ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

TRẦN HỮU DỰ

CÁC KỸ THUẬT PHÂN TÍCH VÀ LẤY TIN TỰ ĐỘNG TỪ WEBSITE

Ngành: Công nghệ Thông tin

Chuyên ngành: Kỹ Thuật Phần Mềm

Mã số: 60 48 0103

TÓM TẮT LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Hà Nội -2016

MỤC LỤC

MỞ ĐẦU

Trong thực tế, tất cả mọi người, tất cả các ngành nghề, mọi lĩnh vực đều có nhu cầu thuthập và cập nhật thông tin, tin tức nhanh nhất có thể.Nguồn cung cấp các thông tin đó đến từ nhiều nguồn các nhau, tại nhiều địa điểm khác nhau... trong đó việc cập nhật tin tức từ các website cũng là một trong số những nguồn cung cấp thông tin cần thiết và quan trọng. Do đó đã thúc đẩy nghiên cứu các kỹ thuật phân tích và lấy tin tự động từ các website. Nghiên cứu các kỹ thuật phân tích và lấy tin tự động từ các website tới người dùng một cách kịp thời có ý nghĩa lý luận và thực tiễn.

Mục tiêu chính của luận văn là nghiên cứu tổng thể các kỹ thuật phân tích và đọc tin tự động từ website. Mục đích nhằm nắm bắt được các công nghệ và các kỹ thuật phân tích và lấy tin tự động từ file RSS của website để xây dựng ứng dụng chạy trên thiết bị di động thông báo cho người dùng biết có tin mới một cách tự động và trong thời gian nhanh nhất có thể.

Bố cục của luận văn tuân theomẫu của trường Đại Học Công Nghệ- ĐHQGHN; Luận văn gồm có 4 chương chính ngoài ra còn có phần mở đầu, kết luận và tài liệu tham khảo. Trong đó chương 1 sẽ nêu lên nhu cầu cập nhật tin tức của người dùng và ứng dụng của việc đọc tin tự động.

Chương 2: sẽ giới thiệu về cấu trúc và kỹ thuật làm việc với XML, RSS 2.0, các phương pháp kỹ thuật Interface DOM, Interface SAXđể phân tích RSS 2.0.

Chương 3: sẽ trình bày chi tiết hơn về các hàm được cung cấp dùng trong kỹ thuật phân tích website với Tool Jsoup.

Chương 4: là phần thực nghiệm xây dựng ứng dụng tự động cập nhật thông báo về nội dung mới từ các website.

Kết luận: Tóm lược lại các kết quả đã nghiên cứu được về các kỹ thuật phân tích XML đọc tin tự đông.

Tài liệu tham khảo: các nguồn tài liệu đã sử dụng thực hiện luận văn.

Xin trân trọng cảm ơn

Tác giả: Trần Hữu Dự

CHƯƠNG 1 GIỚI THIỆU

1.1. Nhu cầu cập nhật tin tức của người dùng

Hầu hết mọi người quan tâm đến nhiều trang website có thay đổi nào trên một lịch trình không thể đoán trước nội dung ví dụ về các trang web như là các trang web tin tức, các trang cộng đồng và thông tin tổ chức tôn giáo, các trang thông tin sản phẩm, các trang web y tế, và weblog...

Trước đây, Email thông báo về những thay đổi là một giải pháp sớm để vấn đề này. Sau đó RSS ra đời là một cách tốt hơn để cung cấp, thông báo về nội dung mới và thay đổi.Thông báo thay đổi đến từ nhiều trang web khác nhau được xử lý một cách dễ dàng, kết quả được trình bày một cách có tổ chức, có cấu trúc tốt và khác biệt từ email.

1.2. Ứng dụng của đọc tin tự động

Ứng dụng trong thương mại:Ứng dụng đọc tin tự động là một ứng dụng trực tuyến tiết kiệm thời gian tốt nhất và rất dễ sử dụng. Với công nghệ đọc và phân tích cú pháp của RSS là một cách để cho người dùng lướt web có thể có được những tin tức mới nhất hoặc "tóm" được những bản hợp đồng du lịch trọn gói giá rẻ nhưng chất lượng, mua được những tấm vé xem hòa nhạc mong muốn hoặc mua được những mặt hàng thiết yếu tại các siêu thị lớn cùng với chương trình khuyến mãi hấp dẫn và gần như là tất cả mọi thứ mà mọi người thường mua thông qua mạng Internet[9].

Ứng dụng trong y tế: Nhu cầu cập nhật thông tin y học về các loại bệnh mới, các phương pháp điều trị mới trở thành mối quan tâm hàng đầu của nhiều bệnh nhân.xNếu như họ sử dụng ứng dụng để thường xuyên cập nhật tin tức từ 1 website về các biểu hiện tiến triển của bệnh, lời khuyên từ bác sĩ, thông tin về các loại thuốc mới điều trị tiểu đường, hay các loại thực phẩm tốt đối với họ thì rất là hữu ích.

