

HỆ THỐNG TỔNG HỢP TIN TỨC

Sinh viên: Đào Hoàng Sơn

Cán bộ hướng dẫn: TS. Nguyễn Ngọc Hóa

Khóa luận Công nghệ Thông tin, Đại học Công nghệ, Đại học Quốc gia Hà Nội


Tháng 5 - 2012

Nội dung

- Đặt vấn đề
- Thiết kế hệ thống
- Cài đặt
- Thực nghiệm
- Kết luận

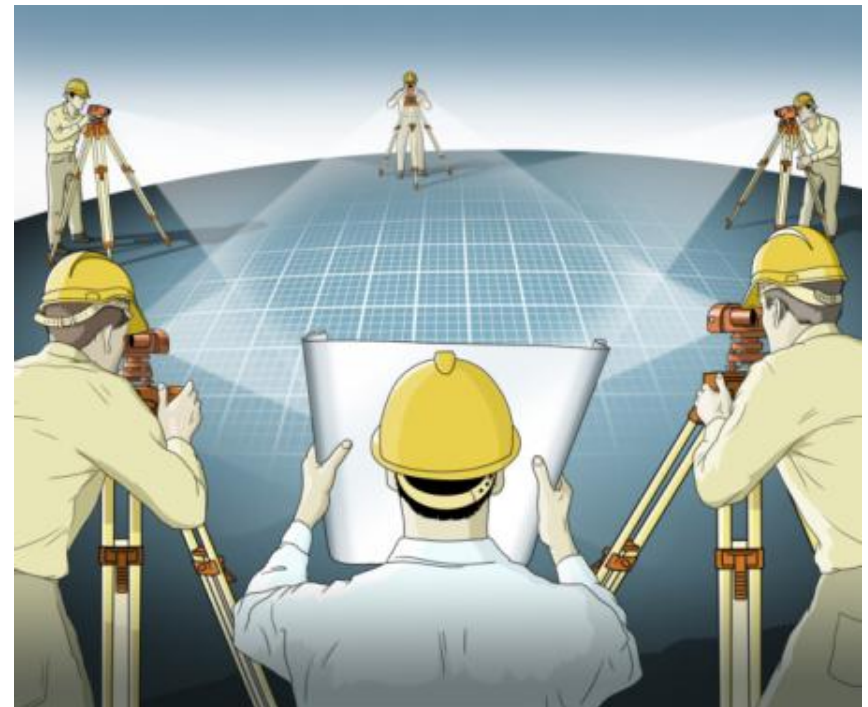
Đặt vấn đề [1/2]: Hiện trạng



- Quá nhiều trang tin, báo, tạp chí
- Thông tin trùng lặp, chồng chéo
- Khó theo dõi tin tức
- Dịch vụ hiện có:
 - BáoMới.com
 -  Starbuzz

Đặt vấn đề [2/2]: Mục tiêu

- Hệ thống tổng hợp tin tức:
 - Thu thập tin tức từ nhiều nguồn
 - Xử lý tiếng Việt
 - Thông kê xu hướng
- Horizontal scaling



