông thường của người dùng mạng xã hội  
(trích từ báo cáo thường niên của WebM Vietnam 2013)

## Mạng xã hội tại Việt Nam



















## Bối cảnh nghiên cứu

Ngày nay, với sự bùng nổ của mạng xã hội, dựa trên nghiên cứu của Nielsen – công ty nghiên cứu hàng đầu Việt Nam và thế giới, cứ 5 người dùng internet sẽ có 3 người sử dụng mạng xã hội. Từ năm 2008 đến năm 2010, số người dùng mạng xã hội tăng 110%, mỗi ngày trung bình người dùng facebook sẽ sử dụng 55 phút để truy cập mạng xã hội. Sự phát triển nhanh chóng đó khiến mạng xã hội trở thành thị trường thứ hai thu hút sự quan tâm của các nhà đầu tư, các chuyên viên phân tích thị trường.

Theo thống kê của công ty phân tích thị trường WebM tại Việt Nam năm 2013, 66% người dùng sử dụng mạng xã hội để xem nhận xét, đánh giá về một sản phẩm, 26% thảo luận trực tuyến trên mạng xã hội đề cập tới một sản phẩm hoặc thương hiệu. Một người dùng có tầm ảnh hưởng trên mạng xã hội có thể gây ảnh hưởng đến quyết định của ít nhất 30 người dùng khác nếu họ đề nghị về một sản phẩm hoặc thương hiệu.

Hình ‑ - Sức lan tỏa cửa mạng xã hội

## Động lực nghiên cứu

Để thực hiện một nghiên cứu, khi chưa có hệ thống, các nhà nghiên cứu phải sử dụng rất nhiều nhân lực để làm công việc thu thập dữ liệu. Việc thu thập dữ liệu bằng sức người có chi phí rất lớn, độ phủ không cao, thường chỉ có thể tập trung vào những nguồn dữ liệu lớn và chiếm nhiều thời gian. Hơn nữa, sau công đoạn thu thập dữ liệu sẽ có thể dẫn đến hàng loạt lỗi nhập liệu dẫn đến sai sót ở kết quả. Ví dụ: nhóm nghiên cứu sẽ phải duyệt khắp các bài viết trong các website có số lượng người dùng cao (tinhte.vn, hdvietnam.com, vnexpress.net…) trong suốt quãng thời gian cần thống kê, có thể là một tuần, một tháng, hoặc một năm. Nhóm nghiên cứu sẽ phải nhập liệu các con số như ngày viết bình luận, số lượt người xem, số lượt yêu thích, tên tác giả… và khả năng cập nhật ngay lập tức gần như là không có.

Chính vì thế, một hệ thống thu thập thay thế con người với độ phủ rộng, độ chính xác cao, có thể cập nhật liên tục và khả năng thất lạc ở mức chấp nhận được là thứ rất cần thiết.

## Mục tiêu đề tài

Đề tài xây dựng hệ thống lắng nghe mạng xã hội Zie sử dụng sức mạnh công nghệ để thu thập những bình luận, ý kiến của người dùng về một sản phẩm thương hiệu hoặc bất kỳ vấn đề nào có thể theo dõi thông qua internet và mạng xã hội. Từ đó, hệ thống hỗ trợ những công cụ trích lọc, sắp xếp và thể hiện dữ liệu một cách trực quan phục vụ cho các nhà nghiên cứu thị trường.

# NHỮNG SẢN PHẨM TƯƠNG TỰ

Hiện nay, trên thế giới và ở Việt Nam đã có những sản phẩm có mục đích tương tự, mỗi sản phẩm đều có những điểm mạnh đáng để học tập và những điểm yếu cần phải khắc phục:

## NM InCite



|  |  |
| --- | --- |
| Địa chỉ | <http://nmincite.com> |
| Mô tả | NM InCite là hệ thống theo dõi mạng xã hội được sử dụng bởi công ty nghiên cứu thị trường Nielsen[[1]](#footnote-1). |
| Điểm mạnh | Có độ phủ lớn nhất trong các sản phẩm được đề tài đề cập tới.  Có giao diện chức năng rất phức tạp.  Hỗ trợ nhiều loại báo cáo. |
| Điểm yếu | Khả năng thu thập không thực sự tốt.  Dữ liệu thu thập rộng nhưng thiếu chiều sâu.  Có rất nhiều thiếu sót. |
| Số lượng website bao phủ | Trên 500 website tại Việt Nam. |
| Mục tiêu thị trường | Cả thế giới. |
| Hoạt động | Đã đóng cửa từ tháng 3 năm 2013 |

## Ubervu



|  |  |
| --- | --- |
| Địa chỉ | <https://www.ubervu.com/> |
| Mô tả | UberVu là công cụ theo dõi theo thời gian thực có quy mô lớn và lịch sử hình thành bền vững. UberVu hiện được tin dùng bởi khá nhiều doanh nghiệp chuyên thực hiện những chiến dịch quảng bá trên mạng xã hội. |
| Điểm mạnh | Có khả năng theo dõi theo thời gian thực.  Có giao diện đẹp.  Hỗ trợ nhiều loại báo cáo, dễ sử dụng. |
| Điểm yếu | Chỉ có khả năng hoạt động trên một số nguồn lớn (facebook, twitter).  Thị trường chủ yếu không phải là Việt Nam |
| Số lượng website bao phủ | Chỉ các mạng xã hội lớn |
| Mục tiêu thị trường | Cả thế giới |
| Hoạt động | Vẫn tiếp tục được phát triển |

## Sysomos



|  |  |
| --- | --- |
| Địa chỉ | <http://www.sysomos.com/> |
| Mô tả | Sysomos là công cụ theo dõi mạng xã hội theo thời gian thực, được sử dụng bởi thương hiệu đắt giá nhất hành tinh Coca-cola. |
| Điểm mạnh | Có khả năng theo dõi theo thời gian thực.  Có giao diện đẹp.  Có hệ thống cơ sở dữ liệu là các trang tin, các trang blog đồ sộ. |
| Điểm yếu | Không theo dõi được facebook.  Không theo dõi được một số forum, website lớn ở Việt Nam.  Hỗ trợ ít định dạng báo cáo. |
| Số lượng website bao phủ | Twitter và các blog, trang tin lớn ở Việt Nam |
| Mục tiêu thị trường | Cả thế giới |
| Hoạt động | Vẫn tiếp tục được phát triển |

## Noti5



|  |  |
| --- | --- |
| Địa chỉ | <http://noti5.vn> |
| Mô tả | Noti5 là sản phẩm theo dõi thảo luận trên một số forum lớn ở Việt Nam và được phát triển bởi người Việt. |
| Điểm mạnh | Có khả năng theo dõi theo thời gian thực.  Có khả năng nhận diện cảm xúc theo tiếng Việt. |
| Điểm yếu | Khả năng thu thập thiếu cả chiều rộng và chiều sâu.  Số lượng các website được đo lường khá ít.  Hỗ trợ ít định dạng báo cáo.  Quy trình nạp và xuất còn chưa tự động hóa hoàn toàn |
| Số lượng website bao phủ | Facebook và các forum lớn ở Việt Nam, tổng khoảng 20 website. |
| Mục tiêu thị trường | Việt Nam |
| Hoạt động | Vẫn tiếp tục được duy trì nhưng phát triển chậm |

# PHÂN TÍCH CHỨC NĂNG

## Tổng quan chức năng

Hệ thống lắng nghe mạng xã hội có thể chia ra làm hai phần, một chỉ lo nhiệm vụ thu thập, một chỉ lo nhiệm vụ truy vấn. Hai hệ thống sẽ có thể hoạt động hoàn toàn độc lập với nhau và không gây ảnh hưởng lẫn nhau khi một trong hai hệ thống có sai sót. Hơn nữa, theo nhu cầu ngày càng cao, chúng ta cũng có thể dễ dàng phân tán một phần hệ thống (thu thập hoặc truy vấn) mà không cần tốn chi phí cho cả hai.

### Hệ thống thu thập

Một hệ thống thu thập cần phải có các chức năng sau:

* Hiểu được các chuẩn kết nối của Word Wide Web.
* Đọc hiểu HTML.
* Lưu trữ dữ liệu một cách có cấu trúc.
* Mô phỏng hành động của người dùng thật xuất sắc.

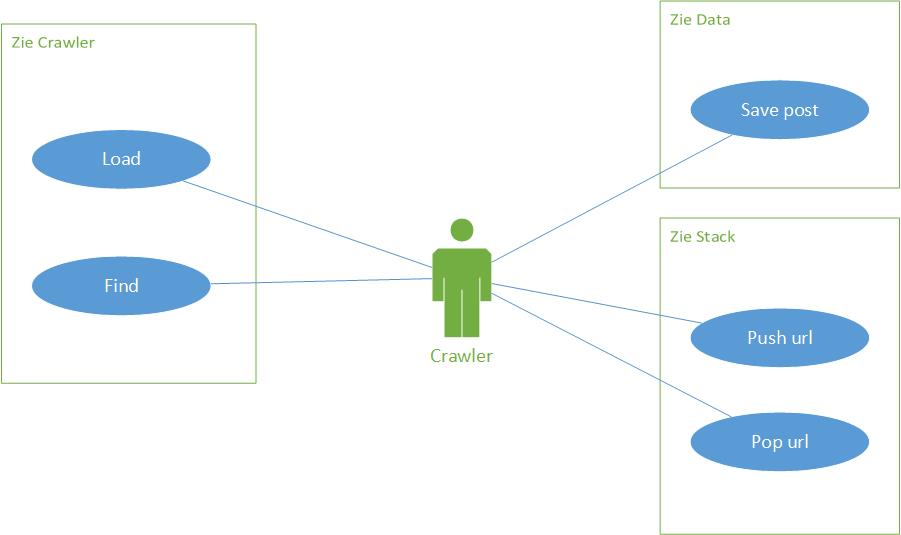
### Hệ thống báo cáo

Dựa trên đặc điểm nghiệp vụ, hệ thống báo cáo phục vụ công tác nghiên cứu thị trường cần phải đạt được các chức năng sau:

* Truy vấn dữ liệu bằng query.
* Thống kê dữ liệu theo ngày (Trend[[2]](#footnote-2)).
* Thống kê dữ liệu theo theo luồng thảo luận (Thread[[3]](#footnote-3)).
* Thống kê dữ liệu theo nguồn (Source).
* Thống kê dữ liệu theo loại mạng xã hội (Social Media Type[[4]](#footnote-4) hay SMT).

## Sơ đồ Use case

### Use case Crawler



Hình ‑ – Use case Crawler

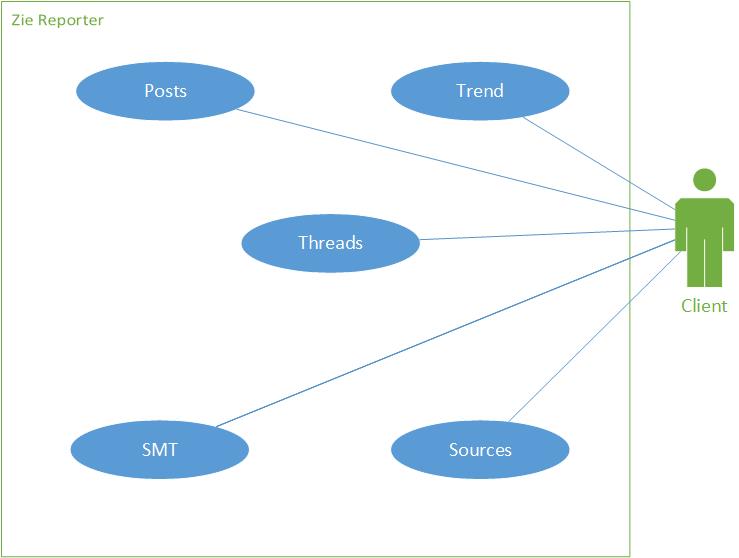
Danh sách Actor

|  |  |  |
| --- | --- | --- |
| STT | Tên Actor | Ý nghĩa/Ghi chú |
| 1 | Crawler | Crawler thực hiện thu thập dữ liệu. |

Danh sách Use case

| STT | Tên Use case | Ý nghĩa/Ghi chú |
| --- | --- | --- |
| 1 | Load | Tải một trang web từ địa chỉ url và thực hiện đọc cấu trúc HTML trả về từ trang web. |
| 2 | Find | Tìm một node HTML được trả về từ trang web thông qua hàm load. |
| 3 | Save post | Lưu một thảo luận tìm được từ việc truy cập trang web. |
| 4 | Push url | Nạp một liên kết tìm được vào hệ thống. |
| 5 | Pop url | Lấy liên kết cần xử lý tiếp theo từ hệ thống. |

### Use case Client



Hình ‑ - Use case Client

Danh sách Actor

|  |  |  |
| --- | --- | --- |
| STT | Tên Actor | Ý nghĩa/Ghi chú |
| 1 | Client | Người dùng đầu cuối hoặc một hệ thống khác truy cập tới Zie (có thể là Front-end) |

Danh sách Use case

| STT | Tên Use case | Ý nghĩa/Ghi chú |
| --- | --- | --- |
| 1 | Posts | Lấy danh sách các thảo luận. |
| 2 | Trend | Lấy danh sách số lượng các thảo luận tính theo ngày. |
| 3 | Sources | Lấy danh sách các nguồn có thảo luận đông nhất. |
| 4 | Threads | Lấy danh sách các luồng thảo luận đông nhất. |
| 5 | SMT | Lấy danh sách các loại mạng xã hội có số lượt thảo luận đông nhất. |

Actor Client ở đây có thể là người dùng thực sự, thông thường là admin hoặc người có khả năng quản trị hệ thống thao tác trực tiếp với Zie. Hoặc trong trường hợp cá biệt, một người dùng đầu cuối có thể được cấp quyền sử dụng Zie.

Trong trường hợp thông thường, Zie sẽ là một hệ thống ngầm và Actor Client là một hệ thống khác giao tiếp trực tiếp với User và nhận dữ liệu xử lý từ Zie. Client không nhất thiết phải cài đặt trên cùng một máy với Zie.

### Use case Save post

| Tên Use case | Save post | |
| --- | --- | --- |
| Mục đích sử dụng | Lưu một thảo luận vừa được thu thập | |
| Mô tả | Crawler thực hiện lưu trữ một thảo luận vừa thu thập được hoặc thông qua giao diện, người dùng thực hiện lưu trữ một thảo luận. | |
| Tác nhân chính | Crawler. | |
| Các tác nhân khác | Người dùng đầu cuối | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Id của thảo luận là một số lớn hơn 0 vừa được lưu trữ vào hệ thống. | |
| Kết quả khác | Kết quả | Nguyên nhân |
| -1 | Lệnh gửi tới không đúng cú pháp. |
| -2 | Không thể truy cập tới cơ sở dữ liệu. |
| -3 | Không đủ tham số. |
| -4 | Định dạng ngày tháng không chính xác. |
| Dòng sự kiện chính | Crawler truy cập vào website cần thu thập dữ liệu.  Crawler phân tích dữ liệu và nạp vào hệ thống các thảo luận vừa nhận diện được. | |
| Dòng sự kiện khác | Không có. | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

### Use case Push url

| Tên Use case | Push url | |
| --- | --- | --- |
| Mục đích sử dụng | Nạp một url vừa phát sinh vào stack | |
| Mô tả | Crawler xử lý một trang trong website và phát hiện được các liên kết tới những trang khác trong website và thực hiện nạp các liên kết đó vào hệ thống. | |
| Tác nhân chính | Crawler. | |
| Các tác nhân khác | Không có. | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Id của url vừa được nạp vào hệ thống là một số lớn hơn 0. | |
| Kết quả khác | Kết quả | Nguyên nhân |
| -1 | Lệnh gửi tới không đúng cú pháp. |
| -2 | Không thể truy cập tới cơ sở dữ liệu. |
| -3 | Không đủ tham số. |
| Dòng sự kiện chính | Crawler truy cập vào website cần thu thập dữ liệu.  Crawler phân tích dữ liệu và nạp vào hệ thống các liên kết tới các website khác. | |
| Dòng sự kiện khác | Người dùng nạp vào hệ thống một đường dẫn chưa được phát hiện bởi Crawler.  Đường dẫn ngày lập tức được nạo vào hệ thống và được xử lý ngay bởi Crawler. | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

### Use case Pop url

| Tên Use case | Pop url | |
| --- | --- | --- |
| Mục đích sử dụng | Lấy url tiếp theo sẽ xử lý theo thứ tự ưu tiên của Zie Stack. | |
| Mô tả | Crawler hoàn tất xử lý một trang web và gửi yêu cầu lên hệ thống hỏi địa chỉ trang web tiếp theo cần xử lý. | |
| Tác nhân chính | Crawler. | |
| Các tác nhân khác | Không có. | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Id của url, url và referer của url tiếp theo cần được xử lý. | |
| Kết quả khác | Kết quả | Nguyên nhân |
| 0 | Tạm thời chưa có url mới cần được xử lý |
| Dòng sự kiện chính | Crawler khởi động hoặc sau khi hoàn tất xử lý một trang web. Crawler tiếp tục hỏi hệ thống địa chỉ trang web kế tiếp cần xử lý là gì. | |
| Dòng sự kiện khác | Không có. | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