Ứng dụng trong trường học:Các trường đại học, cao đẳngthường duy trì một kênh liên lạc với sinh viên và giảng viên bởi 1 website. Khi sinh viên có mối quan tâm tới một trong các thông tin trên thì hoàn toàn có thể sử dụng ứng dụng như công cụ theo dõi cập nhật thông tin cho mình một cách hiệu quả và kịp thời nhất.

Có thể thấy rằng ý nghĩa trong thực tiễn của RSS nói chung và Ứng dụng đọc tin tự động nói riêng là rất có ý nghĩa, quan trọng và cần thiết.

CHƯƠNG 2
CÁC KỸ THUẬT PHÂN TÍCH VÀ LẤY TIN TỰ ĐỘNG

2.1. Giới thiệungôn ngữ mở rộng đánh dấu XML

Khái niệm XML: XML (viết tắt của từ tiếng Anh Extensible Markup Language, "ngôn ngữ đánh dấu mở rộng") [10]là ngôn ngữ đánh dấu với mục đích chung do W3C đề nghị, để tạo ra các ngôn ngữ đánh dấu khác nhau.

Đặc điểm của XML: XML cung cấp một phương tiện dùng văn bản (text) để mô tả thông tin, áp dụng một cấu trúc kiểu cây cho thông tin. Đơn vị cơ sở của XML là các ký tự theo định nghĩa của Universal Character Set(bộký tự toàn cầu). Các ký tự được kết hợp theo các tổ hợp chuỗi hợp lệ để tạo thành tài liệu XML.Tài liệu XML được soạn thảo tạo, bảo trì môt cách đơn giản, thuận tiện bởi sự phổ biến của các phần mềm soạn thảo văn.

Khai báo định dạng file:Dòng đầu tiên trong file XML phải là dòng khai báo XML phần tùy chọn này dùng để nhận dạng đây là một file tài liệu XML. Khai báo có thể được viết như sau:

- Khai báo dạng đơn giản như sau *<?xml?>* .

Tạo phần tử gốc trong tài liệu: Thẻ bắt đầu và thẻ kết thúc của phần tử gốc bao quanh toàn bộ nội dung của file tài liệu XML. Và chỉ có duy nhất một phần tử gốc trong một file dữ liệu, và thẻ "đóng" này để chứa đựng tất cả nội dung của file tài liệu XML.

Đặt tên các phần tử tùy biến: Với ngôn ngữ XML cho phép đặt tên các phần tử một cách tùy biến, sau đó định nghĩa theo Document Type Definition hoặc theo lược đồ XML tương ứng với các tên đó.

Lồng ghép các phần tử: Lồng ghép là đặt một phần tử này nằm bên trong một phần tử khác.

Thêm các thuộc tính: Các thuộc tính hoàn toàn có thể được thêm vào các phần tử, bao gồm cặp tên-giá trị, với giá trị được đặt bên trong hai dấu ngoặc kép ("), ví dụ: type="dessert".

2.2. Giới thiệu Kỹ thuật RSS

RSS (Really Simple Syndication)[6]có nghĩa là "dịch vụ cung cấp thông tin thực sự đơn giản", dùng trong việc chia sẻ tin tức Web (Web syndication) được dùng bởi nhiều website tin tức và weblog với công nghệ của RSS.

Định nghĩa RSS: RSS là viết tắt của một trong những thuật ngữ sau:

- Really Simple Syndication: "Dịch vụ cung cấp thông tin thực sự đơn giản".

- Rich Site Summary, RDF Site Summary[13]: Tóm lược thông tin phong phú, tóm lược thông tin theo định dạng RDF.
- RDF - Resource Discovery Framework:(tạm dịch là "cơ cấu khám phá tài nguyên"): là định dạng của phiên bản RSS 1.0. Ở một số trang web biểu tượng RDF sử dụng thay cho RSS.

Lịch sử phát triển RSS: Trước RSS, có nhiều định dạng khác cũng từng được dùng cho vấn đề chia sẻ thông tin.Lịch sử phát triển của RSS trải qua các giai đoạn như sau:

- RDF (*Resource Description Framework*) Site Summary, phiên bản đầu tiên của RSS, được tạo ra bởi Dan Libby củaNetscape vào tháng Ba 1999.
- Nhóm RSS-DEV tiếp tục đưa ra RSS 1.0vào tháng 12 năm 2000.
- Vào tháng 9 năm 2002, Winer cho ra bản cuối cùng của RSS 0.92, bây giờ gọi là *RSS 2.0*[6].