Nội dung

- Đặt vấn đề
- **Thiết kế hệ thống**
- Cài đặt
- Thực nghiệm
- Kết luận

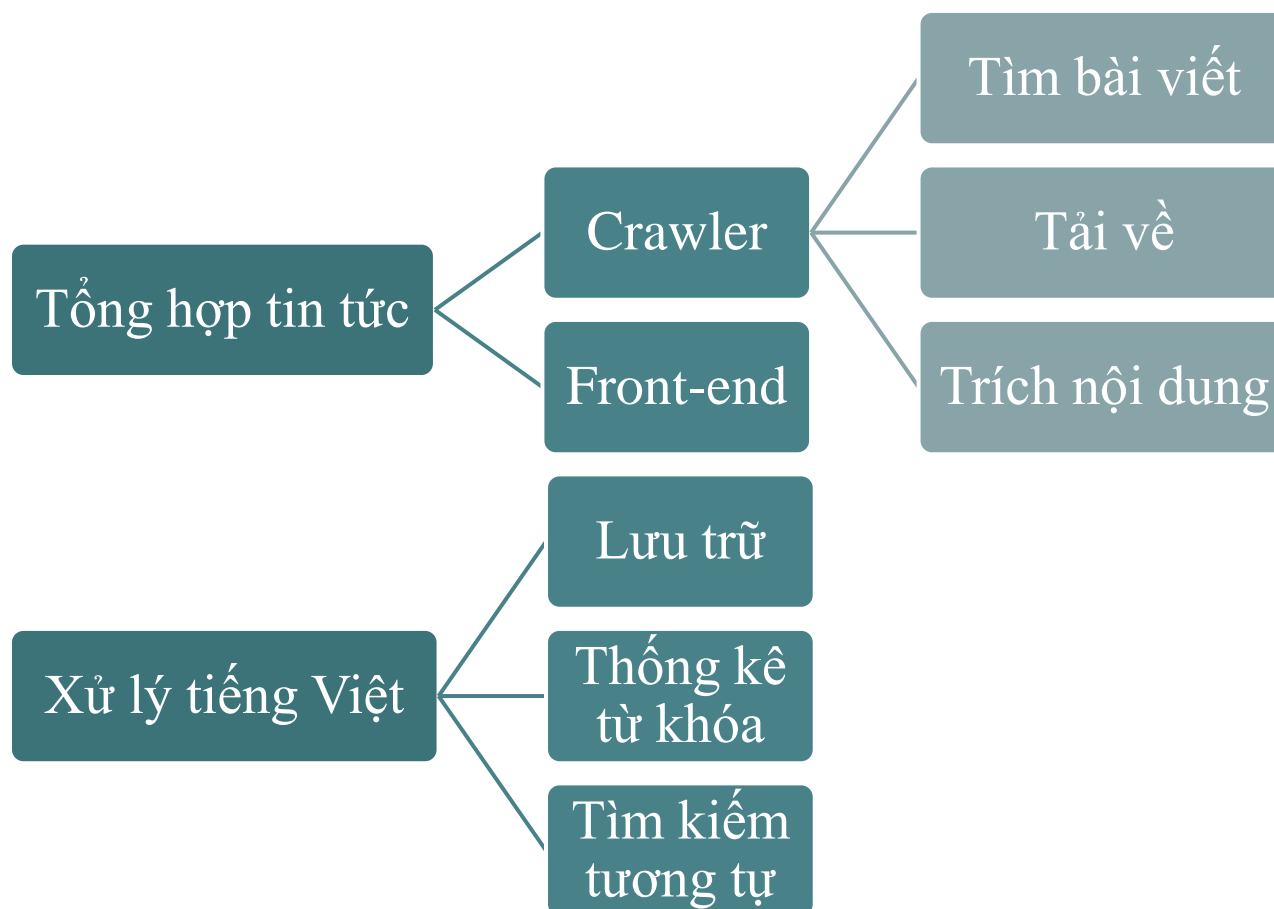
Thiết kế hệ thống [1/3]: Luồng xử lý

Tải về

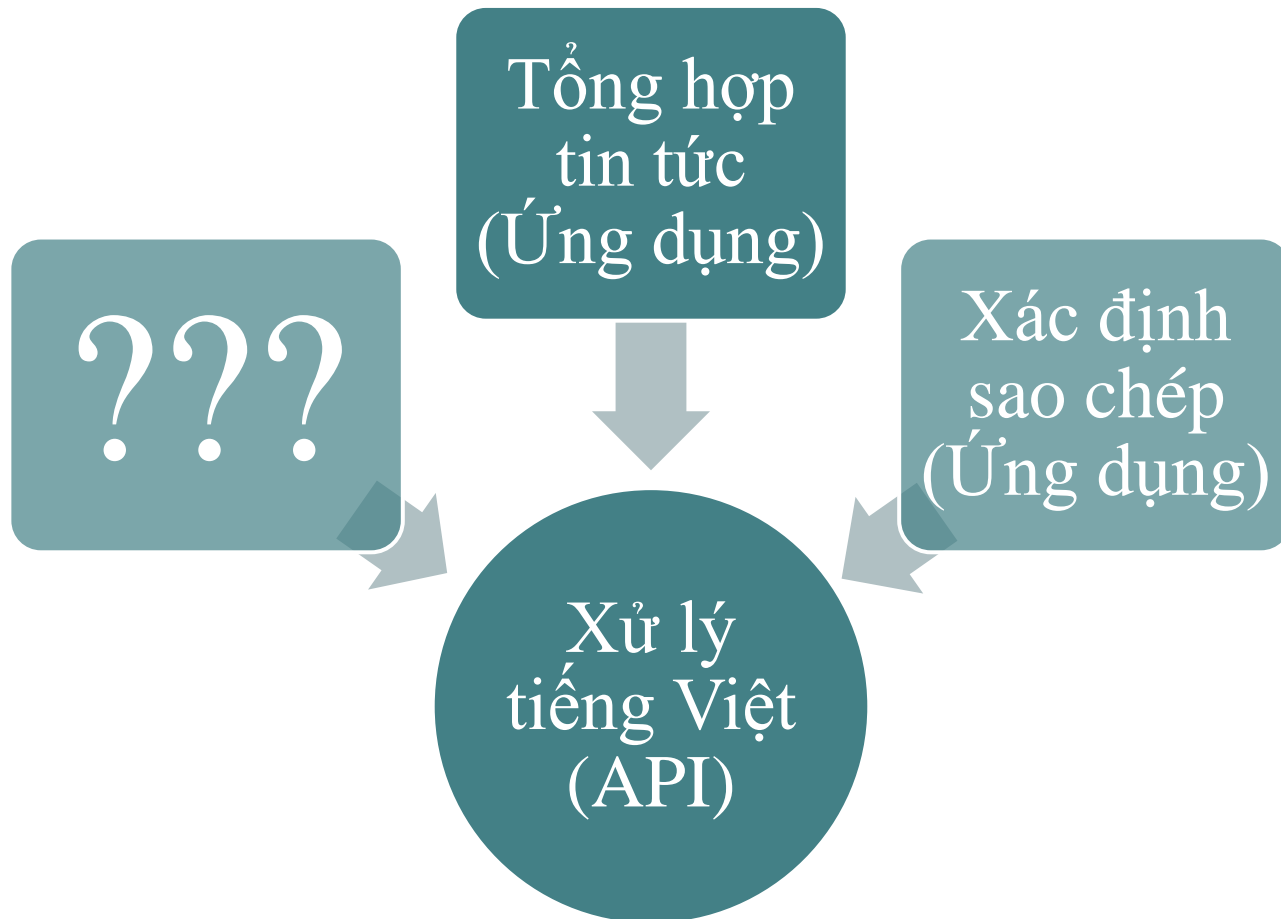
*Xử lý
tiếng Việt*

Hiển thị

Thiết kế hệ thống [2/3]: Các thành phần



Thiết kế hệ thống [3/3]: API và Ứng dụng



Nội dung

- Đặt vấn đề
- Thiết kế hệ thống
- **Cài đặt**
- Thực nghiệm
- Kết luận

Cài đặt [1/8]

- Hệ thống xử lý tiếng Việt (<http://bit.ly/koluto-github-1>)
 - Ngôn ngữ lập trình: JavaScript (Node.js)
 - Lưu trữ dữ liệu: MongoDB, Redis, MySQL
- Crawler (<http://bit.ly/koluto-github-2>)
 - Ngôn ngữ lập trình: Python
 - Lưu trữ dữ liệu: Hệ thống file, MySQL
- Front-end (<http://bit.ly/koluto-github-3>)
 - Ngôn ngữ lập trình: PHP (Yii)
 - Lưu trữ dữ liệu: MySQL, Memcached