### Use case Posts

| Tên Use case | Posts | |
| --- | --- | --- |
| Mục đích sử dụng | Xuất dữ liệu là các thảo luận trong một khoảng thời gian quy định và thỏa một query do người dùng nhập vào. | |
| Mô tả | Người dùng là nhà nghiên cứu sử dụng hệ thống Front-end để nhập query, chọn thời gian cần truy xuất dữ liệu và tạo một báo cáo dữ liệu là những thảo luận trong khoảng thời gian được chọn thỏa query của người dùng. | |
| Tác nhân chính | Người dùng. | |
| Các tác nhân khác | Không có. | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Danh sách các thảo luận thỏa điều kiện | |
| Kết quả khác | Kết quả | Nguyên nhân |
| 1 | Search engine không trả về kết quả và gửi kèm lỗi. |
| Kết quả rỗng | Không có thảo luận phù hợp trong khoảng thời gian lựa chọn. |
| Dòng sự kiện chính | Người dùng truy xuất vào front-end và thực hiện tạo một báo cáo mới có kiểu báo cáo là Message. | |
| Dòng sự kiện khác | Không có. | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

### Use case Trend

| Tên Use case | Trend | |
| --- | --- | --- |
| Mục đích sử dụng | Xuất dữ liệu là số lượng thảo luận trong một khoảng thời gian quy định tính theo ngày và thỏa một query do người dùng nhập vào. | |
| Mô tả | Người dùng là nhà nghiên cứu sử dụng hệ thống Front-end để nhập query, chọn thời gian cần truy xuất dữ liệu và tạo một báo cáo dữ liệu là số lượng những thảo luận trong khoảng thời gian được chọn tính theo ngày thỏa query của người dùng. | |
| Tác nhân chính | Người dùng. | |
| Các tác nhân khác | Không có. | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Danh sách các các ngày và số lượng thảo luận thỏa điều kiện trong các ngày đó. | |
| Kết quả khác | Kết quả | Nguyên nhân |
| 1 | Search engine không trả về kết quả và gửi kèm lỗi. |
| Kết quả rỗng | Không có thảo luận nào trong thời gian phù hợp để thực hiện thống kê. |
| Dòng sự kiện chính | Người dùng truy xuất vào front-end và thực hiện tạo một báo cáo mới có kiểu báo cáo là Trend. | |
| Dòng sự kiện khác | Không có. | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

### Use case Sources

| Tên Use case | Sources | |
| --- | --- | --- |
| Mục đích sử dụng | Xuất dữ liệu là top 10 website có số lượng thảo luận trong một khoảng thời gian quy định và thỏa một query do người dùng nhập vào cao nhất. | |
| Mô tả | Người dùng là nhà nghiên cứu sử dụng hệ thống Front-end để nhập query, chọn thời gian cần truy xuất dữ liệu và tạo một báo cáo dữ liệu là những website có số lượt thảo luận cao nhất. | |
| Tác nhân chính | Người dùng. | |
| Các tác nhân khác | Không có. | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Danh sách các website có các thảo luận thỏa điều kiện nhiều nhất | |
| Kết quả khác | Kết quả | Nguyên nhân |
| 1 | Search engine không trả về kết quả và gửi kèm lỗi. |
|  | Kết quả rỗng | Không có thảo luận nào trong thời gian phù hợp để thực hiện thống kê. |
| Dòng sự kiện chính | Người dùng truy xuất vào front-end và thực hiện tạo một báo cáo mới có kiểu báo cáo là Source. | |
| Dòng sự kiện khác | Không có | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

### Use case Threads

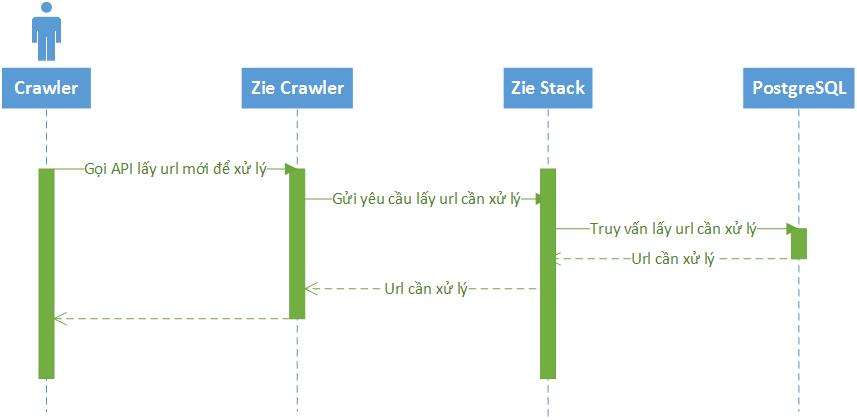
| Tên Use case | Threads | |
| --- | --- | --- |
| Mục đích sử dụng | Xuất dữ liệu là các luồng có số lượng thảo luận cao nhất trong một khoảng thời gian quy định và thỏa một query do người dùng nhập vào. | |
| Mô tả | Người dùng là nhà nghiên cứu sử dụng hệ thống Front-end để nhập query, chọn thời gian cần truy xuất dữ liệu và tạo một báo cáo dữ liệu là những luồng có số lượng thảo luận cao nhất trong khoảng thời gian được chọn thỏa query của người dùng. | |
| Tác nhân chính | Người dùng. | |
| Các tác nhân khác | Không có. | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Danh sách các luồng thỏa điều kiện | |
| Kết quả khác | Kết quả | Nguyên nhân |
| 1 | Search engine không trả về kết quả và gửi kèm lỗi. |
| Kết quả rỗng | Không có thảo luận phù hợp trong thời gian được chọn để thực hiện thống kê. |
| Dòng sự kiện chính | Người dùng truy xuất vào front-end và thực hiện tạo một báo cáo mới có kiểu báo cáo là Thread. | |
| Dòng sự kiện khác | Không có. | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

### Use case SMT (Social Media Type)

| Tên Use case | SMT | |
| --- | --- | --- |
| Mục đích sử dụng | Xuất dữ liệu là các loại mạng xã hội có số lượng thảo luận cao nhất trong một khoảng thời gian quy định và thỏa một query do người dùng nhập vào. | |
| Mô tả | Người dùng là nhà nghiên cứu sử dụng hệ thống Front-end để nhập query, chọn thời gian cần truy xuất dữ liệu và tạo một báo cáo dữ liệu là những loại mạng xã hội có số lượng thảo luận cao nhất trong khoảng thời gian được chọn thỏa query của người dùng. | |
| Tác nhân chính | Người dùng. | |
| Các tác nhân khác | Không có. | |
| Điều kiện/trạng thái hệ thống khi bắt đầu Use case | Không có. | |
| Kết quả mong đợi | Danh sách các loại mạng xã hội thỏa điều kiện | |
| Kết quả khác | Kết quả | Nguyên nhân |
| 1 | Search engine không trả về kết quả và gửi kèm lỗi. |
| Kết quả rỗng | Không có thảo luận phù hợp trong thời gian được chọn để thực hiện thống kê. |
| Dòng sự kiện chính | Người dùng truy xuất vào front-end và thực hiện tạo một báo cáo mới có kiểu báo cáo là SMT. | |
| Dòng sự kiện khác | Không có. | |
| Yêu cầu đặc biệt | Không có. | |
| Mức độ ưu tiên | 1 | |

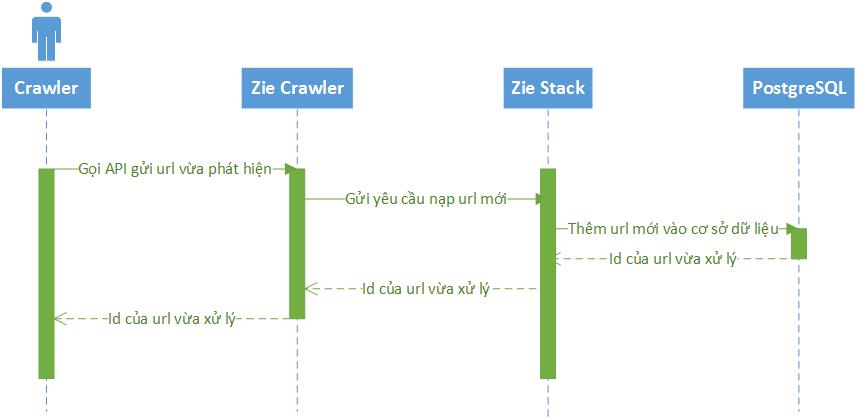
## Sơ đồ tuần tự

### Sơ đồ tuần tự Use case Pop url



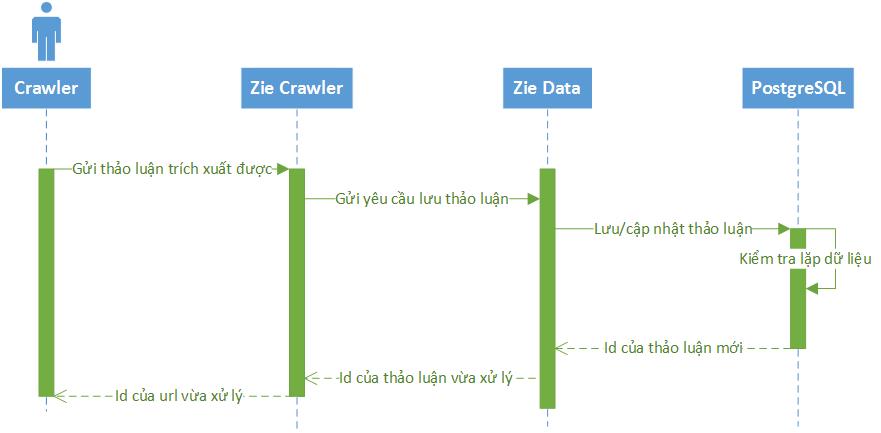
Hình ‑ - Sơ đồ tuần tự Use case Pop url

### Sơ đồ tuần tự Use case Push url



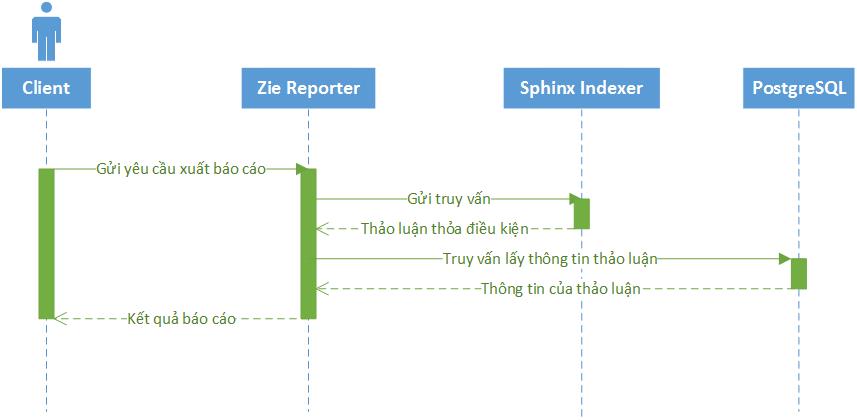
Hình ‑ - Sơ đồ tuần tự Use case Push url