Cấu trúc định dạng tập tin RSS 2.0[6]:Một file RSS là do một phần tử (element) ***<channel> </channel>*** và các phần tử con của nó tạo nên.

Thành phần <channel> của RSS

Một kênh**<channel>** thường có ba phần tử đểbiết thông tin của kênh bao gồm: <title> (tiêu đề): Tên của kênh hoặc nguồn cấp tin<link> (liên kết): URL của trang web hoặc vùng trang web có liên kết với kênh này <description> (mô tả): Mô tả ngắn gọn về kênh đó.

- **Thành phần <Item> bản tin**

Các **<Item>** là phần quan trọng nhất của một nguồn cấp tin. Mỗi <Item>có thể là một bài weblog. Các <Item>có thể thường xuyên thay đổi nội dung, có thể chứa bao nhiêu <Item>cũng được. Một <Item>thường chứa ba phần tử như sau:

- <title>: Phần tử này là tên của <Item>.
- <link>: Phần tử này là URL của <Item>.
- <description>: Phần tử này thường là một bản tóm tắt.

2.3. Phương pháp Interface DOM phân tích RSS 2.0

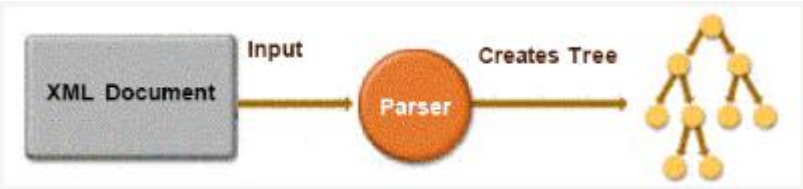
Khái niệm: Document Object Model [1] là một phương pháp tiếp cận dữ liệu từ XML, trong đó tất cả các thành phần của XML như thẻ, thuộc tính, text… đều được xem như là các đối tượng.

Các đối tượng trong DOM: Tất cả các thành phần trong XML được xem như là 1 "node". Trong đó root là documentnode, các thẻ là element node, thuộc tính là attribute node, text là text node, ghi chú là comment node….

Mối quan hệ giữa các node được mô tả bằng thuật ngữ " parent" và "child" (cha và con). Thuật ngữ "leaf node" dùng để chỉ các node không có child node

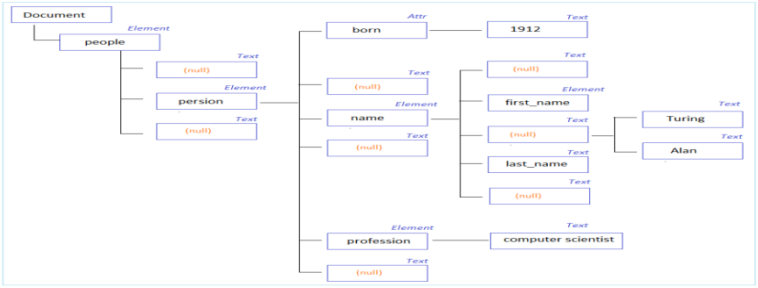
Các đặc điểm của DOM: Truy cập tài liệu XML như là một cấu trúc cây, việc truy cập đượctới hầu hết các nút element và các nút text.

Hoạt động của DOM: Hình 2.3.1 cho thấy đầu vào là một tài liệu XML được bộ phân tích bởi mô hình DOM và một sẽ cây được tạo ra trong bộ nhớ mangthông tin của tài liệu đó .Việc phân tích tài liệu XML bây giờ đưa về phân tích,xử lý các nút của cây.



Hình 2.1 - Mô hình hoạt động DOM[3]

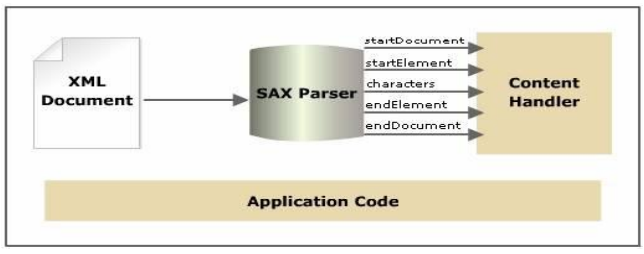
Cấu tạo cây DOM và các kiểu nút(Node): Tài liệu XML được hình dung như là một cây. Một cây được làm từ các nút cây (nodes), có 12 kiểu nút cây khác nhau.
Mô hình cây DOM được thể hiện dưới các dạng node (xem Hình 2.2):



Hình 2.2 - Mô hình tổng thể cây DOM[3]

2.4. Phương pháp Interface SAX phân tích RSS 2.0

Simple API for XML (SAX)[4]là một phương pháp phân tích dữ liệu XML dựa vào sự kiện trong quá trình đọc file XML từ trên xuống dưới.