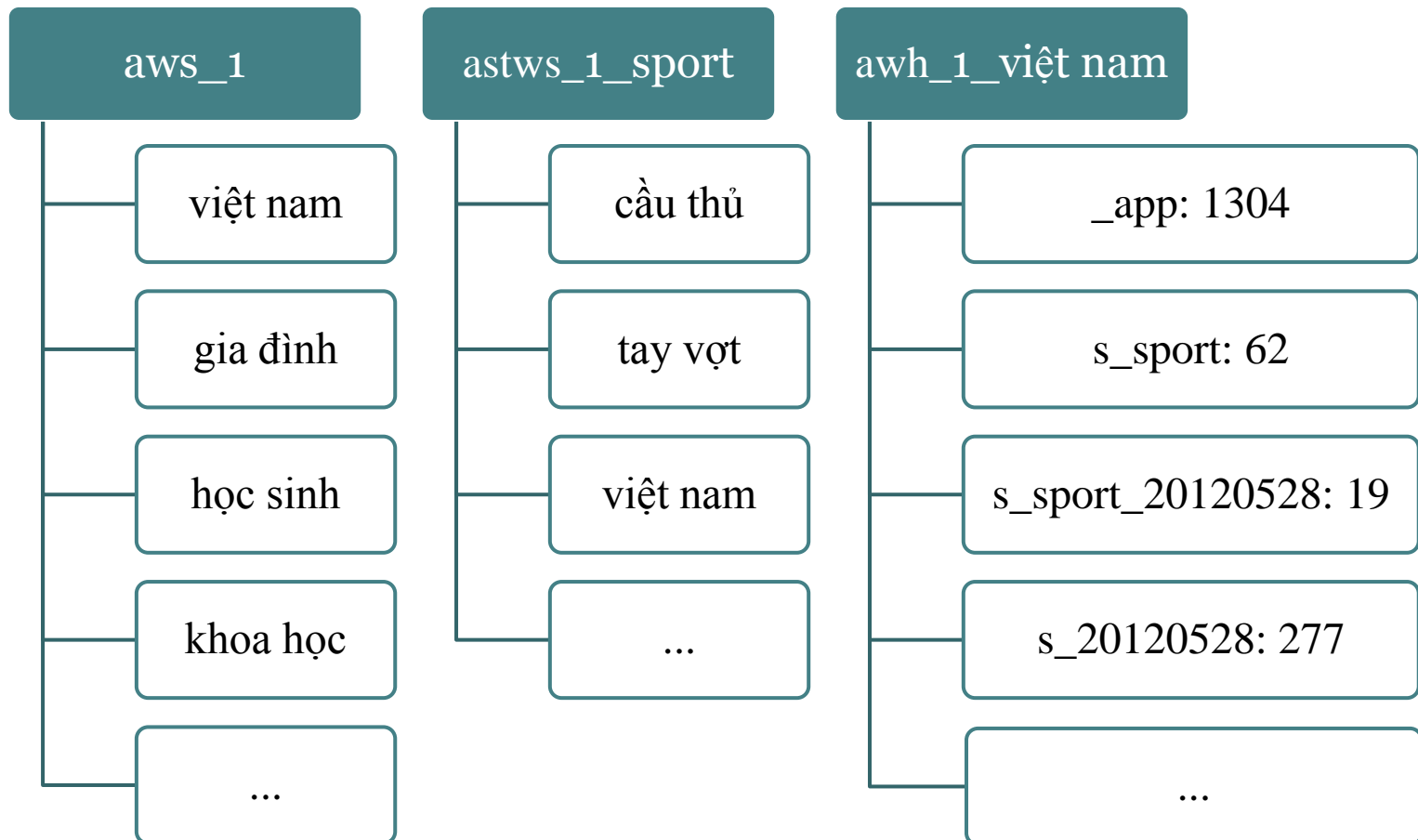
Cài đặt [2/8]: Trích xuất từ khóa

- Input: Văn bản
 - Output: Tập các từ khóa
 - Thuật toán
 1. Loại bỏ kí tự đặc biệt, trắng
 2. Loại bỏ stop word (73 từ)
 3. Tạo tập các từ đơn S1
 4. Ghép các từ đơn thành bộ đôi trong tập S2, bộ ba S3
 5. loại bỏ các từ trùng lặp S2, S3
 6. Output là $S2 \cup S3$
- Input: “Người Việt Nam và các bạn Lào...”
 1. “người việt nam và các bạn Lào”
 2. “người việt nam bạn Lào”
 3. $S1 = [\text{người, việt, nam, bạn, Lào}]$
 4. $S2 = [\text{người việt, việt nam, bạn Lào}], S3 = [\text{người việt nam}]$
 5. N/A
 6. Output: [người việt, việt nam, bạn Lào, người việt nam]

Cài đặt [3/8]: Thống kê từ khóa

- Sử dụng Redis:
 - set “aws_{appId}”: từ trong toàn ứng dụng
 - set “astws_{appId}_{section}”: từ trong mục
 - hash “awh_{appId}_{word}”:
 - key “_app”: bộ đếm trong toàn ứng dụng
 - key “s_{section}”: bộ đếm trong mục

Cài đặt [4/8]: Thống kê từ khóa



Cài đặt [5/8]: Thống kê từ khóa

- Thống kê toàn ứng dụng:
 - `SORT aws_{appId} BY
awh_{appId}_*->_app DESC`
- Thống kê một mục:
 - `SORT astws_{appId}_{section} BY
awh_{appId}_*->s_{section} DESC`

Cài đặt [6/8]: Tìm kiếm văn bản tương tự

- Sử dụng MongoDB, ứng dụng MapReduce
- Input: Văn bản T, tập văn bản $S = \{S_1, S_2, \dots, S_n\}$