### Sơ đồ tuần tự Use case Save post



Hình ‑ - Sơ đồ tuần tự Use case Save Post

### Sơ đồ tuần tự Use case Posts



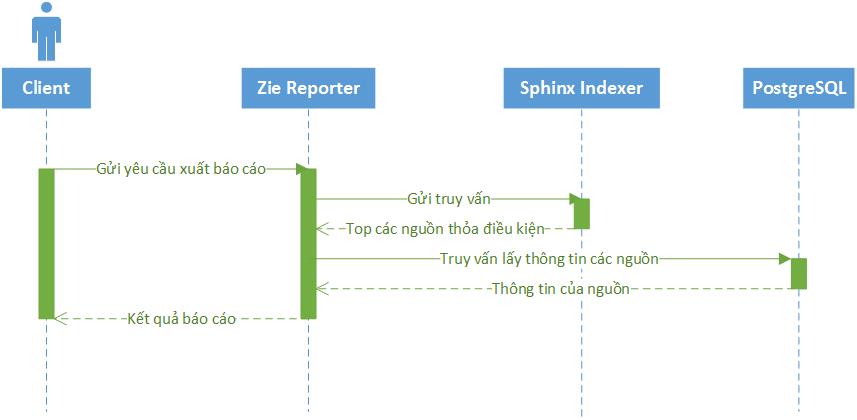
Hình ‑ - Sơ đồ tuần tự Use case Post

### Sơ đồ tuần tự Use case Trend



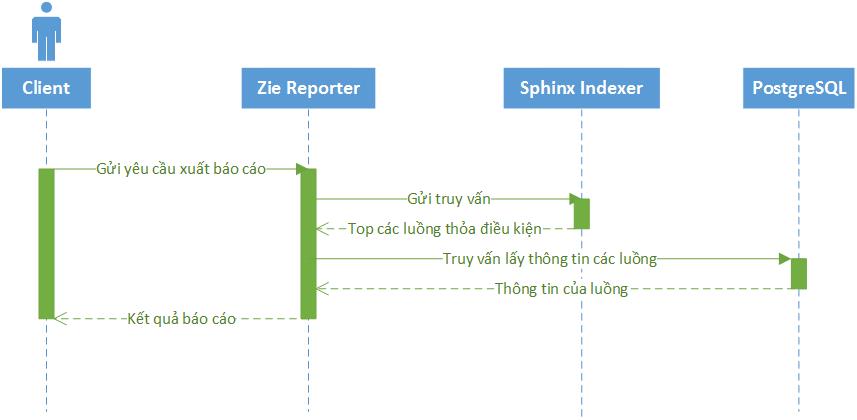
Hình ‑ - Sơ đồ tuần tự Use case Trend

### Sơ đồ tuần tự Use case Sources



Hình ‑ - Sơ đồ tuần tự Use case Sources

### Sơ đồ tuần tự Use case Threads



Hình ‑ - Sơ đồ tuần tự Use case Threads

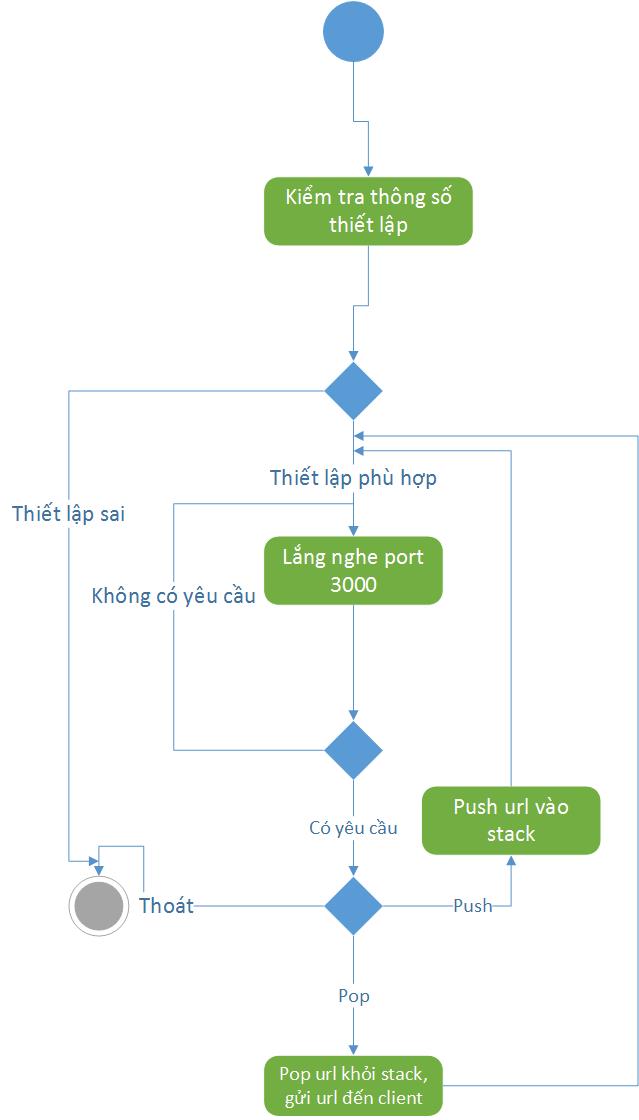
### Sơ đồ tuần tự Use case SMT



Hình ‑ - Sơ đồ tuần tự Use case SMT

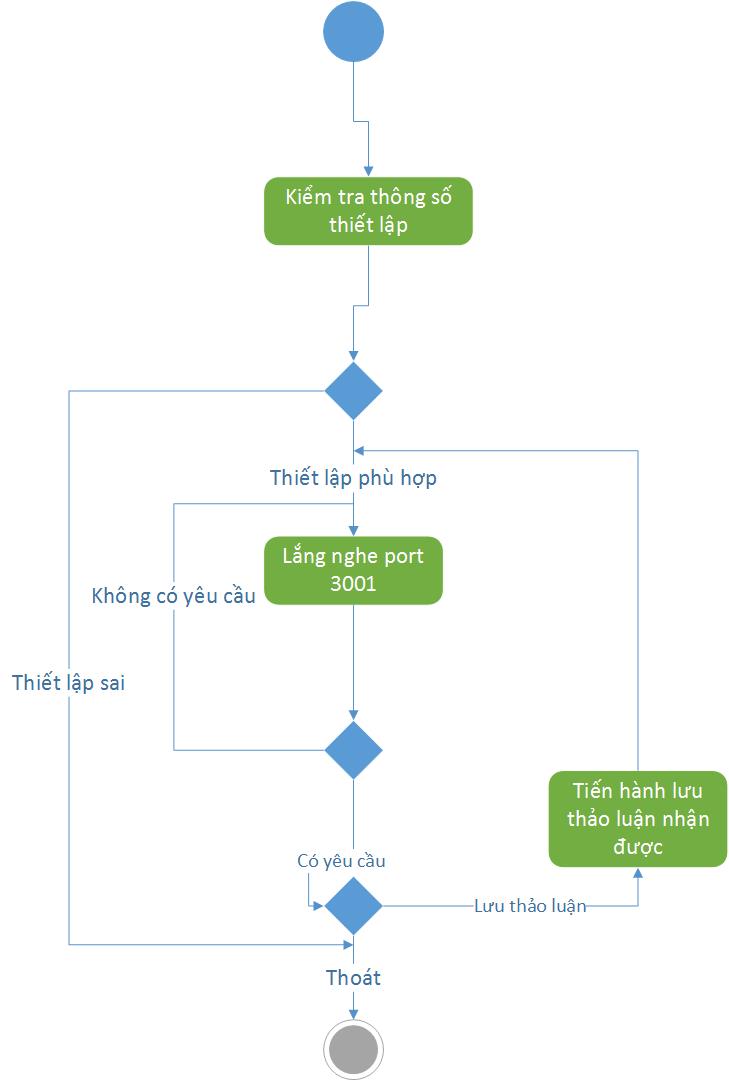
## Sơ đồ hoạt động

### Zie Stack



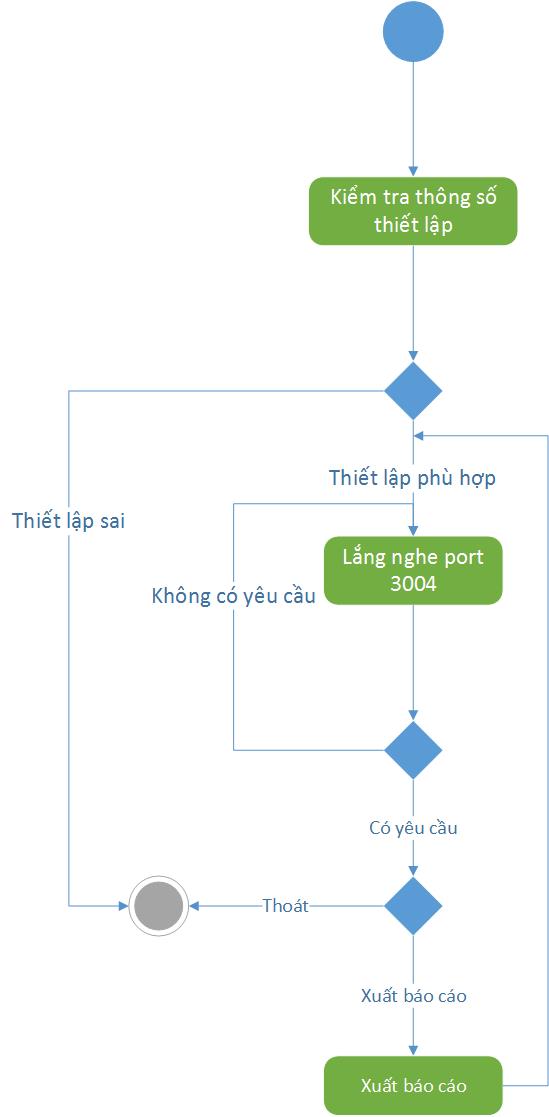
Hình ‑ - Sơ đồ hoạt động Zie Stack

### Zie Data



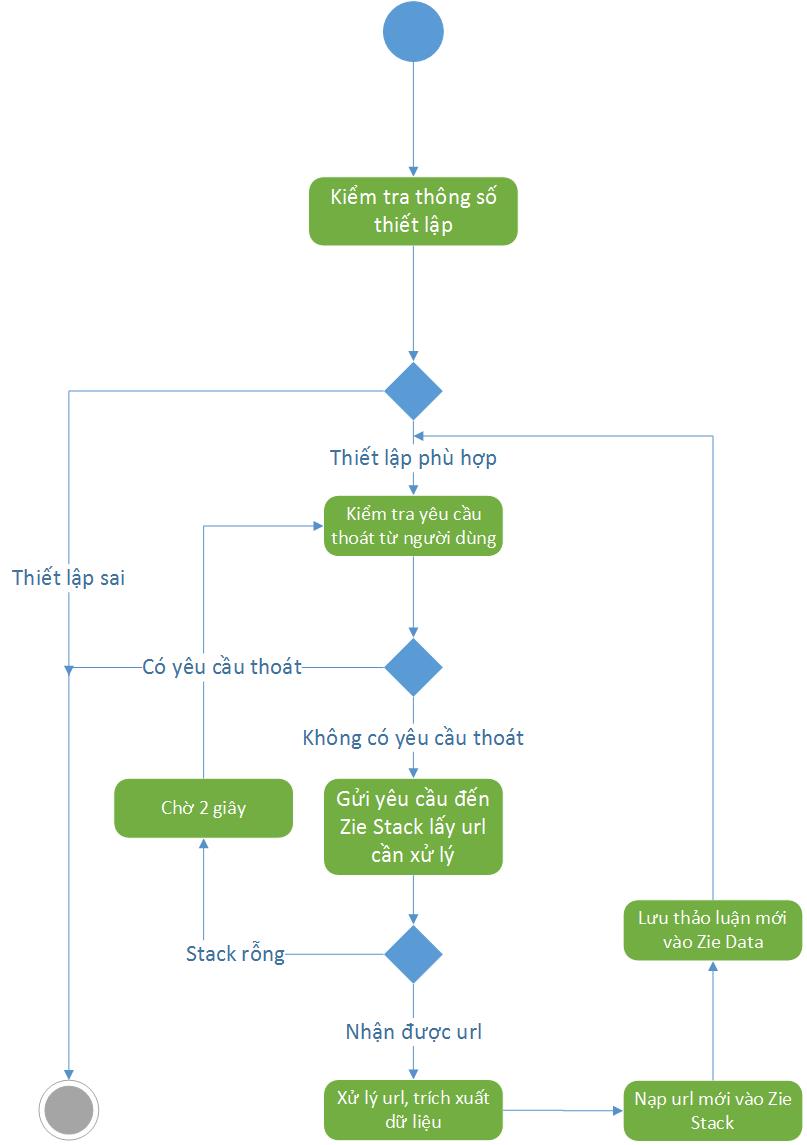
Hình ‑ - Sơ đồ hoạt động Zie Data

### Sơ đồ hoạt động Zie Reporter



Hình ‑ - Sơ đồ hoạt động Zie Reporter

### Sơ đồ hoạt động Zie Crawler



Hình ‑ - Sơ đồ hoạt động Zie Crawler

# HƯỚNG TIẾP CẬN

## Một hệ thống duy nhất vừa đảm nhiệm lưu trữ, vừa đảm nhiệm truy xuất

Phương pháp sử dụng một hệ thống duy nhất dùng cho cả việc lưu trữ và truy xuất dữ liệu có điểm lợi là tiết kiệm được chi phí cho thiết bị lưu trữ, mỗi thể hiện chỉ được lưu trữ một lần duy nhất, có khả năng hoạt động theo thời gian thực.

Tuy nhiên, hệ thống có điểm yếu là hiệu suất của cả quá trình lưu trữ và truy xuất sẽ bị ảnh hưởng. Hệ thống lưu trữ phải thường xuyên index lại dữ liệu dành cho hệ thống truy xuất. Hệ thống truy xuất cần phải kết nối dữ liệu dạng thô do cần cho việc cập nhật lại thông tin ở hệ thống lưu trữ.

Để đáp ứng được nhu cầu cần có một hệ thống quản trị cơ sở dữ liệu quan hệ mạnh, hỗ trợ Full-text nâng cao, hỗ trợ lưu trữ dữ liệu thu thập khổng lồ.

Hệ quản trị cơ sở dữ liệu đề xuất: Oracle[[5]](#footnote-5), PostgreSQL[[6]](#footnote-6).

Hệ quản trị cơ sở dữ liệu Oracle hỗ trợ Oracle Text tương tự Full-text search, hỗ trợ các toán tử boolean, hàm tìm kiếm NEAR[[7]](#footnote-7) và khả năng tìm kiếm dựa trên từ điển đồng nghĩa (tuy nhiên chưa hỗ trợ tiếng Việt). Ngoài ra Oracle bản Enterprise hỗ trợ chức năng Partitioning phù hợp để lưu trữ dữ liệu với dung lượng lớn.

Hệ quản trị cơ sở dữ liệu PostgreSQL hỗ trợ Full-text search, tuy nhiên tại version hiện tại (9.3) PostgreSQL mới chỉ hỗ trợ các toán tử boolean. Chưa đáp ứng được nhu cầu truy vấn dữ liệu nâng cao. Tương tư Oracle, PostgreSQL hỗ trợ Partitioning tuy nhiên phức tạp hơn trong cài đặt và có khả năng tùy biến cao hơn.

## Hai hệ thống, một đảm nhiệm lưu trữ, một đảm nhiệm truy xuất

Sử dụng hai hệ thống một cho truy xuất và còn lại cho lưu trữ sẽ có thể tối ưu hiệu suất cho từng hệ thống. Một hệ thống lưu trữ không cần phải index dữ liệu như một hệ thống chỉ truy xuất. Một hệ thống chỉ truy xuất không cần phải lưu trữ dữ liệu dạng thô như hệ thống chỉ lưu trữ.

Hướng tiếp cận hai hệ thống có điểm mạnh là tăng cường hiệu suất nhưng có điểm yếu là cần gấp đôi chi phí cho thiết bị lưu trữ. Sẽ có một độ trễ nhất định giữa hệ thống lưu trữ và hệ thống truy xuất trong quá trình đồng bộ. Độ phức tạp cao.

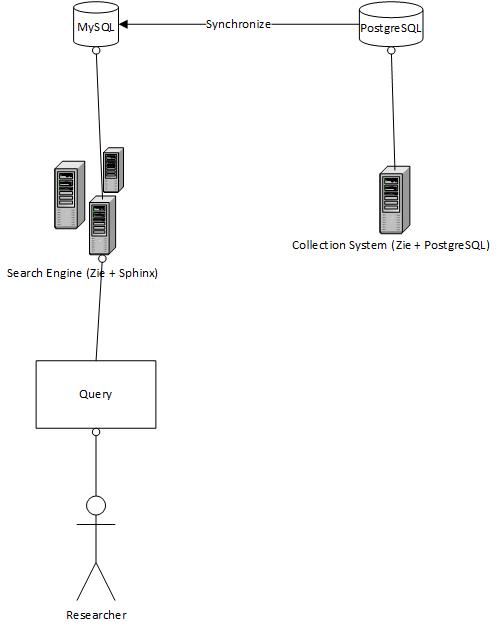
Hệ thống lưu trữ đề xuất: Oracle, PostgreSQL.

Hệ thống truy xuất đề xuất: Sphinx, Lucence.

Trong mô hình hai hệ thống, điểm mạnh hỗ trợ Full-text search[[8]](#footnote-8) nâng cao của Oracle không còn giá trị nữa, thay vào đó là chi phí cao cho bản Enterprise để có được chức năng Partitioning. Trong khi đó, với PostgreSQL chi phí là bằng không.

Sphinx là hệ thống tìm kiếm mạnh, được viết bằng C++ và hỗ trợ trực tiếp MySQL[[9]](#footnote-9), PostgreSQL. Hỗ trợ các toán tử boolean, hàm NEAR và các toán tử phức tạp khác. Có tốc độ truy xuất và index dữ liệu cực nhanh.

Lucence là hệ thống tìm kiếm được viết bằng Java (đã được port qua các ngôn ngữ khác). Lucence hỗ trợ các toán tử tìm kiếm phức tạp. Tuy nhiên, việc kết hợp giữa Lucence và PostgreSQL không đơn giản như giữa Sphinx và PostgreSQL.



Hình ‑ - Sơ đồ mô tả hai hệ thống tách rời (Sphinx + Postgres)

## Đánh giá liên kết[[10]](#footnote-10) dựa trên liên kết đến chính nó

Các liên kết hay siêu liên kết trong một website sẽ dẫn đến một trang chứa nội dung hoặc đến gần trang chứa nội dung mà website muốn gửi đến người đọc. Thông qua liên kết, người dùng website có thể được dẫn đến nơi chứa nội dung bài viết mà người dùng cần. Chính vì vậy, việc đi theo liên kết và đánh giá một liên kết dựa trên các liên kết khác đến nó là có cơ sở.

Ví dụ: Google đánh giá một website có thứ hạng cao dựa vào số lượng website liên kết đến website đó.

Nếu đánh giá liên kết dựa trên số lượng liên kết đến chính nó, các liên kết được dẫn đến nhiều sẽ được ưu tiên xử lý trước và các liên kết ít được dẫn đến hơn sẽ ít được xử lý hơn. Cách đánh giá này tuy có cơ sở nhưng dùng đến đánh giá từng trang trong website thì không chính xác. Trong một website, trang chủ thường được dẫn đến bởi tất cả các trang thành phần, chính vì vậy, trang chủ sẽ có thứ hạng cao nhất, tiếp đến là các trang chuyên mục rồi mới đến các trang nội dung. Theo thứ tự như vậy, hệ thống sẽ đọc được dữ liệu cuối cùng, và sẽ ưu tiên phát tán (đi vào các liên kết tìm được) hơn.

## Đánh giá liên kết dựa trên độ mới của nội dung

Các liên kết chứa nội dung mới nhất sẽ được ưu tiên đưa lên hàng đầu và được ưu tiên xử lý trước. Để đánh giá được chính xác, các liên kết đó cần phải được xử lý và có dữ liệu chứa trong liên kết hoặc phải tìm được cơ sở xác định độ mới của liên kết.

Đánh giá dựa trên độ mới sẽ sinh ra hai trường hợp: một là liên kết chứa dữ liệu được đánh giá chính xác bằng thời gian của dữ liệu đó (thường là thảo luận trong trang web đó), hai là liên kết chưa từng được truy xuất và không xác định được chính xác độ mới của dữ liệu chứa trong liên kết.

Liên kết có thể được đánh giá chính xác nếu có dữ liệu, và có thể được dự đoán nếu chưa có dữ liệu. Giải thuật dự đoán sẽ ảnh hưởng trực tiếp đến chất lượng hệ thống, một hệ thống sử dụng tốt tài nguyên sẽ truy cập tối đa đến những trang mang lại dữ liệu có ích.

Việc đánh giá liên kết có thể dựa trên độ mới của liên kết sinh ra liên kết đó hoặc độ mới của liên kết được sinh ra bởi liên kết đó hoặc chính một thông tin nào đó trong liên kết.

## Thu thập dữ liệu theo chiều sâu

Thu thập dữ liệu là bài toán phức tạp vì cần đảm bảo được dữ liệu là độc lập, không trùng lắp, dữ liệu phải đủ đáp ứng được yêu cầu của chiến dịch nghiên cứu.

Đối với thu thập dữ liệu theo chiều sâu, hệ thống sẽ thu thập dữ liệu của một website từ mới nhất tới cũ nhất, sau đó mới tiến hành thu thập trên website khác, cũng như cập nhật lại.

Thu thập dữ liệu theo chiều sâu giúp hệ thống thu thập đầy đủ dữ liệu của cả website nguồn từ khi website phát triển, khả năng bỏ sót dữ liệu là rất thấp. Tuy nhiên, để có được một hệ thống đủ mạnh để bao phủ 100% hàng loạt các website lớn khác khiến chi phí vượt lên rất cao. Tính đến tháng 12 năm 2013, website hdvietnam.com có hơn 4 triệu bài viết, con số đó ở tinhte.vn và vozforums.com là 18 triệu và 30 triệu bài viết.

## Thu thập dữ liệu theo chiều rộng

Thu thập dữ liệu theo chiều rộng đòi hỏi phải có một đánh giá các liên kết trước khi quyết định thu thập (liên kết nào cần trước, liên kết nào cần sau), đòi hỏi hệ thống phải có sự phân phối dữ liệu cần thu thập chính xác nếu không muốn hệ thống chỉ thu thập được những dữ liệu vô ích.