Hình 2.3 - Mô hình SAX xử lý parser XML[4]

Một điểm cần lưu ý rằng SAX không ghi nhớ XML như một cấu trúc cây trên bộ nhớ. SAX chỉ xây dựng cấu trúc của nhiệm vụ đang thi hành trên bộ nhớ do đó SAX sẽ thực hiện nhanh hơn và ít tốn tài nguyên hơn.

SAX gồm có 2 interface chính: XMLReader là interface dùng để đọc XML và ContentHandler dùng để nhận dữ liệu từ XML. Hai interface này giải quyết 90% nhu cầu của người dùng với SAX.

Các trình khách cung cấp một phân tích lớp của DefaultHandler.Phân lớp này được sử dụng để lấn quyền những phương pháp trên và xử lý dữ liệu.Quá trình này có thể bao gồm việc lưu trữ dữ liệu vào trong cơ sở dữ liệu, hoặc viết chúng ra một luồng dữ liệu.

2.5. Phân tích nội dung XML trong Android sử dụng XmlPullParser

Sơ lược về XmlPullParser[12, 16]:Android không cung cấp hỗ trợ cho StAX API của Java. Tuy nhiên Android lại đi kèm với một trình phân tích XmlPullParser làm việc tương tự như StAX.Nó cho phép mã ứng dụng "kéo" hoặc tìm kiếm các sự kiện từ trình phân tích, trái ngược với trình phân tích SAX tự động đẩy các sự kiện cho trình xử lý.

Ưu điểm của XmlPullParser: Phương pháp này với khá nhiều ưu điểm đạt được như đơn giản trong xử lý mã nguồn, sử dụng ít tài nguyên bộ nhớ thiết bị, tốc độ xử lý nhanh hơn phương pháp DOM và SAX do tập trung vào dữ liệu hiện hành để xử lý, tính năng lọc dữ liệu tương đối đơn giản dễ sử dụng.

Đặc điểm của XmlPullParser[12, 16]:Bộ phân tích của XmlPullParser xem tài liệu XMLchỉ có các phần tửvà các nút văn bản tương ứng với phần tử để phân tích cú pháp. Các thuộc tính không tạo ra một sự kiệncó thể lấy các thuộc tính ra từ một phần tử.

CHƯƠNG 3
KỸ THUẬT PHÂN TÍCH WEBSITE VỚI JSOUP

3.1. Giới thiệu

Ngày nay khối lượng thông tin lưu trữ trên Internet ngày càng tăng chóng mặt, việc trích rút được các thông tin chính xác, gọn gẽ từ kho dữ liệu trên là vô cùng có ý nghĩa. Đã có rất nhiều ứng dụng và nhiều Tools ra đời như: Search Engine[18], RSS[15], Feedback[17], jSon[14], Jsoup[5].

Trong đó Tool Jsoup là một thư viện mã nguồn mở của Java. Jsoup là một công cụ phân tích cú pháp mã HTML, cung cấp nhiều Packages và nhiều API thuận tiện, an toàn để thao tác với HTML.

3.2. Định nghĩa Jsoup

Jsoup là Java HTML Parser[5]. Nói cách khác Jsoup là một thư viện được sử dụng để phân tích tài liệu HTML. Jsoup cung cấp các API dùng để lấy dữ liệu và thao tác dữ liệu từ URL hoặc từ file HTML. Nó sử dụng các phương thức giống với DOM, CSS, JQuery để lấy dữ liệu và thao tác với dữ liệu.

Thành phần của Jsoup API

Jsoup có 6 Packages và nhiều Class khác nhau. Trong khung khổ của luận văn này chỉ xin phép giới thiệu 3 lớprất quan trọng đó là các lớp:

- org.jsoup.Jsoup;
- org.jsoup.nodes.Document ;
- org.jsoup.nodes.Element;

Sau đây là các mô tả kỹ thuật về các phương thức được cung cấp bởi các lớp:

Jsoup.java

Phương thức	Mô tả
static Connection connect(String url)	Tạo và trả về đối tượng Connection kết nối của URL
static Document parse(File in, String charsetName)	Phân tích 1 file tài liệu html với chỉ định mã hóa.
static Document parse(File in, String charsetName, String baseUri)	Phân tích file tài liệu html với chỉ định mã hóa, và baseUri
static Document parse(String html)	Phân tích mã html chuyển thành Document

Bảng 3.1 - Các phương thức của lớp Jsoup.java

Document.java

Phương thức	Mô tả
Element body()	Truy nhập vào phần tử body
Charset charset()	Trả về charset được sử dụng trong tài liệu này
void charset(Charset charset)	Sét charset sử dụng cho tài liệu này.
Element head()	Truy cập vào phần tử head.
String location()	Trả về URL của tài liệu này.
String nodeName()	Trả về node name của node này.
Document normalise()	Normalise the document.
String outerHtml()	Trả về outer HTML của node này.