- Map:
 1. Lấy n-gram ($n = 5, w = 50$) của T và S_i là N_T và N_i
 2. $R = N_T \cap N_i, r = |R|/\min(|N_T|, |N_i|)$
 3. Nếu $r > 0.5$, thực hiện emit `_id` của S_i
- Reduce:
 - Trả lại kết quả đầu tiên trong tập giá trị (`_id` duy nhất)

Cài đặt [7/8]: n-gram

- Input: “cộng hòa xã hội chủ nghĩa Việt Nam”, $n = 5$, $w = 50$
 1. “cộnghòaxãhộichủnghĩaViệtNam”
 2. $W = [\text{cộngh}, \text{ộnghò}, \text{nghòa}, \text{ghòax}, \text{hòaxã}, \text{òaxãh}, \dots]$
 - “cộngh” = $\text{ord}(\text{'c'}) + \text{ord}(\text{'ộ'}) + \dots = 8313$
 - “ộnghò” = $\text{ord}(\text{'ộ'}) + \text{ord}(\text{'n'}) + \dots = 8456$
 - ...
 - “ộichủ” = $\text{ord}(\text{'ộ'}) + \text{ord}(\text{'i'}) + \dots = 16116$
 - ...
 3. [ộichủ]
- Output: [ộichủ]

Cài đặt [8/8]: Trích nội dung tin bài

- Input: HTML
- Output: Văn bản thuần
- Thuật toán
 1. Xây dựng cây thành phần HTML
 2. Tập T chứa các thành phần $\langle p \rangle$, $\langle div \rangle$, $\langle span \rangle$, $\langle blockquote \rangle$ có chứa dấu chấm (“.”)
 3. Tập P là tập chứa cha của $T_1, T_2, \dots, T_n \in T$
 4. Xét $|P|$
 - = 0, không tìm thấy văn bản
 - = 1, trả lại nội dung của P_0
 - > 1, trả lại nội dung của $P_i \in P$ chứa nhiều văn bản nhất