Thu thập dữ liệu theo chiều rộng giúp hệ thống bao phủ đều các website nguồn. Sức mạnh của hệ thống yêu cầu là không quá cao, chỉ cần đáp ứng được số lượng bài viết mới xuất hiện hàng ngày trên các website nguồn. Tuy nhiên, thu thập theo chiều rộng sẽ đánh đồng tất cả các website nguồn, khó thu thập tập trung trên một website nào đó theo nhu cầu của chiến dịch nghiên cứu thị trường, đòi hỏi chi phí cho giải thuật đánh giá liên kết hay nguồn dữ liệu, hơn nữa, nhu cầu dữ liệu cho các chiến dịch nghiên cứu là khác nhau, thu thập theo chiều rộng đòi hỏi phải xem các chiến dịch nghiên cứu có nhu cầu dữ liệu như nhau.

## Kết luận

* Sử dụng một hệ thống duy nhất đơn giản hóa quy trình xây dựng hệ thống, tuy nhiên sẽ làm giảm hiệu suất hệ thống, chi phí tăng cường cho bản quyền Oracle Enterprise[[11]](#footnote-11).
* Sử dụng hai hệ thống có độ phức tạp nhất định, hiệu suất hệ thống được đảm bảo. Chi phí thừa dành cho hệ thống chỉ là cần thêm thiết bị lưu trữ so với một hệ thống.
* Đánh giá liên kết dựa trên độ mới của liên kết sẽ có thể dẫn hệ thống tới những nơi chứa dữ liệu có ích cao hơn dựa trên liên kết đến chính nó.
* Thu thập dữ liệu kết hợp cả hai hướng theo chiều rộng và theo chiều sâu. Trong đó, chiều sâu cần được giới hạn và đánh giá dựa trên độ phủ của chiều rộng, việc thu thập theo chiều sâu khi chiều rộng đã đảm bảo độ phủ sẽ tối ưu hóa hiệu suất của hệ thống mà vẫn đảm bảo nhu cầu dữ liệu mới nhất.

Từ những tiếp cận trên, nhóm đã đi đến quyết định sử dụng mô hình Sphinx-PostgreSQL làm cơ sở cho hệ thống Zie, đánh giá liên kết bằng độ mới dữ liệu mà liên kết đó đem lại kết hợp dự đoán độ mới của liên kết, đồng thời kết hợp thu thập theo chiều sâu và chiều rộng.

# NHỮNG VẤN ĐỀ GẶP PHẢI KHI HIỆN THỰC HỆ THỐNG ZIE

Để dễ dàng trong việc trình bày những vấn đề có thể gặp phải khi hiện thực hệ thống Zie, đề tài cần thống nhất một số khái niệm sau để tránh gây nhầm lẫn.:

* Gọi các website nguồn là Source[[12]](#footnote-12).
* Những thành phần truy xuất, xử lý website và chuyển thành dữ liệu thô là Crawler.
* Hệ thống lưu trử, xử lý dữ liệu thô đưa thành dữ liệu trực quan là Zie.
* Mỗi đơn vị bài viết của một người dùng vào một thời gian nhất định và phân biệt hoàn toàn với các bài viết khác mà hệ thống thu thập sẽ được gọi là Post[[13]](#footnote-13).

### Lấy dữ liệu từ Source

#### Đặt vấn đề

Mỗi website sẽ có hàng loạt những thông tin, những bài viết của người dùng, chúng ta có thể chia nhỏ thông tin ra theo URL (Uniform Resource Locator). Mỗi URL có thể chứa một hay nhiều bài viết mà chúng ta cần lấy. Đơn vị nhỏ nhất trong Zie phải là Post chứ không thể là URL. Vấn đề này đòi hỏi hệ thống phải có khả năng tạo những kết nối tới các giao thức thông dụng được dùng trong Word Wide Web. Đồng thời, phải có khả năng biên dịch các mã web (HTML, XML) để phân tích được dữ liệu.

Mỗi URL sau một thời gian nhất định sẽ có thể bị thay đổi dữ liệu, dữ liệu bị thay đổi có thể là thêm mới hoặc cập nhật lại, ví dụ cùng một URL trong forum, có thể hôm nay có 3 bài viết, nhưng 5 phút sau sẽ có thêm người viết bài dẫn đến có 4 bài viết. Trong trường hợp này, Zie cũng cần phải nhận biết và có khả năng cập nhật dữ liệu mới.

#### Hướng giải quyết

Từ yêu cầu xử lý dữ liệu dẫn đến chúng ta phải hình thành một thành phần là Crawler, Crawler sẽ biên dịch được các trang web. Vấn đề biên dịch trang web là hoàn toàn khả thi, chúng ta chỉ cần lưu trữ được trang web ở dạng DOM[[14]](#footnote-14) (Document Object Model). Có rất nhiều hệ thống, thư viện hỗ trợ việc này, ngoài ra, dành cho những trường hợp đặc biệt, ta có thể đưa về xử lý dạng chuỗi với dữ liệu trả về từ Source vốn chỉ là một chuỗi các mã HTML có thể đọc được dễ dàng và xử lý rất tốt bằng Regular Expresion[[15]](#footnote-15).

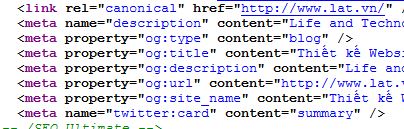
Sau một thời gian phù hợp Zie cần phải duyệt lại các URL đã từng xử lý, để cập nhật mới. Thời gian này phải được tính toán phù hợp để cân bằng giữa việc thu thập các URL mới và cập nhật các URL đã có.

### Nhận diện Post

#### Đặt vấn đề

Khi lưu trữ được một Post, sẽ có trường hợp hệ thống sẽ truy xuất lại và lại thu thập cùng Post đó. Nguyên nhân có thể do Source thay đổi cấu trúc URL hoặc cập nhật lại dữ liệu.

Post không thể được nhận diện qua URL vì nhiều URL có thể đưa đến cùng một nội dung (URL redirection). Tuy Word Wide Web đã đưa ra chuẩn các website có nhiều hơn một URL dẫn đến một nội dung cần có tham số meta canonical[[16]](#footnote-16), nhưng không phải tất cả đều tuân thủ nguyên tác này.



Hình ‑ - Chuẩn canonical meta tag

Các Post có thể có quan hệ với nhau, Post này sẽ là reply của Post khác. Cần tổ chức lưu trữ hiệu quả.

#### Hướng giải quyết

Post có thể được nhận diễn qua mã hash của nội dung Post, và một tham số nữa như chuyên mục của Post. Sẽ không có hai bài viết cùng một nội dung trong một chuyên mục. Trong trường hợp này, khả năng nhận dạng có thể giúp loại bỏ những bài viết lặp nội dung (spam) từ các website.

Việc lưu trữ quan hệ giữa các Post có thể sử dụng mô hình Nested Set Model[[17]](#footnote-17) đem lại hiệu quả truy xuất rất cao, tuy nhiên, trong thực tế chứng minh dữ liệu nghiên cứu không quá quan tâm về vấn đề này mà chủ yếu chỉ cần thông tin về bài viết sinh ra chuyên mục (Original Post) và các bài viết thành phần của chuyên mục đó.

### Nhận diện ngôn ngữ

#### Đặt vấn đề

Các website lớn hoặc các hệ thống mạng xã hội có thể chứa nhiều Post ở nhiều ngôn ngữ khác nhau. Hệ thống cần phải nhận diện được để đưa ra những thống kê theo một quốc gia, hoặc vùng lãnh thổ hợp lý.

Tuy nhiên, thói quen của người dùng không phải chỉ dùng một ngôn ngữ trong lúc viết Post, nên việc nhận diện ngôn ngữ cần phải có khả năng phát hiện được tất cả các ngôn ngữ người dùng sử dụng.



Hình ‑ - Người dùng viết lẫn lộn các ngôn ngữ

#### Hướng giải quyết

Trường hợp một: Source hỗ trợ nhiều ngôn ngữ:

Cần phải có một bộ nhận diện ngôn ngữ, có thể thông qua từ điển, hoặc bảng mã các ký tự của ngôn ngữ đó, những đặc trưng của ngôn ngữ. Về việc này, đã có các bộ ứng dụng phát hiện ngôn ngữ miễn phí được phét triển cho người dùng sử dụng, chúng ta có thể tận dụng lại, và chỉnh sửa những thứ cần thiết. Hoặc sử dụng những API chất lượng nhưng giá cả phải chăng như <http://detectlanguage.com/>

Theo tự nhiên, các bài post thường chỉ chứa cao nhất hai hoặc ba ngôn ngữ khác nhau, các trường hợp nhiều hơn ba ngôn ngữ là rất hiếm và phần lớn là lỗi của bộ nhận diện ngôn ngữ chưa được chính xác. Chính vì thế, chúng ta chỉ cần lưu trữ lại ba ngôn ngữ có tỷ lệ nhận diện cao nhất trong Post.

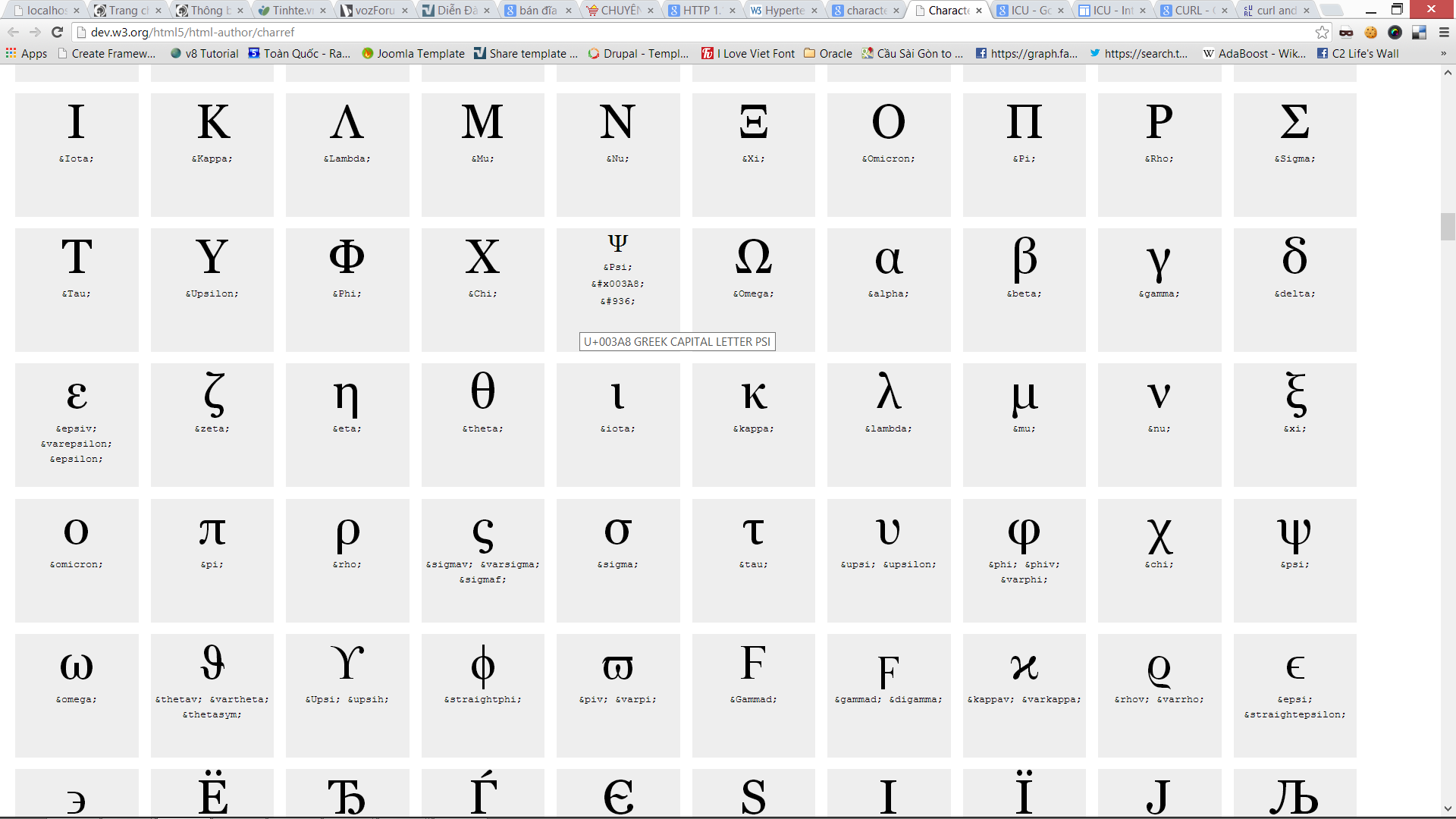
Trường hợp hai: Source chỉ chứa một ngôn ngữ

Trường hợp này có thể xử lý đơn giản là gán ngôn ngữ duy nhất cho Source, tuy nhiên, vẫn có thể có trường hợp ngoại lệ là Source sẽ phát triển thêm một ngôn ngữ mới, hoặc thói quen người dùng sẽ viết bài trên nhiều ngôn ngữ. Chúng ta có thể an toàn hơn bằng cách đặt ưu tiên cho một ngôn ngữ nhất định.

### Tính tương thích của hệ thống

#### Đặt vấn đề

Hiện nay, Word Wide Web đã phát triển rất xa và có rất nhiều nguyên tắc được lập ra trong việc phát triển và sử dụng Word Wide Web. Chính vì thế, tính tương thích của hệ thống là điều rất quan trọng, hệ thống cần phải đọc hiểu được các chuẩn Word Wide Web hiện đại (HTTP/1.0, HTTP/1.1), có khả năng giải mã các chuẩn encode phổ biến (UTF-8, ISO-8859-1, Windows-1252…), có khả năng kết nối tới các giao thức mã hóa (HTTPS, SFTP, FTPS…) và phải giải mã được các character entity[[18]](#footnote-18) (&nbsp;, &#222…) trong HTML.



Hình ‑ - Một phần bảng mã character entity từ w3.org

#### Hướng giải quyết

Trường hợp: đọc hiểu các chuẩn encode phổ biến

Tính đến năm 2013 chuẩn encode phổ biến nhất trong tiếng Việt là UTF-8 nên có thể ưu tiên giải mã chuẩn encode này trước, ngoài ra trong cộng đồng mã nguồn mở cũng đã có phát triển bộ thư viện ICU (International Components for Unicode) đến version 52.1 đã hỗ trợ rất nhiều chuẩn encode khác nhau.

Trường hợp: hỗ trợ nhiều giao thức kết nối

Có thể sử dụng CURL là giải pháp tối ưu, CURL là bộ thư việc mã nguồn mở được phát triển và tin dùng bởi rất nhiều ứng dụng lớn trên thị trường hiện nay. CURL hỗ trợ hầu hết các giao thức cần thiết trong Word Wide Web như HTTP, HTTPS, IMAP, IMAPS….

Trường hợp: giải mã các character entity

Hiện nay, trên thế giới có khá nhiều thư viện hỗ trợ việc giải mã các character entity, nhưng chưa có một thư viện nào hỗ trợ thực sự đầy đủ. Giải mã các chracter entity có thể xây dựng lại dựa trên nguyên tắc của character entity đó và bộ từ điển phù hợp cho character entity tương ứng. Các character entity đã được liệt kê đầy đủ tại <http://w3.org> nên việc thực hiện cũng không chiếm quá nhiều thời gian.

### Xử lý dữ liệu lớn

#### Đặt vấn đề

Dữ liệu thu thập được từ hệ thống theo thời gian sẽ là hàng triệu, chục triệu, thậm chí là trăm triệu Post (website tinhte.vn có hơn 18 triệu bài viết từ năm 2006 đến nay, vozforums.com là hơn 30 triệu bài). Với một nguồn dữ liệu khổng lồ sẽ làm cho hệ thống dễ dàng quá tải và mất khả năng kiểm soát. Xóa dữ liệu cũ cũng không phải là một biện pháp hay vì dữ liệu cũ vẫn là dữ liệu giá trị và có thể nói là cực kỳ giá trị trong phân tích.

#### Hướng giải quyết

Biện pháp hay và hợp lý nhất đó chính là chia nhỏ dữ liệu theo thời gian đồng thời phân tán hệ thống, những dữ liệu càng gần thời gian hiện tại sẽ có tốc độ cập nhật cao hơn và được ưu tiên xử lý nhiều hơn, các dữ liệu đã lâu cũng sẽ có ít Post mới, ít được chỉnh sửa từ Source nên chúng ta cũng có thể bỏ qua các xử lý cập nhật, trích xuất để tối ưu tài nguyên cho việc cập nhật dữ liệu mới nhất.

Tố chức Zie có khả năng phân tán để chia nhỏ xử lý, chia nhỏ dữ liệu. Trong thực tế chứng minh, mỗi Post thường có dung lượng trunh bình là 1 KB, không quá 1000 từ, tuy nhiên vẫn có những bài Post lên tới 100KB, tương đương khoảng 100 nghìn từ đối với dạng mã hóa là UTF-8. Thì phải với 10 triệu Post, chúng ta chỉ tốn khoảng 10 GB lưu trữ, chủ yếu là phải lưu trữ các quan hệ thật tốt để đảm bảo tốc độ truy xuất.

Hướng các Crawler phát tán URL của một website một cách có chọn lọc, không phán tán bừa bãi sẽ hướng được hệ thống tới những URL chất lượng chỉ chứa dữ liệu cần thiết, giảm thiểu dữ liệu thừa, giảm thiểu tiêu tốn tài nguyên hệ thống.