Bảng 3.2- Các phương thức của lớp Document.java

3.3. Các phương thức DOM

Jsoup có một vài phương thức gần giống với các phương thức trong mô hình DOM (Phân tích tài liệu XML)

Phương thức	Mô tả
Element getElementById(String id)	Tìm một phần tử cho bởi ID
Elements getElementsByTag(String tag)	Tìm phần tử, bao gồm và cả đệ quy
Elements getElementsByClass(String className)	Tìm phần tử có className
Elements getElementsByAttribute(String key)	Tìm kiếm các phần tử
Elements siblingElements()	Trả về các phần tử anh em với phần tử

Bảng 3.2- Các phương thức của DOM

Các phương thức lấy dữ liệu trên Element

Phương thức	Mô tả
String attr(String key)	Trả về giá trị thuộc tính cho bởi key
String id()	Trả về thuộc tính ID, nếu có
String className()	Trả về chuỗi chữ giá trị của thuộc tính "class”
Set<String> classNames()	Trả về tất cả các class names.
String text()	Trả về một văn bản kết hợp text của nó

Bảng 3.3- Các phương thức lấy dữ liệu của Element

Các phương thức vận dụng HTML

Methods	Description
Element append(String html)	Nối thêm HTML vào trong phần tử này
Element prepend(String html)	Nối thêm HTML vào phần tử này.
Element appendText(String text)	Tạo và nối một TextNode mới vào phần tử này.
Element prependText(String text)	Tạo và nối một TextNode mới vào phía trước

Bảng 3.3 - Các phương thức vận dụng của HTML

3.4. Các phương thức giống Css, jQuery

Các phần tử JSoup hỗ trợ cú pháp giống với CSS (hoặc JQuery) giúp tìm kiếm các phần tử phù hợp, những hỗ trợ như vậy là rất mạnh mẽ. Các phương thức lựa chọn có sẵn trong class Document, Element hoặc Elements.

Selector Unit (Bộ lựa chọn)

Các Selector	Mô tả
tagname	Tìm kiếm các phần tử theo tên thẻ. Ví dụ: a
ns|tag	Tìm kiếm các phần tử theo tên thẻ trong một không gian tên (namespace
#id	Tìm kiếm phần tử theo ID, ví dụ #logo
.class:	Tìm kiếm các phần tử theo tên class, ví dụ .masthead
[attribute]	Các phần tử với thuộc tính, ví dụ [href]
[^attr]	Các phần tử với thuộc tính bắt đầu bởi,
[attr=value]	Các phần tử với giá trị thuộc tính, ví dụ [width=500]

Bảng 3.4- CácSelector Unit

Selector kết hợp

Selector	Mô tả
el#id	Phần tử với ID, ví dụ div#logo
el.class	Các phần tử với class, ví dụ div.masthead
el[attr]	Các phần tử với thuộc tính, ví dụ a[href]
Kết hợp bất kỳ	ví dụ a[href].highlight
parent > child	Các phần tử con trực tiếp của phần tử cha,

Bảng 3.5- CácSelector kết hợp

Pseudo selectors

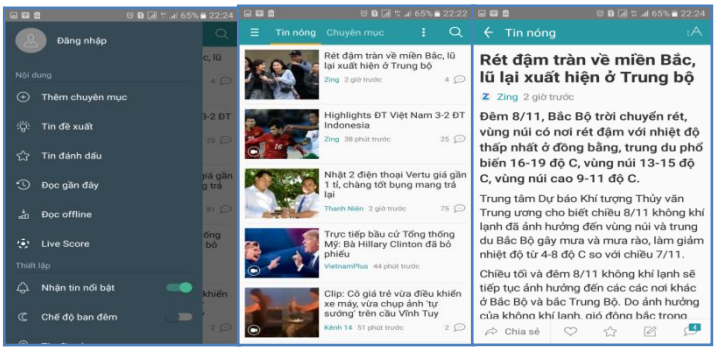
Selector	Mô tả
:lt(n)	Tìm kiếm các phần tử có chỉ số anh em
:gt(n)	Tìm kiếm các phần tử có chỉ số anh em lớn hơn n
:eq(n)	find elements whose sibling index is equal to n
:has(seletor)	Tìm kiếm các phần tử chứa các phần tử khớp với selector
:not(selector)	Tìm kiếm các phần tử không khớp với selector
:contains(text)	Tìm kiếm các phần tử chứa đoạn text đã cho