Nội dung

- Đặt vấn đề
- Thiết kế hệ thống
- Cài đặt
- **Thực nghiệm**
- Kết luận



Thực nghiệm [1/5]: Hệ thống xử lý tiếng Việt

URI	Phương thức	Ý nghĩa
/users	POST	Tạo người sử dụng mới, đồng thời tạo ứng dụng mới với người sử dụng này là người quản trị.
/documents	GET	Lấy danh sách các tài liệu của một ứng dụng.
/documents/:id	GET	Lấy thông tin về một tài liệu với id được chỉ định.
/documents	POST	Đăng một tài liệu mới lên hệ thống xử lý tiếng Việt.
/similar	POST	Tìm các tài liệu tương tự đoạn text được yêu cầu.
/search	POST	Tìm các tài liệu dựa trên từ khóa được chỉ định.
/words	GET	Lấy danh sách các từ khóa trong các tài liệu đã được xử lý.
/words/:word	GET	Lấy thông tin về một từ khóa được chỉ định.
/sections/:section	GET	Lấy thông tin về một chủ đề được chỉ định.

Thực nghiệm [2/5]: Crawler

- Số nguồn RSS: 249
- Số chuyên mục: 10
- Số bài tải về một ngày: ~4000
- Tỷ lệ trích nội dung thành công: ~90%

Thực nghiệm [3/5]: Front-end

Koluto News

Home

Hà Nội, ngày 27 tháng 5 năm 2012

[Trao Giải thưởng Nhà nước và Danh hiệu NSUT cho các nghệ sỹ ...](#)

21:59:46 27/05/2012, cập nhật cách đây 2 giờ Ngày 27/5, lễ trao Giải thưởng Nhà nước và danh hiệu Nghệ sĩ Ưu tú (NSUT) đã diễn ra trọng thể tại Nhà hát Lớn Hà Nội với sự tham gia của đông đảo giới văn nghệ sĩ cả nước. Tới dự và chia vui với các văn nghệ sĩ có Chủ tịch nước Trương Tấn San...

[- Bất chấp thời tiết trước giờ diễn có cơn mưa, hàng chục n...](#)

ghìn khán giả vẫn có mặt tại sân vận động Mỹ Đình từ 4 giờ chiều để có vị trí thuận lợi gần nhất với sân khấu. Chương trình đại nhạc hội MTV EXIT với mục đích tuyên truyền chống lại nạn buôn người trái phép đã diễn ra thành công tốt đẹp vào tối 26/5 tại SVĐ Mỹ Đình. Điều may mắn cho chương trình ...

[Mỹ Tâm nổi bật hơn cả dàn sao ngoại ở MTV Exit - Sánh vai...](#)

bên cạnh Brown Eyed Girls và Simple Plan trong đêm nhạc kêu gọi chống nạn buôn người diễn ra ở Hà Nội, tối 26/5, "Họa mi tóc nâu" vẫn gây được ấn tượng và sức hút mạnh mẽ. >> >> Được chọn làm đại sứ cho chiến dịch chống nạn buôn người của MTV Exit tại Việt Nam, tất nhiên Mỹ Tâm không thể ...

Văn hóa

[Trao Giải thưởng Nhà nước và D...](#)

(HNM) - Xác suất đóng vai trò quan trọng tron...