### Nhiều Source dẫn đến nhiều Crawler

#### Đặt vấn đề

Trong thực tế, số lượng Source cần phải lắng nghe có thể rất lớn, hàng nghìn, chục nghìn, việc cần phải cung cấp mỗi Source một Crawler khác nhau cũng làm cho số lượng Crawler phải tăng lên, hơn nữa, khi các Source có thay đổi cấu trúc website như giao diện, URL, bảo mật, Crawler cũng cần phải được cập nhật.

#### Hướng giải quyết

Trước hết cần có một bộ phận có khả năng phát hiện vấn đề về Crawler bằng cách truy vấn lại cùng một đường dẫn, nhưng phát hiện số lượng Post khác quá nhiều so với ban đầu.

Việc lập trình, hay tạo Crawler cần phải được đơn giản hóa, nếu có thể chỉ cần dùng giao diện đồ họa để tạo Crawler sẽ là một bước tiến vượt bậc. Trong thực tế, Crawler chỉ cần nhận diện những vị trí có dữ liệu, tuy nhiên, có thể Crawler sẽ phải xử lý các vấn đề như chuyển đổi định dạng ngày tháng, xác định hướng phát tán URL, nên thực tế vẫn cần một ngôn ngữ lập trình làm nền hỗ trợ. Ở đây, chúng ta có thể sử dụng các ngôn ngữ Script không cần biên dịch và đính kèm bộ biên dịch vào Zie như (Javascript, Python...).

Crawler cần phải được hỗ trợ khả năng giả dạng người dùng một cách hoàn hảo, vì hiện tại, với các website lớn, việc họ xây dựng các tường lửa để ngăn chặn những kết nối không phải là người dùng nhằm bảo vệ website (ngăn chặn DDOS) là rất dễ hiểu. Tránh tình trạng Source sẽ phát hiện và chặn Crawler.



Hình ‑ - Một tường lửa đơn giản để nhận diện người dùng

# THIẾT KẾ HỆ THỐNG ZIE

Hình ‑ - Mô hình đơn giản Zie

## Zie

Zie là thành phần cốt lõi của hệ thống và có tên trùng với tên hệ thống (Zie). Nhìn tổng quan, Zie là bộ não điều hành mọi hoạt động, bao gồm các hoạt động như chỉ đạo Crawler, thông báo lỗi, lưu trữ dữ liệu, chuẩn hóa dữ liệu. Nói cách khác, Zie là thành phần duy nhất trong hệ thống có khả năng hoạt động độc lập và là thành phần cực kỳ quan trọng. Zie được chia làm ba phần nhỏ như sau:

### Zie Stack

#### Ý tưởng:

Một trong những thành phần quan trọng trong website đó chính là liên kết (link). Khi xây dựng website, thông thường trang chủ sẽ chứa liên kết dẫn người dùng đến các trang thông tin chưa dữ liệu, các trang chứa dữ liệu hoặc sẽ liên kết với trang thông tin khác, hoặc là với trang chủ.

Hình ‑ - Hướng di chuyển của người dùng qua liên kết

Đi theo các liên kết trên website, người dùng sẽ truy cập được vào các trang nội dung mà website muốn gửi đến người dùng. Chính vì vậy, hệ thống sẽ cần một bộ nhớ stack để lưu vết liên kết và xử lý liên kết theo thứ tự LIFO[[19]](#footnote-19) tương tự như trình tự một người dùng thông thường truy cập vào website.

#### Khái niệm

Zie Stack hay gọi tắt là Stack có nhiệm vụ phân phối các liên kết của các trang web cần thu thập đến Crawler, đồng thời Stack sẽ nhận lại các liên kết mới từ Crawler sau khi Crawler hoàn tất xử lý một trang web, các liên kết mới sẽ được nạp vào Stack và sẽ được xử lý theo trình tự LIFO. Quá trình sản sinh ra liên kết mới từ Crawler gọi là Spreading.

Từ những dữ liệu được thu thập bởi Crawler (thời gian, mức độ quan trọng, khả năng sinh mới liên kết, khả năng sinh mới dữ liệu…), Stack sẽ sắp xếp các liên kết theo một thứ tự nhất định, phân loại các liên kết và sẽ gửi lại cho Crawler xử lý các liên kết kế tiếp.

Stack không chỉ đơn giản hoạt động theo nguyên tắc LIFO mà Stack ở đây có một sự thông minh, từ những thông tin được nạp vào cuối cùng có thể sẽ thay đổi thứ tự của những liên kết đã được nạp trước. Các liên kết dẫn tới dữ liệu không cần thiết, hoặc được dự đoán là chứa dữ liệu quá cũ kỹ sẽ được Stack đẩy càng về sau.

Hình ‑ - Quy trình hoạt động của Zie Stack

### Zie Crawler

#### Ý tưởng:

Mỗi website hệ thống sẽ cần có một Crawler khác nhau để thu thập dữ liệu, vì cấu trúc của các website khác nhau là khác nhau. Để có khả năng thu thập được cả thị trường, Zie cần đạt được một số lượng Crawler đáng kể và sẽ sinh mới hoặc cập nhật lại liên tục.

Chính vì thế, công việc lập trình Crawler là lâu dài và cần nhu cầu số lượng, đơn giản hóa công việc lập trình Crawler sẽ giảm bớt chi phí đào tạo, chi phí lập trình, giảm bớt lỗi phát sinh bởi lập trình viên. Zie Crawler ra mắt như một bộ biên dịch cung cấp môi trường lập trình hoàn toàn mới mà ở đó chỉ dành cho Crawler, trong đó rất nhiều thứ đã được tự động hóa để đơn giản Crawler tới mức tuyệt đối.

#### Khái niệm:

Zie Crawler phân biệt hoàn toàn với Crawler. Crawler nằm tách rời bên ngoài Zie trong khi Zie Crawler là một thành phần thuộc Zie. Zie Crawler là bộ biên dịch các Crawler, cung cấp cho Crawler một bộ API giao tiếp với Zie và là thành phần duy nhất giúp cho Crawler có thể liên lạc với Zie (Zie Stack, Zie Data).

Zie Crawler hoạt động như một trình duyệt web. Zie Crawler có khả năng biên dịch một trang web, phân tích các thành phần của trang web và cho phép Crawler dễ dàng lấy thông tin các thành phần của một trang web hoặc gửi các gói tin đặc biệt đến website chủ nhằm mục đích vượt qua các rào cản (firewall) từ website chủ. Ngoài ra, Zie Crawler chỉ cho phép lập trình viên viết Crawler làm nhiệm vụ của Crawler, đảm bảo tách rời với hệ thống, sơ suất từ Crawler sẽ không ảnh hưởng tới hệ thống và bất cứ sự xâm nhập nào từ Crawler đều là không tưởng.

Zie Crawler còn đảm nhận phần lớn các xử lý dữ liệu như:

* Kiểm tra khả năng lỗi dữ liệu.
* Xác định sự tương quan giữa các dữ liệu thu thập được (xác định một bài viết thuộc chủ đề nào, website nào, là bài viết trả lời cho bài viết nào…)
* Xác định ngôn ngữ được sử dụng trong website. Đây là thành phần có chức năng mạnh mẽ và mang lại lợi ích cực kỳ to lớn cho hệ thống.

### Zie Data

#### Ý tưởng:

Ý tưởng duy nhất đưa đến sự ra đời của Zie Data là phân chia xử lý, module hóa chức năng hệ thống.

#### Khái niệm:

Zie Data hay gọi tắt là Data là thành phần tương tự như Zie Stack nhưng chỉ có duy nhất một chức năng là nhận các dữ liệu về bài viết thu thập được bởi Crawler.

Các dữ liệu nhận được đều đã qua xử lý bởi Crawler nên nhiệm vụ của Data là lưu trữ, phân phối đến các cơ sở dữ liệu phân tán. Ngoài ra Data là thành phần đảm bảo tính nhất quán của dữ liệu, Data sẽ đảm bảo không có bất kỳ dữ liệu trùng lắp nào được lưu trữ lại hai lần.

### Zie Reporter

#### Ý tưởng:

Thay cho việc để hệ thống Front-end truy cập trực tiếp vào dữ liệu, việc xây dựng một module trong suốt giữa Front-end và dữ liệu là cần thiết không chỉ nhằm mục đích bảo mật mà còn nhằm mục đích giảm tải xử lý cho Front-end, giảm chi phí xây dựng Front-end và tăng cường sự chủ động của hệ thống trong việc trích xuất dữ liệu.

#### Khái niệm

Zie Reporter hay gọi tắt là Reporter là trung gian giữa Zie và Front-end, Reporter sẽ thể hiện dữ liệu ở dạng Front-end có thể hiểu được và nhận query từ Front-end, biên dịch query từ Front-end để gửi đến hệ thống truy xuất dữ liệu.

### Tổng quan về Zie

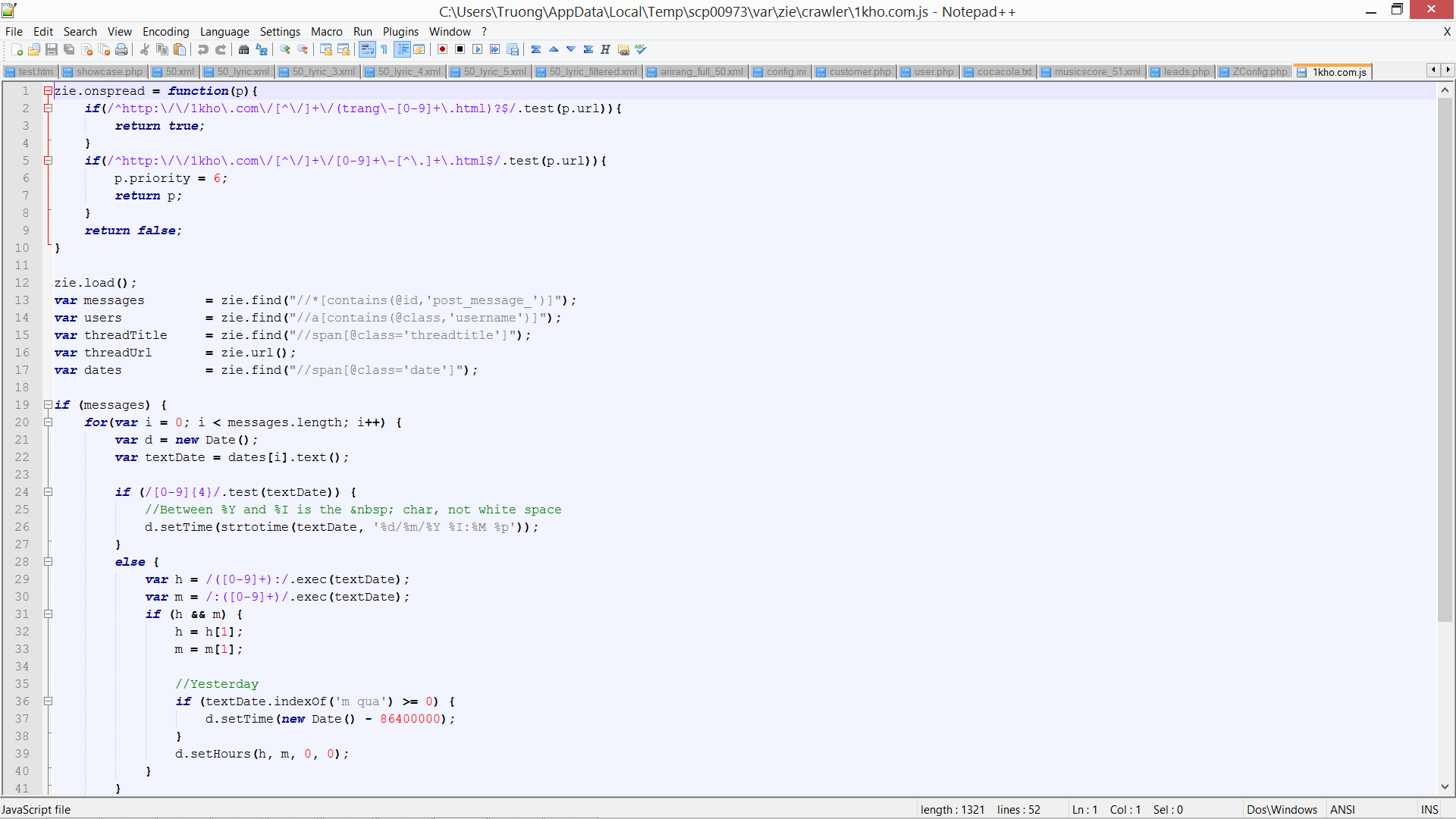
Hình ‑ - Sơ đồ quan hệ đơn giản giữa các thành phần trong Zie

## Crawler

Crawler là tập hợp tất cả các đoạn mã xử lý các website khác nhau để từ liên kết nhận được từ Stack, Crawler sẽ trích xuất được dữ liệu cho Data và nạp lại các liên kết mới cho Stack.

Crawler là thành phần đòi hỏi độ rộng do có rất nhiều, rất nhiều các website cần phải xử lý. Chính vì vậy, Crawler đã được hỗ trợ từ Zie và Crawler sẽ được hỗ trợ một ngôn ngữ lập trình riêng được Zie hiểu mà không cần phải biên dịch, ngôn ngữ lập trình Crawler sẽ đơn giản hóa các thao tác thông thường của một Crawler. Theo các tiêu chí trên, ngôn ngữ lập trình cho Crawler nên là một ngôn ngữ dạng Script, có tính tùy biến cao, uyển chuyển, phù hợp với môi trường xử lý HTML, JavaScript là những ngôn ngữ phổ biến để thiết kế website. Trong mô hình của hệ thống hiện tại, nhóm sử dụng ngôn ngữ lập trình JavaScript làm cơ sở phát triển Crawler.

Hiện tại, Crawler là thụ động, chỉ được gọi bởi Zie khi cần thiết và sau khi hoàn thành nhiệm vụ, Crawler cũng sẽ được tắt đi và ở trang thái chờ.



Hình ‑ – Một Crawler được viết bằng API của Zie

## Front-end

Front-end là thành phần duy nhất có giao diện cho phép người dùng thực hiện trích xuất, thống kê lại dữ liệu có được từ quá trình thu thập của Zie.

## Thiết kế dữ liệu

### Table authors

Lưu trữ thông tin người gửi một bài viết trên internet.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | bigint | Yes | No |
| fullname | character varying(500) | No | No |
| url | bigint | No | No |
| min\_date | date | Yes | No |
| max\_date | date | Yes | No |
| source | bigint | Yes | No |

Bảng ‑ – Cấu trúc table authors

### Table contents

Lưu trữ lại nội dung bài viết trên internet, ngôn ngữ sử dụng, tỷ lệ các ngôn ngữ sử dụng. Đồng thời lưu trữ lại mã hash để đối chiếu.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | bigint | Yes | No |
| max\_date | date | No | No |
| min\_date | date | No | No |
| content | text | No | No |
| lang | smallint[] | Yes | No |
| percent | smallint[] | Yes | No |
| hash | bytea | No | No |
| primitive | bytea | No | No |
| score | real[] | Yes | No |

Bảng ‑ - Cấu trúc table contents

### Table languages

Danh sách các ngôn ngữ hỗ trợ bởi Zie.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | integer | Yes | Yes |
| code | character varying(10) | Yes | No |

Bảng ‑ - Cấu trúc table languages

### Table posts

Danh sách các bài viết, bao gồm thông tin về nội dung, ngày xử lý, ngày viết bài, tác giả viết bài. Đây là bảng kết nối tất cả các bảng thông tin khác. Và cũng là thành phần quan trọng để tính toán báo cáo gửi đến cho Front-end.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | bigint | Yes | No |
| liked | smallint | Yes | No |
| title | character varying(255) | No | No |
| author | bigint | Yes | No |
| date | timestamp without time zone | Yes | No |
| content | bigint | Yes | No |
| thread | bigint | Yes | No |
| url | bigint | Yes | No |
| crawl | timestamp with time zone | No | No |

Bảng ‑ - Cấu trúc table posts

### Table sources

Danh sách các nguồn website được hỗ trợ bởi hệ thống. Đồng thời, bảng còn có trách nhiệm lưu trữ lại những lần xử lý với các website này để thực hiện những xử lý xoay vòng, tránh làm quá tải website nguồn.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | bigint | Yes | Yes |
| domain | character varying | Yes | No |
| state | smallint | Yes | No |
| crawl | timestamp without time zone | Yes | No |
| update | timestamp without time zone | Yes | No |
| updatable | boolean | Yes | No |
| type | smallint | Yes | No |
| out | boolean | Yes | No |
| ccycle | interval | No | No |
| ucycle | interval | No | No |
| cycle | smallint | Yes | No |
| cpop | smallint | No | No |
| upop | smallint | No | No |

Bảng ‑ - Cấu trúc table sources

### Table stacks

Bảng stacks là dữ liệu được xây dựng và sắp xếp bởi Zie Stack. Lưu trữ các thông tin về thời gian dữ liệu của liên kết, nguồn gốc sản sinh ra liên kết. Số lần xử lý liên kết và lỗi xử lý gửi đến từ Crawler.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | bigint | Yes | No |
| push | timestamp without time zone | No | No |
| pop | timestamp without time zone | No | No |
| popped | integer | Yes | No |
| priority | smallint | Yes | No |
| err | boolean | Yes | No |
| err\_message | text | No | No |
| url | bigint | Yes | No |
| parent | bigint | No | No |
| data | timestamp without time zone | Yes | No |
| fresh | interval | Yes | No |
| updatable | boolean | Yes | No |

Bảng ‑ - Cấu trúc table stacks

### Table threads

Mỗi bài viết được phân định thuộc vào một thread hay còn gọi là một luồng. Mỗi luồng sẽ có một bài viết sinh ra luồng gọi là Lead (Original Post).