Bảng 3.6- CácPseudo Selector

CHƯƠNG4
XÂY DỰNG VÀ PHÁT TRIỂN ỨNG DỤNG TỰ ĐỘNG CẬP NHẬT TIN TỨC

4.1. Một số ứng dụng hỗ trợ đọc tin tức từ file RSS hiện có

Báo mới[7]:Là một ứng dụng đọc tin tức online và offline từ trang báo điện tử http://www.baomoi.com.Với các màn hình giao diện của ứng dụng (xem Hình 4.1 từ trái qua phải)

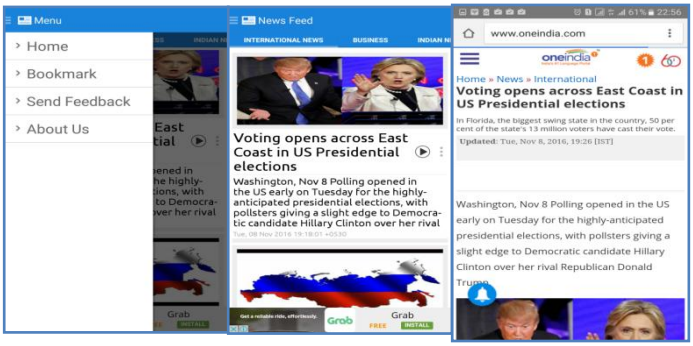


Hình 4.1 - Giao diện của BaoMoi

TinMoi24h[8]cũng là một trong số các ứng dụng được xây dựng công phu lấy nguồn tin từ trang http://tinmoi24.com/**TinMoi24**. Một số giao diện của ứng dụng (xem Hình 4.2 từ trái qua phải từ trên xuống dưới.

Hình 4.2- Giao diện của TinMoi24h

News Feed[2]là một ứng dụng khá phổ biến trên thị trường hiện nay News Feed lấy nguồn tin từ http://www.oneindia.com.Với các chứ năng đơn giản gọn nhẹ tiết kiệm được dung lượng của máy từ trái qua phải của Hình 4.3.



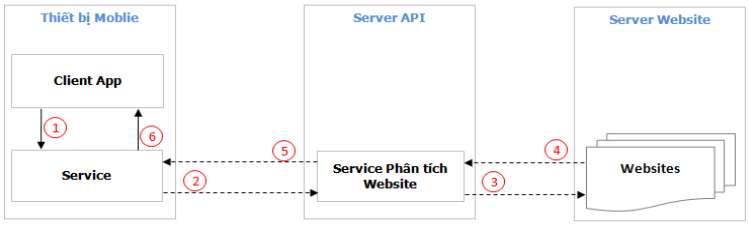
Hình 4.3- Giao diện của News Feed

4.2. Phân tích thiết kế hệ thống đọc tin tự động từ website

- **Thiết kế tổng quan hệ thống**

Ở đây có hai giải pháp được đưa ra cho thiết kế tổng quan hệ thống như sau:

Giải pháp 1: Mô hình Client - Server - Website



Hình 4.4- Mô hình Client-Server-Website

Thiết kế mô hình tổng thể như Hình 4.4 luồng dữ liệu sẽ hoạt động như sau:

(1): Client sẽ gọi 1 Service yêu cầu quét thu thập dữ liệu mới tới Server API,

②: Service gửi đi yêu cầu quét thu thập dữ liệu mới tới Server API

③: Server API nhận yêu cầu tiến hành phân tích RSS và lưu dữ liệu tin tức mới của Website.

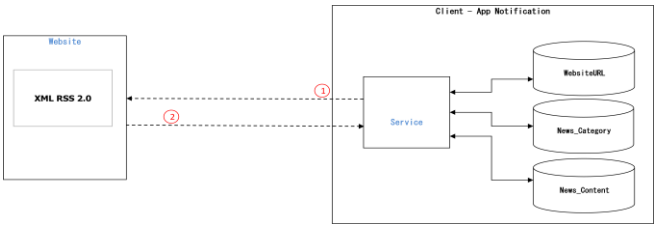
④: Website cung cấp file RSS cho phép Server đọc dữ liệu từ RSS.

⑤: Server API trả về dữ liệu mới cho Client.

⑥: Client tiến hành hiển thị thông tin, và đẩy thông báo tự động tới giao diện người dùng.

Với giải pháp 1 thể hiện được các ưu điểm như Client và Server phân tích website hoạt động độc lập nhau từ đó sẽ giảm tải cho Client. Khi Client Offline thì Server vẫn lấy được tin mới và lưu tại server..