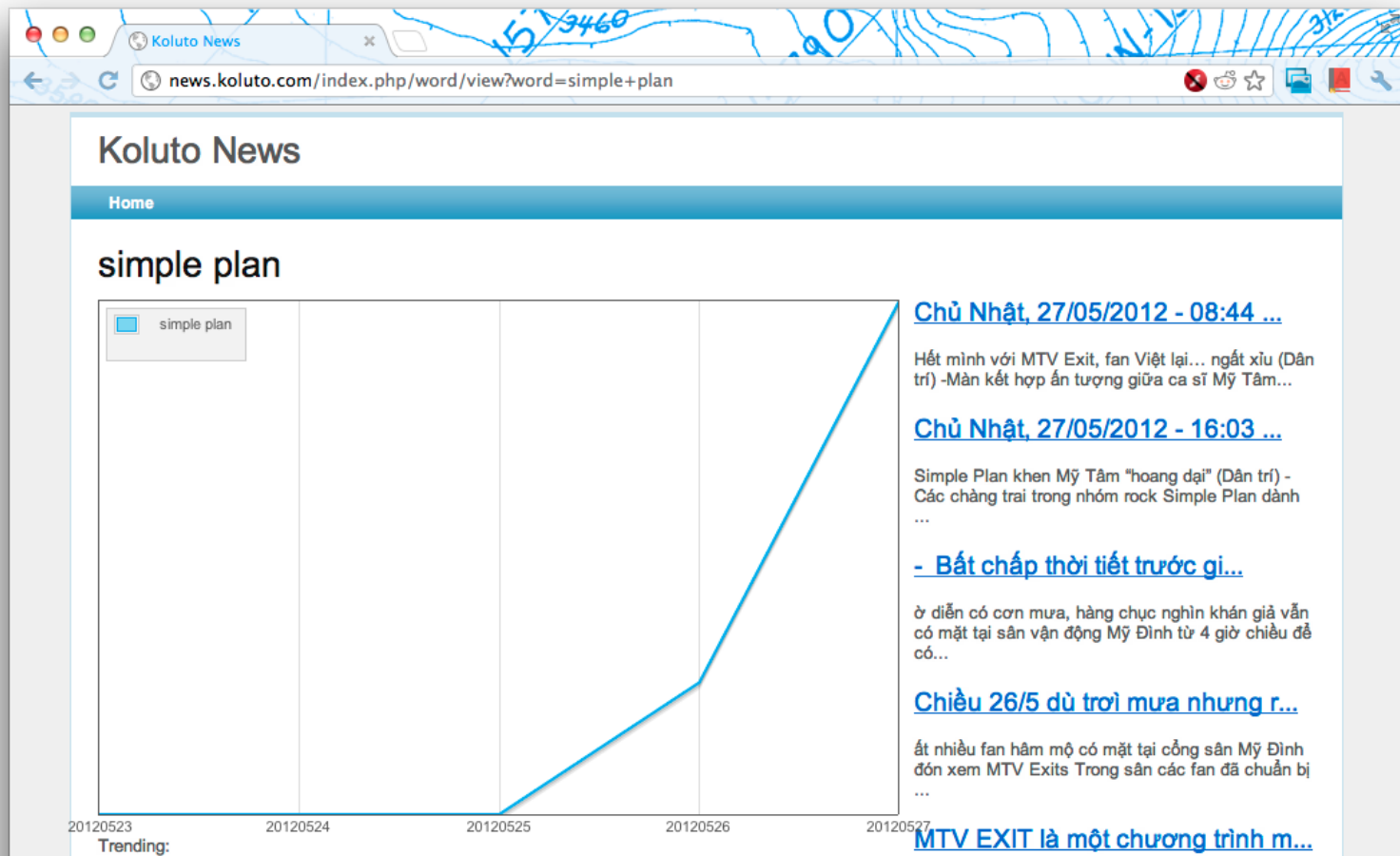
g cuộc sống. Với mỗi cá nhân hay tổ chức, việc lựa chọn khả năng có xác suất cao nhất trong các khả năng có thể xảy ra sẽ giúp cho ta có cơ hội để thà...

[GS Hoàng Quang Thuận nhận Bằng kỷ lục châu Á ...](#)