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | bigint | Yes | No |
| title | character varying(1000) | No | No |
| url | bigint | No | No |
| lead | bigint | No | No |
| min\_date | date | Yes | No |
| max\_date | date | Yes | No |
| lang | smallint[] | Yes | No |

Bảng ‑ - Cấu trúc table threads

### Table urls

Lưu trữ các liên kết được sử dụng, bao gồm các liên kết của bài viết, của thread và cả của stack. Mỗi liên kết sẽ được xác định bằng một mã hash đảm bảo không trùng lấp.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Data type | Not Null? | Primary key? |
| id | bigint | Yes | No |
| url | character varying(3000) | No | No |
| min\_date | date | No | No |
| max\_date | date | No | No |
| hash | bytea | Yes | No |
| source | bigint | Yes | No |
| primitive | bytea | No | No |

Bảng ‑ - Cấu trúc table urls

## Kỹ thuật phân chia dữ liệu

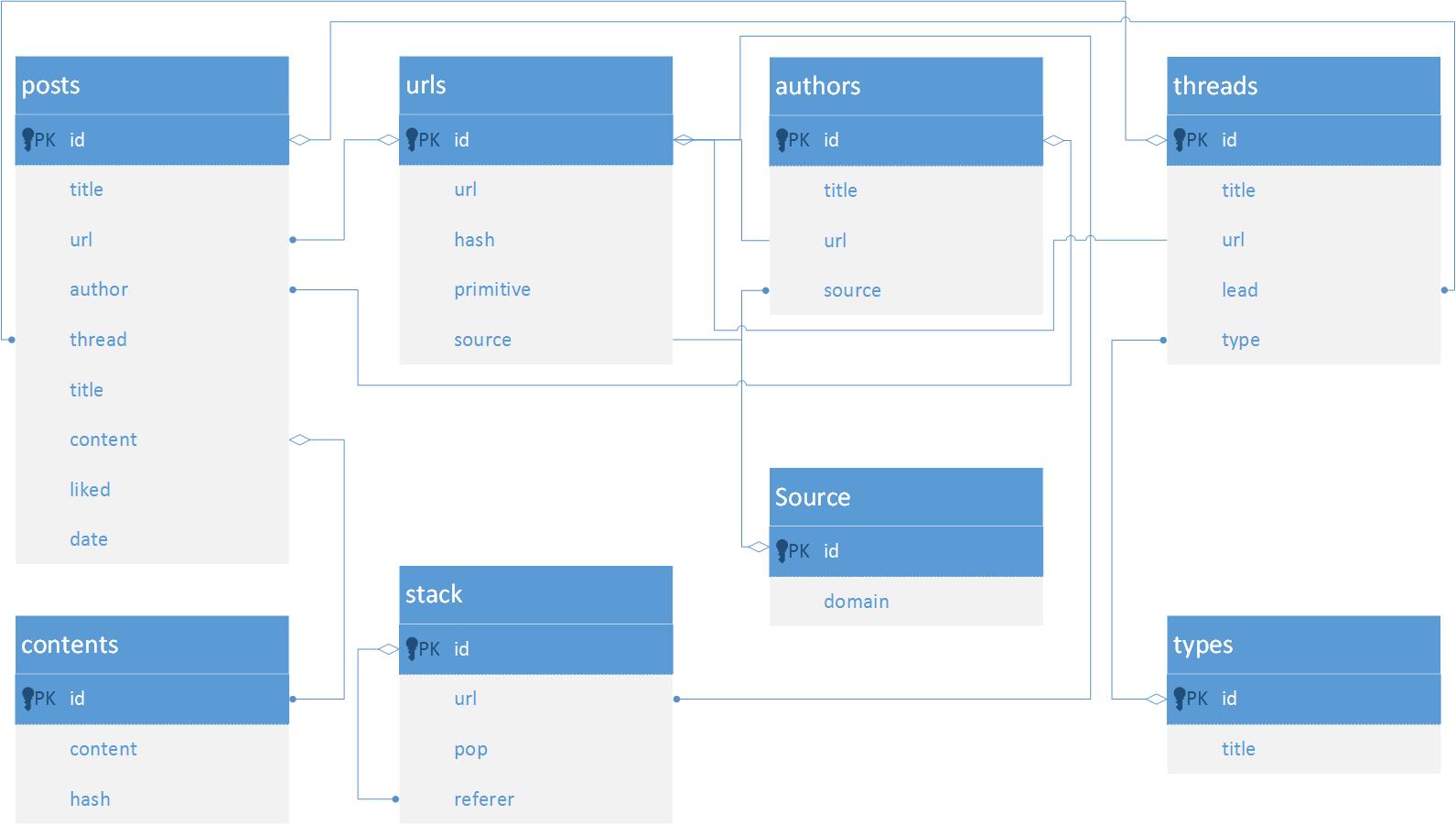
Ứng dụng chức năng Partitioning[[20]](#footnote-20) trong cơ sở dữ liệu. Ở mỗi bảng dữ liệu đều có chứa hai cột min\_date và max\_date lưu trữ thời gian lâu nhất và mới nhất mà thành phần đó được sử dụng. Các bảng dữ liệu sẽ được phân chia ra theo max\_date để giảm kích thược các bảng, đồng thời tăng tốc độ xử lý vì phải truy vấn trên một phân vùng của dữ liệu thay cho toàn bộ dữ liệu.

Ví dụ: table posts sẽ được chia ra làm 3 table

* posts\_2011: lưu trữ những thảo luận có max\_date nhỏ hơn 01-01-2012.
* posts\_2012: lưu trữ những thảo luận có max\_date nhỏ hơn 01-01-2013.
* posts\_2013: lưu trữ những thảo luận có max\_date nhỏ hơn 01-01-2014.

Khi truy xuất tới một loạt các thảo luận trong một khoảng thời gian, hệ thống sẽ có thể nhanh chóng xác định phân vùng của dữ liệu cần thiết và truy xuất trực tiếp trên đó thay cho việc phải truy xuất tất cả dữ liệu.

## Sơ đồ quan hệ



Hình ‑ - Sơ đồ quan hệ

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết luận

Từ những cấu trúc, mô hình và phương pháp tiếp cận nêu trên, nhóm đã xây dựng thành công hệ thống thu thập dữ liệu với độ phủ đạt khoảng 200 website vừa và lớn tại Việt Nam. Hệ thống có khả năng thu thập dữ liệu từ các kênh mạng xã hội lớn như Facebook, Twitter…

Hiện nay, với cấu hình phần cứng:

* CPU: 2 × Intel Xeon E5-2620 (2 Ghz, 6 cores).
* RAM: 4 × 8GB RAM.
* Hardisk: 4 × SSD Intel 335 180 GB – Raid 10.
* Internet Bandwidth: 20 Mbps Up/20 Mbps Down.

Hệ thống có thể thu thập 100.000 thảo luận mới/phút (chưa bao gồm cập nhật lại dữ liệu đã được thu thập). Tốc độ truy xuất dữ liệu đạt dưới 10 giây cho một báo cáo lớn và dưới 1 giây cho một báo cáo trung bình. Tính đến tháng 1 năm 2014, sau 8 tháng xây dựng và thử nghiệm, hệ thống hiện tại đang lưu trữ dữ liệu của hơn 100.000.000 bài viết và tập trung chủ yếu là dữ liệu trong năm 2013.

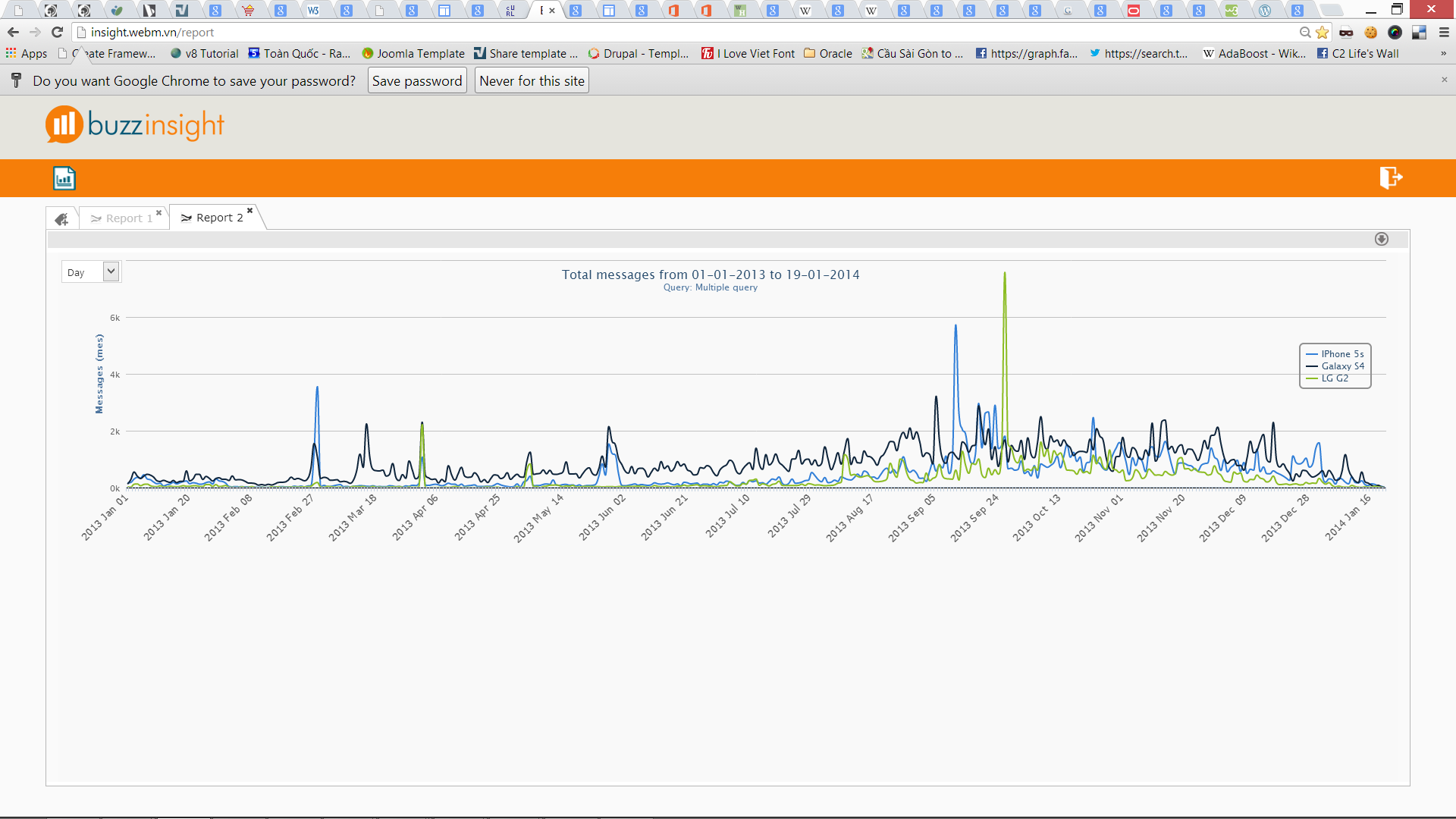
Hệ thống không có tích hợp bất kỳ ứng dụng thương mại nào nên chi phí phát triển là rất thấp. Hoàn toàn có lợi thế cạnh tranh với các đối thủ trên thị trường và hoàn toàn nằm trong khả năng của các nhà nghiên cứu thị trường thậm chí là cá nhân.

Nhóm đã cài đặt hệ thống tại địa chỉ <http://insight.webm.vn> với giao diện front-end đơn giản nhằm mục đích thử nghiệm hệ thống và đặc biệt hiện nay đang được sử dụng bởi công ty phân tích thị trường WebM Corp, đồng thời, hệ thống cũng đã nhận được một số thành tích:

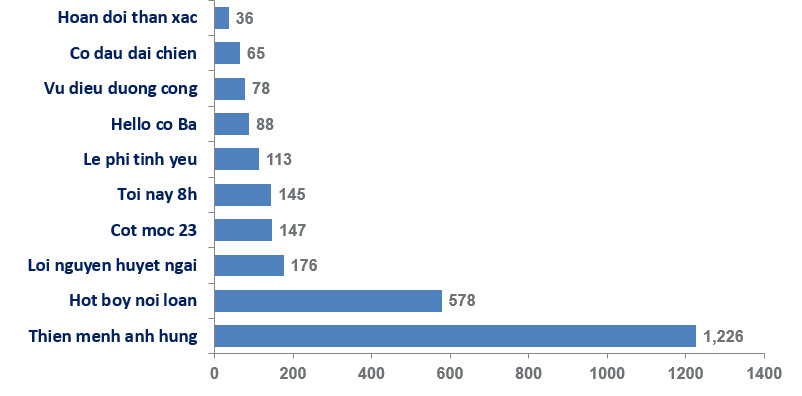
* Giải nhì trong hội nghị khoa học trẻ 2013 tại trường Đại Học Công Nghệ Thông Tin – ĐH Quốc Gia TP Hồ Chí Minh.
* Giải ba cuộc thi sinh viên nghiên cứu khoa học Eureka 2013 của thành phố Hồ Chí Minh.

Về kỹ thuật, hệ thống tính đến tháng 1 năm 2014 đang sử dụng hệ quản trị cơ sở dữ liệu mở PostgreSQL 9.3, hệ thống tìm kiếm mở Sphinx 2.1.4, xây dựng bằng ngôn ngữ C++ trên hệ điều hành CentOS 6.4. Hệ thống sử dụng 100% thư viện nguồn mở.

Dưới đây là hai mẫu báo cáo được xây dựng nhờ dữ liệu phân tích từ hệ thống:



Hình ‑ - Mức độ quan tâm của người dùng với 3 sản phẩm Iphone 5S, Galaxy S4 và LG G2.



Hình ‑ - Mức độ quan tâm của người dùng internet đến phim tết 2013

## Hướng phát triển

Trong tương lai, hệ thống sẽ được tập trung cái thiện hiệu suất ngày một tốt hơn và phát triển các module nâng cao như:

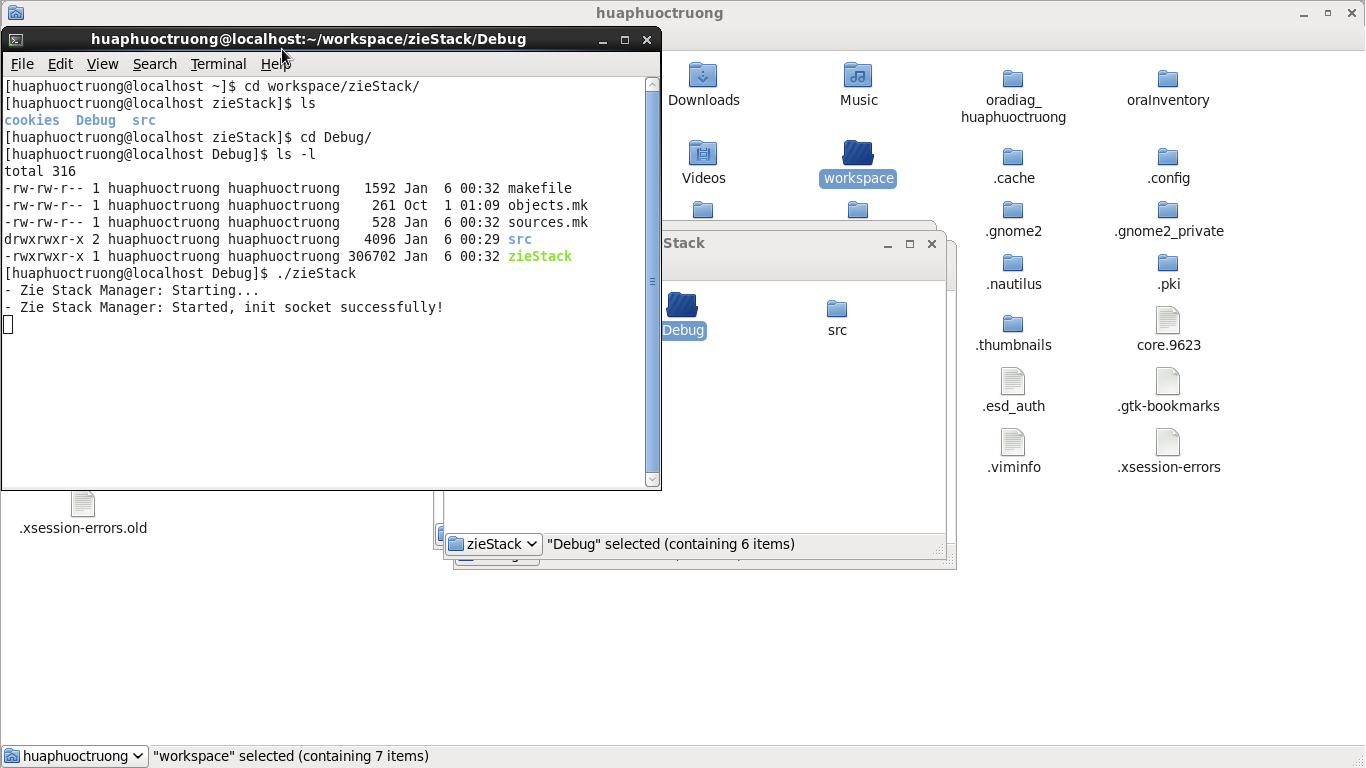
* Phân loại thảo luận dựa trên các dữ liệu thu thập được, các nhà nghiên cứu sẽ có thể truy xuất được những dữ liệu có giá trị cao nhất.
* Nhận diện cảm xúc của thảo luận.
* Thu thập thông tin người dùng.