Giải pháp 2: Mô hình Client - Website



Hình 4.5-Mô hình Client-Website

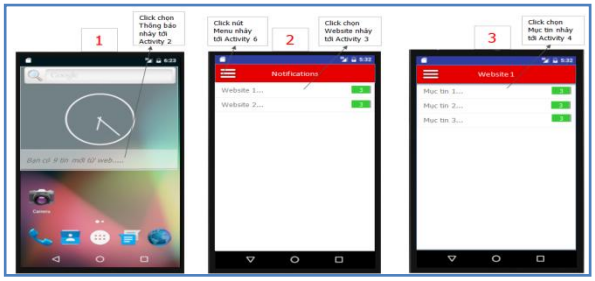
Với thiết kế mô hình tổng thể như trên luồng dữ liệu sẽ hoạt động như sau:

①: Client sẽ gọi 1 Service yêu cầu quét thu thập dữ liệu mới tới Website, tiến hành phân tích RSS, đọc dữ liệu và lưu xuống DB Client.

②: Website cung cấp file RSS và dữ liệu cho Client.

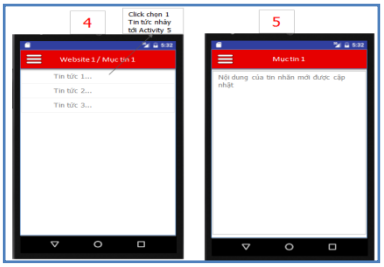
Ưu điểm thể hiện mạnh ở chỗ không tập trung xử lý tại một nơi mà xử lý riêng lẻ đối với từng Client nên tránh được hiện tượng quá tải. Tuy nhiên có nhược điểm là phụ thuộc vào Internet, 3G, offline không tự động lấy được tin mới, phụ thuộc vào phần cứng của Client.

- **Thiết kế giao diện**



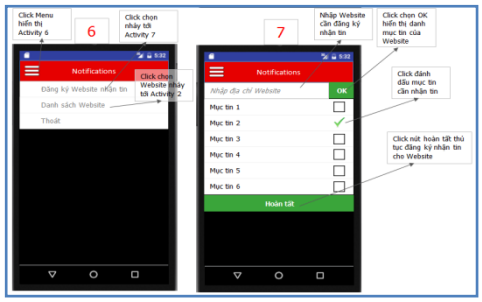
Hình 4.6-Thiết kế giao diện hiển thị Danh sách website

- (Xem Hình 4.6) là 3 màn hình giao diện được thiết kế đánh số thứ tự 1, 2, 3, trong đó thứ tự 1 là giao diện nhận Notifi từ service, thứ tự 2 là tải dữ liệu từ database và hiển thị danh sách website, số lượng tin tức mới.



Hình 4.7- Thiết kế giao diện hiển thị nội dung của tin tức

Hình 4.7 là 2 màn thiết kế giao diện hiển thị danh sách các tin tức và nội dung chi tiết của tin tức

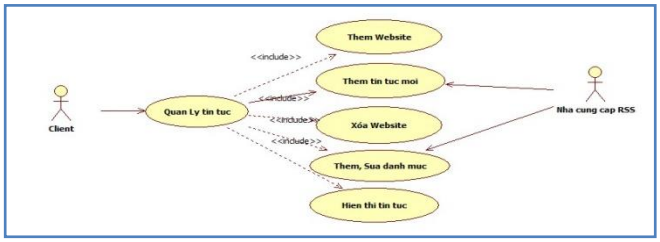


Hình 4.8- Thiết kế giao diện đăng ký website nhận tin

Hình 4.8 là 2 màn hình thiết kế giao diện cho menu truy cập các chức năng và màn hình giao diện đăng ký website nhận tin mới

- **Phân tích chức năng hệ thống**

  - **Biểu đồ Use Case:** Biểu đồ User-Case thể hiện sự tương tác giữa người dùng và hệ thống. Từ đó xác định được hệ thống cần phải làm gì.

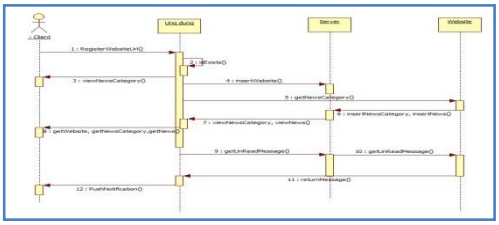


Hình 4.9 - Biểu đồ User-Case

  - **Đặc tả các Use-case:** Hệ gồm 1 Use - case quản lý tin tức cho phép người sử dụng khi mở ứng dụng lên quản lý tin tức bao gồm các việc như: thêm mới, cập nhật trạng thái cho từng danh mục tin, xóa địa chỉ website không muốn

nhận tin. Hiển thị danh sách các website đã đăng, hiển thị danh sách các danh mục tin, hiển thị danh sách các tin tức và đọc tin tức mới nhất.