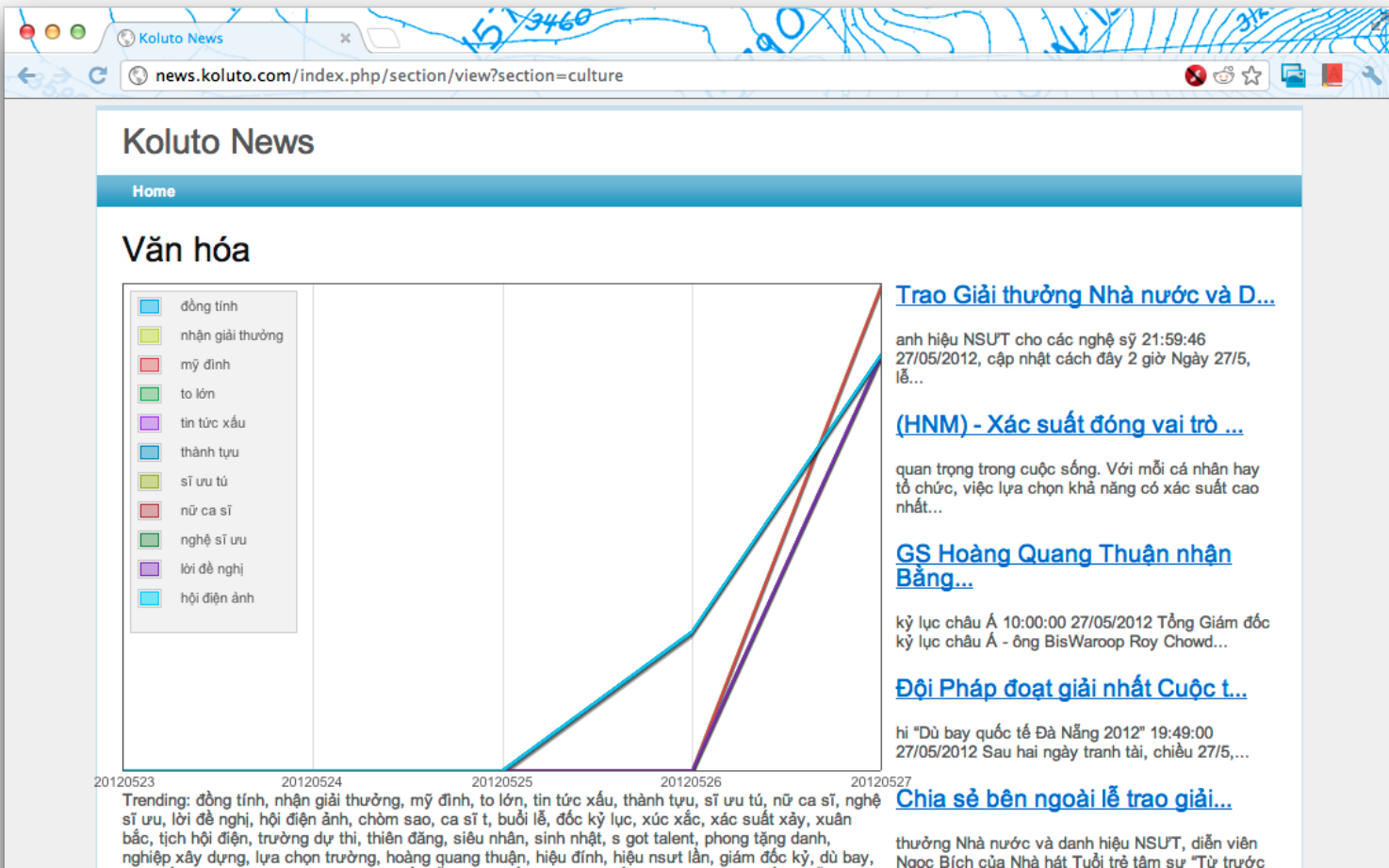
Trending

- làng nghề
- danh hiệu nsut
- simple plan
- trái cây
- hồ dân
- viên chức
- nghìn đồng
- vụ tai nạn

Thực nghiệm [4/5]: Front-end



Thực nghiệm [5/5]: Front-end



Nội dung

- Đặt vấn đề
- Thiết kế hệ thống
- Cài đặt
- Thực nghiệm
- **Kết luận**

Kết luận [1/2]

- Ưu điểm:
 - Hệ thống hoàn chỉnh, chạy ổn định
 - Tin tổng hợp, thống kê có giá trị
- Nhược điểm:
 - Tốc độ chưa cao
 - Giao diện trang tin chưa đẹp mắt

Kết luận [2/2]: Hướng phát triển

- Cải thiện thuật toán
- Bổ sung chức năng
- Tăng tốc độ xử lý
- Triển khai trên cluster



Cám ơn thầy cô đã lắng nghe