## 

# CÀI ĐẶT MINH HỌA

## Zie Stack

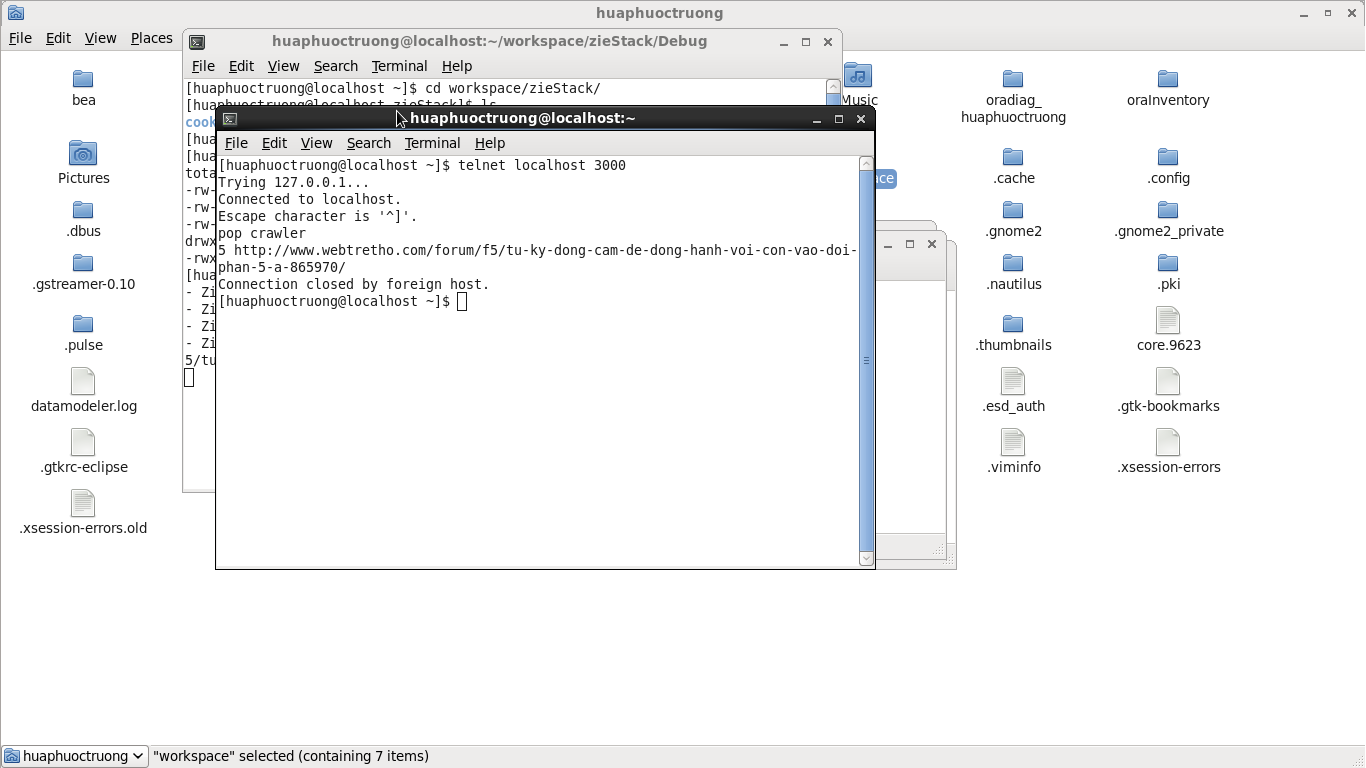
Khởi động Zie Stack bằng cách chạy lệnh ZieStack từ terminal.



Hình ‑ - Khởi động Stack

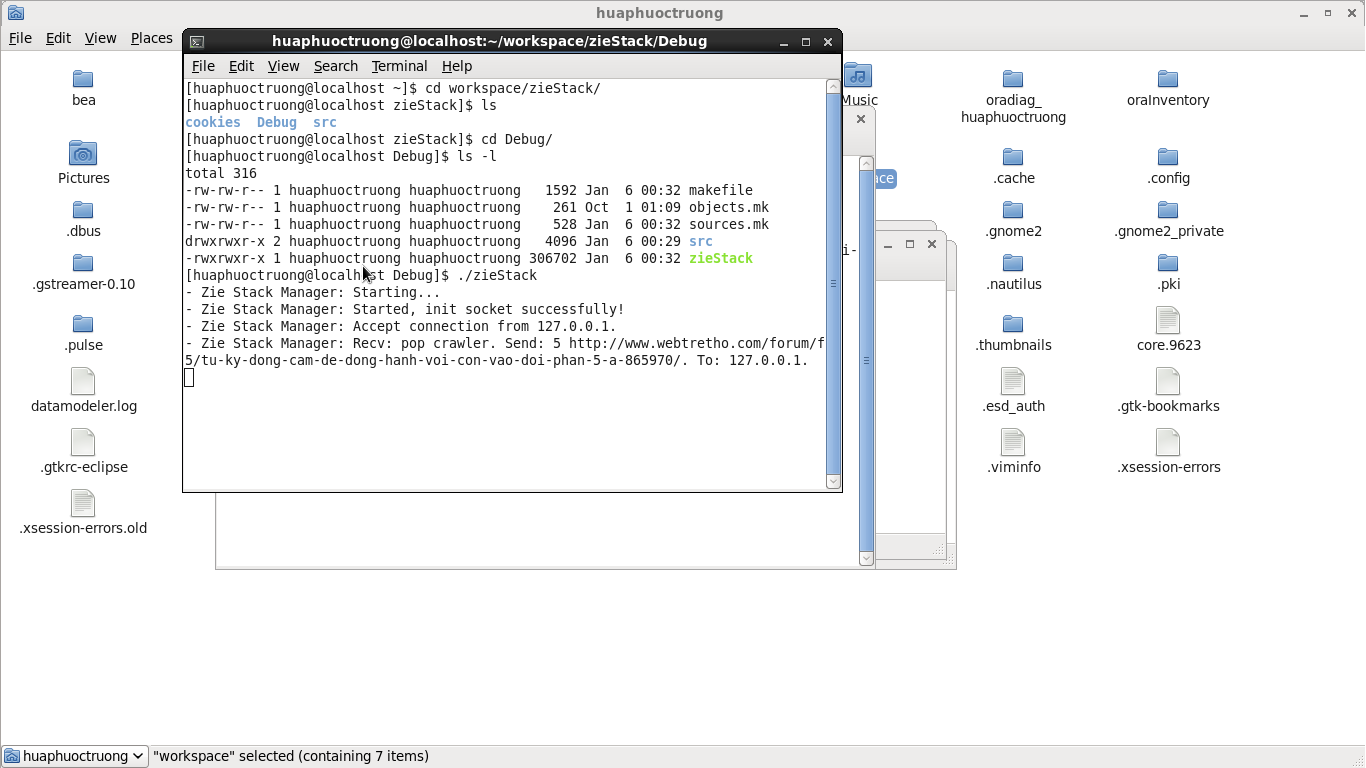
Zie Stack sẽ hiển thị thông báo hoàn tất nếu khởi động thành công, ngược lại, sẽ có thông báo lỗi tùy theo từng trường hợp.

Lấy một url từ stack bằng Zie Stack bằng cách dùng lệnh telnet, kết nối tới port của Zie Stack, mặc định là 3000.



Hình ‑ - Pop Zie Stack

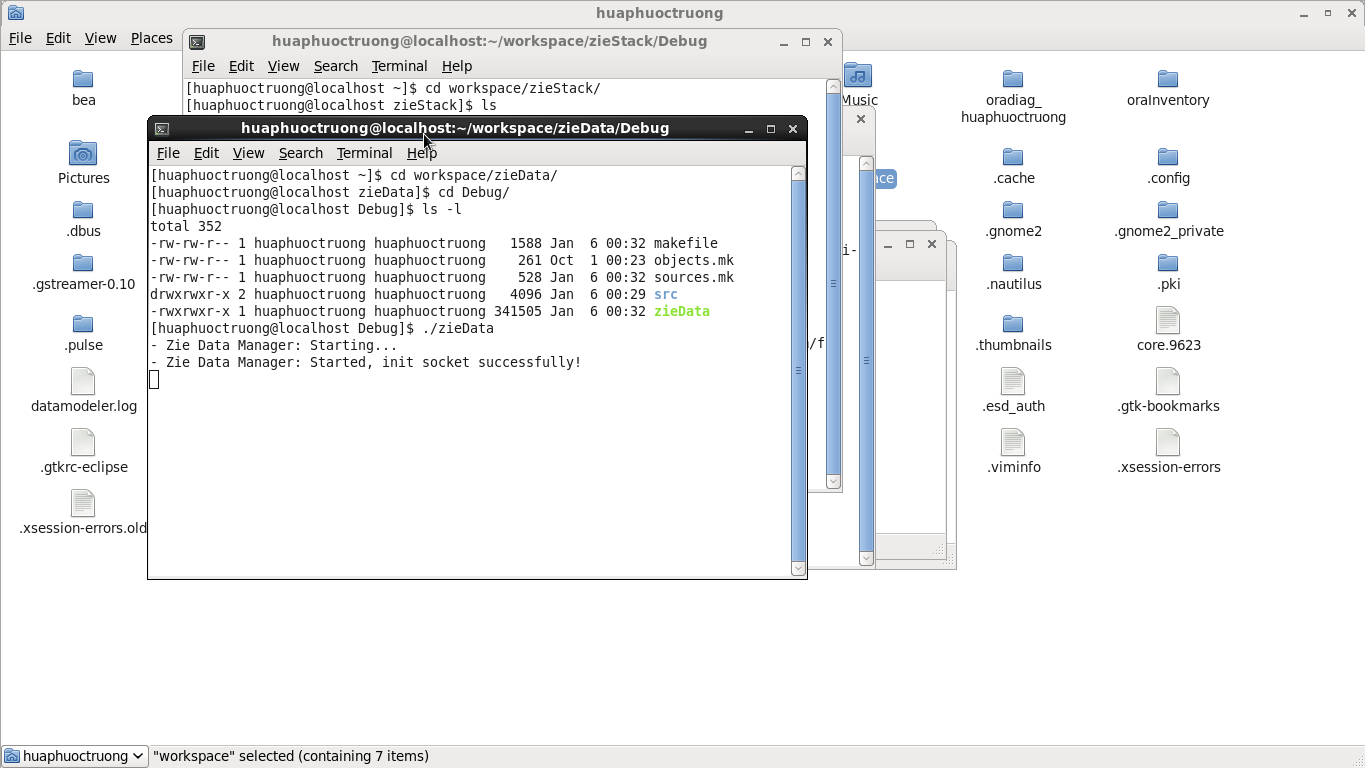
Ngoài ra, khi có truy xuất từ Client hoặc người dùng, Zie Stack cũng sẽ dùng cơ chế Log để thông báo lại tiến trình cho người quản trị:



Hình ‑ - Cơ chế Log trong Zie Stack

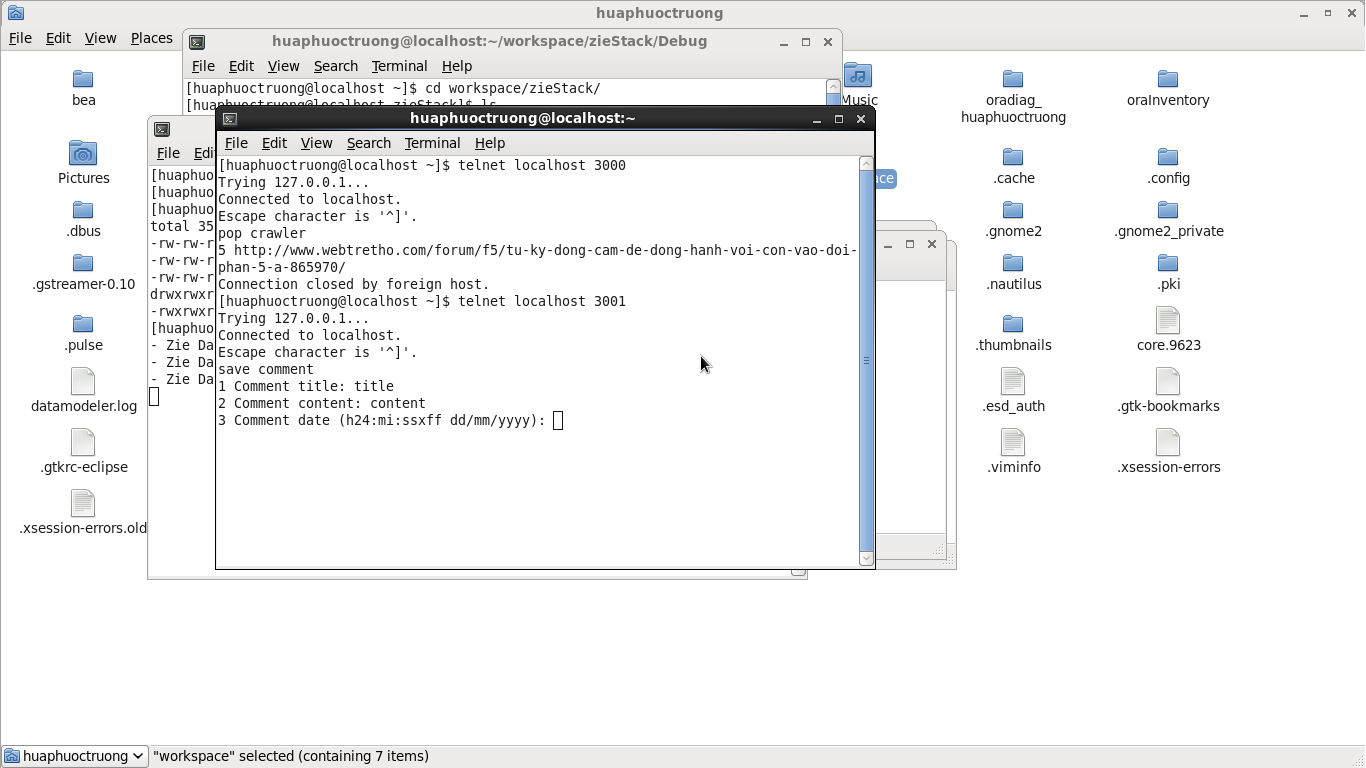
## Zie Data

Tương tự Zie Stack, khởi động Zie Data thông qua lệnh ZieData. Mặc định, Zie Data sẽ dùng port 3001.



Hình ‑ - Khởi động Zie Data

Sử dụng lệnh “save comment” để tiến hành lưu thảo luận từ người dùng. Khi nhập thông qua terminal, Zie Data sẽ có thông báo nhắc nhở người dùng cần điền thông tin gì tiếp theo.

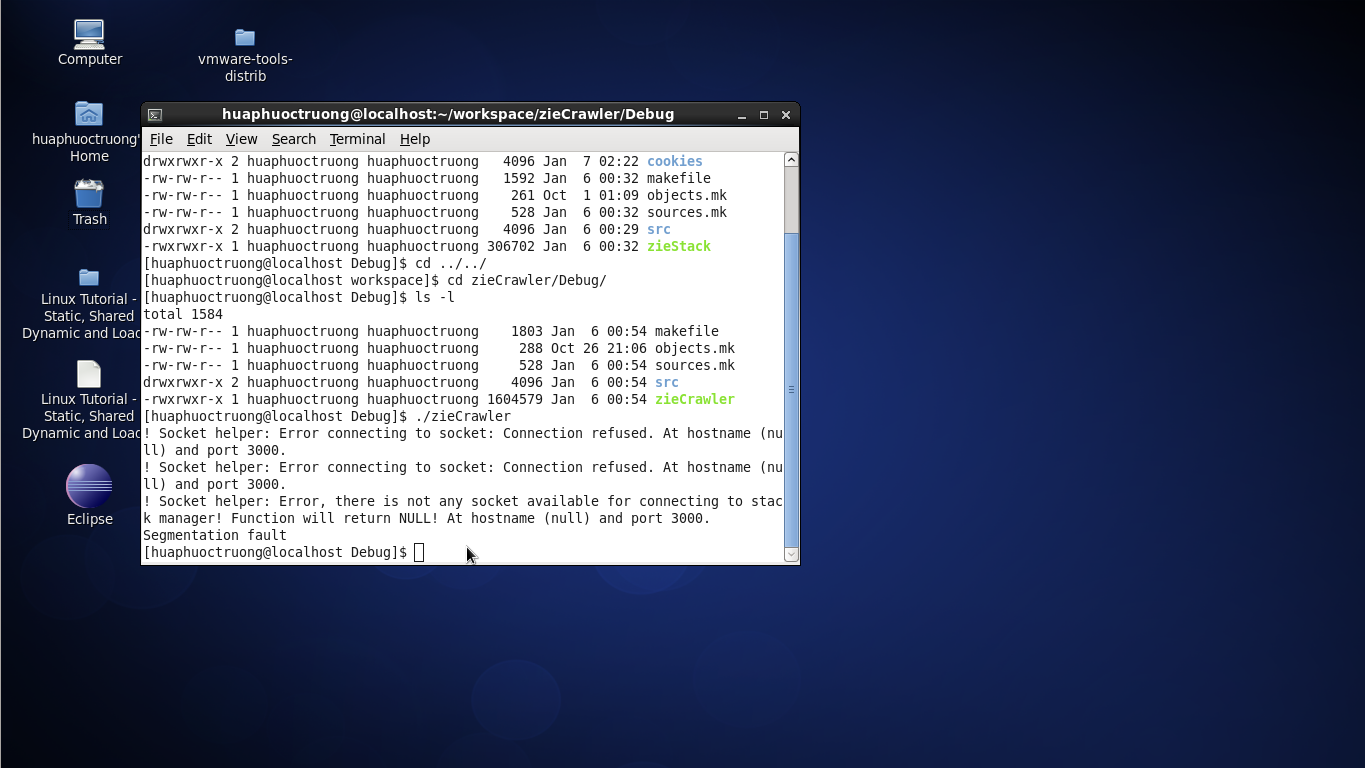


Hình ‑ - Lệnh save của Zie Data

## Zie Crawler

Khởi động Zie Crawler bằng cách dùng lệnh ZieCrawler. Khi khởi động không thành công, Zie Crawler sẽ có thông báo rõ ràng lý do vì sao không khởi động được.

Ví dụ dưới đây là trường hợp Zie Crawler không thể chạy do Zie Data và Zie Stack chưa được khởi động:



Hình ‑ - Zie Crawler thông báo không tìm thấy Stack và Data



Hình ‑ - Crawler chạy thành công

# TÀI LIỆU THAM KHẢO

1. Carlos Castillo, Efective Web Crawling, University of Chile, Chile, 2004.
2. Dr Lyn McCredden, Research and Graduate Studies Convenor, “Influence of the World Wide Web on literature”, School of Literary and Communication Studies - Deakin University, Australia, 1997.
3. Junghoo Cho, Crawling the web: discovery and maintenance of large-scale web data, Stanford University, United State, 2001.
4. Katarzyna Musiał, “Recomemdation system for online social network”, School of Engineering, Sweden, 2006.
5. Sergey Brin and Lawrence Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, Computer Science Department, Stanford University, Stanford, CA 94305.
6. Fork, Exec and Process control

<http://www.yolinux.com/TUTORIALS/ForkExecProcesses.html>

1. Linux Howtos: C/C++ -> Sockets Tutorial

<http://www.linuxhowtos.org/C_C++/socket.htm>

1. PostgreSQL libpq

<http://www.postgresql.org/docs/8.1/static/libpq.html>

1. Sphinx

<http://sphinxsearch.com/docs/current.html>

1. The global structure of an HTML document

<http://www.w3.org/TR/html401/struct/global.html>

1. The global structure of an HTML document

<http://www.w3.org/TR/html401/struct/global.html>

# PHỤ LỤC 1: POSTGRESQL

Zie sử dụng hệ quản trị cơ sở dữ liệu PostgreSQL 9.3 làm nền tảng.



PostgreSQL là hệ quản trị cơ sở dữ liệu nguồn mở, miễn phí và được hỗ trợ bởi nhiều nền tảng khác nhau, đồng thời Postgres còn là hệ quản trị cơ sở dữ liệu được khuyến nghị sử dụng bởi nhà nước và được cộng đồng internet xem là sự thay thế đáng giá cho Oracle.

PostgreSQL có thể sử dụng, cài đặt dễ dàng trên Linux, chiếm ít tài nguyên hệ thống hơn Oracle, tuy nhiên, điểm khiếm khuyết của Postgres là không có tính năng Logging như Oracle có thể bảo vệ dữ liệu an toàn. Nhưng tính năng này hoàn toàn có thể thay thế bằng một chế độ Auto backup.

# PHỤ LỤC 2: COMPACT LANGUAGE DETECTOR



Đây là thành phần phát hiện ngôn ngữ của trình duyệt web Chrome, hoàn toàn mở và có tốc độ xử lý rất nhanh, sai số chấp nhận được. Có tổ chức đơn giản, dễ dàng phát triển.

# PHỤ LỤC 3: PQ

PQ là bộ API viết bằng C hỗ trợ giao tiếp với cơ sở dữ liệu Postgres và hoàn toàn mở, tương đương với OCCI trong Oracle. Tuy không thực sự mạnh mẽ nhưng PQ đáp ứng được các yêu cầu tối thiểu của Zie.

# PHỤ LỤC 4: SPHINX CLIENT



Hệ thống sử dụng một bộ Index riêng tác rời hoàn toàn với dữ liệu là Sphinx. Sphinx mang lại tốc độ truy xuất trên cả tuyệt vời. Trong đó Sphinx Client là thư viện hỗ trợ giao tiếp với Sphinx trên C và hoàn toàn mở.

# PHỤ LỤC 5: ICU



ICU (International Components for Unicode) là thư viện được viết trên C/C++ hỗ trợ phần lớn các chuẩn mã hóa trong Unicode. ICU hiện đang được sử dụng trong các sản phẩm của Apache, IBM, Google… như: DB2, Lotus, Chrome, Android OS, iOS, MacOS, Safari…

# PHỤ LỤC 6: CHUẨN GIAO TIẾP TRONG ZIE

1. **Nguyên tắc giao tiếp trong Zie**

Trong Zie, các module giao tiếp với nhau qua socket. Mỗi module có thể hoạt động trên một PC khác nhau hoặc cùng một PC nhưng trên các Port khác nhau. Trong đó, ngoại trừ đặc biệt trường hợp của Zie Stack không nên khởi động nhiều instance trên các server khác nhau nhằm tránh trùng lắp thì các module khác đều có khả năng hoạt động trên nhiều PC, nhiều instance khác nhau.

Một lệnh trong Zie có cấu trúc như sau:

[Tên lệnh] [Tham số 1] [Tham số 2] […] [Tham số n]

Trường hợp tham số có chứa ký tự khoảng trắng thì cấu trúc lệnh như sau:

[Tên lệnh][CR+LF][Tham số 1][CR+LF][Tham số 2][CR+LF][Tham số 3][CR+LF]…[CR+LF][Tham số n]

Một kết quả trả về trong Zie sẽ có cấu trúc như sau:

[State number] [Thông điệp]

Trong đó:

* State number: là thông số trạng thái của kết quả, thường cho kết quả không âm khi xử lý thành công và cho kết quả âm khi có lỗi xảy ra. Mỗi giá trị State number sẽ tương đương với một kết quả khác nhau và đảm bảo không trùng lắp.
* Thông tin: là một chuỗi báo xử lý thành công, hoặc một chuỗi thông báo lỗi trong trường hợp có lỗi xảy ra trong quá trình xử lý.

5. **Zie Stack**

Mặc định, Zie Stack hoạt động trên port 3000.

Nhằm đảm bảo các Crawler không cùng xử lý một url nên cần thiết phải có một module đóng vai trò như người tổng hợp các Crawler, điều hướng các Crawler. Khi một Crawler được khởi động, Zie Crawler Stack sẽ cung cấp cho Crawler một địa chỉ mà Crawler cần đi đến để tiến hành truy vét dữ liệu.

Trong quá trình truy vét, Crawler có thể sẽ phát hiện những liên kết mới. Ở đây, Zie Crawler Stack đóng một vai trò mới là người tiếp nhận các url từ Crawler, sắp xếp, lưu trữ url vào Stack và trả lại cho Crawler khi cần thiết.

Những lệnh được Zie Crawler Stack hỗ trợ:

| Lệnh | Tham số | Ý nghĩa | Kết quả trả về |
| --- | --- | --- | --- |
| crawl | Không có. | Lấy liên kết kế tiếp cần được xử lý từ Zie Stack. | Một số nguyên báo trạng thái kết quả (state number) trong đó nếu là số dương lớn hơn không là kết quả truy vấn thành công và cũng là giá trị identity của url trong stack.  Đi kèm với state number là chuỗi url phân cách bởi khoảng trắng. |
| push | Liên kết mới cần nạp vào Zie Crawler Stack. | Nạp một liên kết mới vào Zie Stack. | Một state number đi kèm với chuỗi thông báo kết quả. Nếu state number bằng 0 tức là quá trình nạp thành công, ngược lại, nếu state number là số âm, quá trình nạp thất bại. |
| update | Không có | Lấy liên kết kế tiếp cần được cập nhật lại từ Zie Stack. | Một số nguyên báo trạng thái kết quả (state number) trong đó nếu là số dương lớn hơn không là kết quả truy vấn thành công và cũng là giá trị identity của url trong stack.  Đi kèm với state number là chuỗi url phân cách bởi khoảng trắng. |

Mã lỗi trong Zie Stack

|  |  |
| --- | --- |
| Mã lỗi | Thông điệp |
| -1 | Lệnh không đúng |
| -2 | Không thể truy vấn tới cơ sở dữ liệu |
| -3 | Không thể thoát Zie Stack |

Trong quá trình yêu cầu một metric hoặc một thống kê, sẽ có trường hợp thống kê quá lớn dẫn tới quá trình xử lý phải diễn ra trong thời gian dài và đòi hỏi các tiến trình cần phải được xử lý một cách có thứ tự. Zie Statist Stack có nhiệm vụ quản lý các yêu cầu thống kê, điều hướng Zie Statist.

1. **Zie Data**

Zie Data mặc định hoạt động trên port 3001.

Zie Data đóng vai trò là cầu nối trung gian giữa cơ sở dữ liệu và Crawler. Một Crawler khi tìm được dữ liệu phù hợp như thông tin người dùng, ý kiến, lời nhận xét… sẽ gửi tới Zie Data theo một cấu trúc định sẵn, tại đây Zie Data sẽ xử lý lại dữ liệu thô thành dữ liệu có cấu trúc và lưu trữ vào cơ sở dữ liệu.

|  |  |  |  |
| --- | --- | --- | --- |
| Lệnh | Tham số | Ý nghĩa | Kết quả trả về |
| post | Dữ liệu thô một thảo luận bao gồm các thông tin như: tiêu đề, nội dung, ngày đăng thảo luận, liên kết thảo luận, số lượt like, trang thái thảo luận banned/active, username của người đăng tin, liên kết chuyên mục của thảo luận, tiêu đề chuyên mục của thảo luận.  Mỗi thông tin sẽ được truyền cách nhau bởi cặp ký tự Return và Endline. | Thông báo cho Zie Data biết đang có dữ liệu thô của một thảo luận cần được lưu trữ vào cơ sở dữ liệu. | Kết quả cuối cùng với state number có giá trị bằng 0 khi quá trình lưu trữ thành công. Ngược lại state number sẽ mang giá trị âm.  Ngoài ra, trong các quá trình trung gian, có thể Zie Data sẽ đưa ra những thông báo nhắc nhở nếu người dùng bỏ sót dữ liệu. |

Mã lỗi và mã thông báo trong Zie Data

|  |  |
| --- | --- |
| Mã lỗi / mã thông báo | Thông điệp |
| -1 | Lệnh không đúng |
| -2 | Không thể truy cập tới cơ sở dữ liệu |
| -3 | Không đủ tham số |
| -4 | Định dạng ngày tháng năm không chính xác, và hãy nhập lại. |
| -5 | Không thể thoát Zie Data |
| 1 | Hãy nhập tiêu đề thảo luận |
| 2 | Hãy nhập nội dung thảo luận |
| 3 | Hãy nhập ngày đăng của thảo luận |
| 4 | Địa chỉ url của thảo luận |
| 5 | Số lượt người chú ý đến thảo luận này |
| 6 | Thảo luận này có đang bị cấm ở trang nguồn (0/1) |
| 7 | Tên đăng nhập của người đăng thảo luận |
| 8 | Nguồn của thảo luận |
| 9 | Tên luồng của thảo luận |
| 10 | Địa chỉ của luồng của thảo luận |

1. Nielsen: công ty nghiên cứu thị trường hàng đầu thế giới có xuất phát điểm từ Ấn Độ, hiện nay đang cung cấp dịch vụ trên 100 quốc gia với hơn 40.000 nhân viên. [↑](#footnote-ref-1)
2. Trend: thuật ngữ chuyên ngành nghiên cứu thị trường, ý chỉ xu hướng phát triển của một chiến dịch. [↑](#footnote-ref-2)
3. Thread: nhiều bài viết trong một diễn dàn có liên quan với nhau được gọi là một thread, trong nghiệp vụ phân tích, thread dùng để chỉ các bài viết đang bàn luận về cũng một chủ đề, không hạn chế là chỉ trong diễn đàn. [↑](#footnote-ref-3)
4. Social Media Type: là thuật ngữ chuyên ngành nghiên ý chỉ các loại mạng xã hội, bao gồm: diễn đàn, facebook, twitter, tin tức, rao vặt, website của các nhà cung cấp. [↑](#footnote-ref-4)
5. Oracle: hệ quản trị cơ sở dữ liệu quan hệ phổ biến trên thế giới được phát triển bởi tập đoàn Oracle với hơn 50 năm kinh nghiệm. [↑](#footnote-ref-5)
6. PostgreSQL: hệ quản trị cơ sở dữ liệu quan hệ và đối tượng nguồn mở được khoa điện toán của đại học California tại Berkeley phát triển. [↑](#footnote-ref-6)
7. NEAR: hàm tìm kiếm một từ gần một từ khác. Ví dụ word1 NEAR/3 word2 nghĩa là giữa word1 và word2 có không quá 3 từ khác. [↑](#footnote-ref-7)
8. Full-text search: là một kỹ thuật tìm kiếm trên cở sở dữ liệu là các bài viết dựa trên dữ liệu đã được index. [↑](#footnote-ref-8)
9. MySQL: là hệ quản trị cơ sở dữ liệu tự do nguồn mở phổ biến nhất thế giới bởi tốc độ truy xuất cực nhanh. [↑](#footnote-ref-9)
10. Liên kết: là cách dịch của URL (Uniform Resource Location) sang tiếng Việt, ý chỉ các anchor trong một website. [↑](#footnote-ref-10)
11. Oracle Enterprise: là một sản phẩm thuộc Oracle Database, có phí bản quyền lên tới 47.500 USD cho một processor. [↑](#footnote-ref-11)
12. Source: tập hợp các website nguồn bao gồm tất cả các loại website như mạng xã hội, diễn đàn, trang tin, trang rao vặt, website của các nhà cung cấp… [↑](#footnote-ref-12)
13. Post: là cách gọi chung của bài báo cáo này đối với các bài viết trong forum, các bài báo, các thảo luận, comment trong các mạng xã hội hay cả những feedback của người dùng. [↑](#footnote-ref-13)
14. DOM (Document Object Model): là một API thường có dạng một cây cấu trúc dữ liệu, được dùng để truy xuất các tài liệu dạng HTML và XML. [↑](#footnote-ref-14)
15. Regular Expresion: là một biểu thức thường được dùng để mô tả một mẫu chuỗi. [↑](#footnote-ref-15)
16. Meta canonical: là một tham số thuộc header của HTML, có tác dụng chỉ định một loạt các URL khách nhau thực ra chỉ là một. [↑](#footnote-ref-16)
17. Nested Set Model là cấu trúc dữ liệu dạng cây nhưng được tổ chức để phù hợp lưu trữ trong cơ sở dữ liệu quan hệ. [↑](#footnote-ref-17)
18. Chracter entity: là những ký tự hoặc nhóm ký tự dùng để thay thế một ký tự đã được sử dụng trong HTML với một ý nghĩa đặc biệt nào đó. Ví dụ: &lt; tương đương "<", &gt; tương đương ">". [↑](#footnote-ref-18)
19. LIFO: last in first out. [↑](#footnote-ref-19)
20. Partitioning: là kỹ thuật phân đoạn dữ liệu được hỗ trợ bởi các hệ quản trị cơ sở dữ liệu phức tạp. [↑](#footnote-ref-20)