- **Biểu đồ tuần tự (Sequence Diagram):** Hoạt động của hệ thốngnhìn một cách tổng thể (xem Hình 4.10)gồm các hành động cơ bản như sau.



Hình 4.10 -Biểu đồ tuần tự toàn hệ thống

- **Thiết kế cơ sở dữ liệu**

Cơ sở dữ liệu của chương trình không quá lớn, hơn nữa do ứng dụng chạy trên thiết bị di động nên hệ quản trị cơ sở dữ liệu được cung cấp bởi Android SqlLite. Sau đây là thiết kế cấu trúc các bảng dữ liệu và mô hình quan hệ giữa các bảng:

- **Thiết kế cấu trúc các bảng dữ liệu**

Bảng website_url: Lưu thông tin về địa chỉ các Website để chứa các kênh tin.

website_url				
STT	Tên trường	Kiểu dữ liệu	Ghi chú	Diễn giải
1	website_id	INTEGER	Tự tăng	Mã địa chỉ website
2	website_title	TEXT		Tiêu đề của website
3	website_url	TEXT	Khác rỗng	Đường dẫn chính của web
4	website_url_rss	TEXT		Đường dẫn tới trang gồm RSS
5	website_count_news	INTEGER		Tổng số tin của web chưa đọc

Bảng 4.1: Bảng website_url (địa chỉ website)

Bảng news_category: Lưu thông tin về các danh mục được chọn để nhận tin mới.

news_category				
STT	Tên trường	Kiểu dữ liệu	Ghi chú	Diễn giải
1	id	INTEGER	Tự tăng	Mã category
2	website_id	INTEGER	Khác rỗng	Mã địa chỉ website
3	website_url	TEXT	Khác rỗng	Đường dẫn chính web
4	news_category_title	TEXT		Tiêu đề của danh mục nhận tin mới
5	news_category_url	TEXT	Khác rỗng	Đường dẫn của danh mục nhận tin mới

6	news_category_rss_status	INTEGER	Trạng thái	0: không nhận, 1: nhận
7	news_category_count_news	INTEGER	Mặc định 0	Tổng số tin chưa đọc

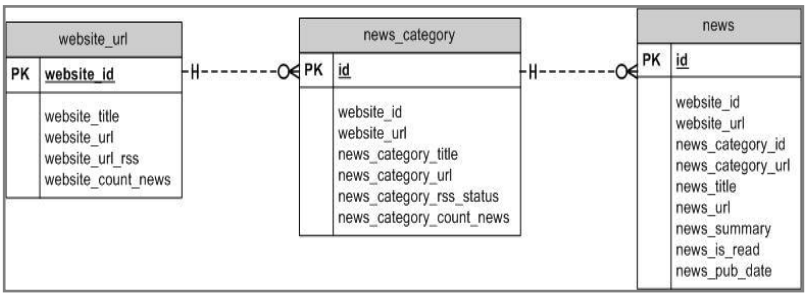
Bảng 4.2 - Bảng news_category(Mục nhận tin mới)

Bảng news: Lưu thông tin các tin tức của một danh mục nào đó.

news				
STT	Tên trường	Kiểu dữ liệu	Ghi chú	Diễn giải
1	id	INTEGER	Tự tăng	Mã địa chỉ của tin tức
2	website_id	INTEGER	Khác rỗng	Mã địa chỉ website
3	website_url	TEXT	Khác rỗng	Đường dẫn chính của web
	news_category_id	INTEGER	Khác rỗng	Mã danh mục
4	news_category_url	TEXT	Khác rỗng	Đường dẫn danh mục
5	news_title	TEXT		Tiêu đề ngắn gọn của tin tức
6	news_url	TEXT		Đường dẫn danh mục
7	news_summary	TEXT		Tóm tắc tin nhắn
8	news_is_read	INTEGER	Mặc định 0	0: chưa đọc, 1: đã được đọc
9	news_pub_date	TEXT		Ngày giờ của tin tức

Bảng 4.3 - Bảng news (Các tin tức mới nhận được)

- **Mô hình quan hệ (ER)**

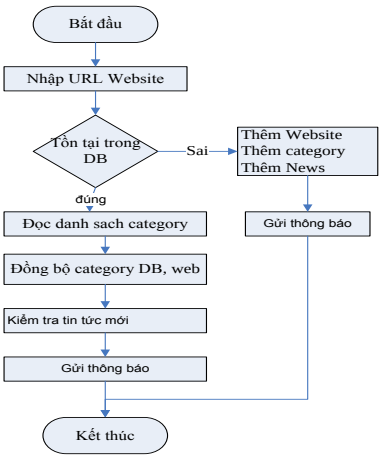


Hình 4.11 - Mô hình quan hệ thực thể(ER)

- **Một số thuật toán được cài đặt**

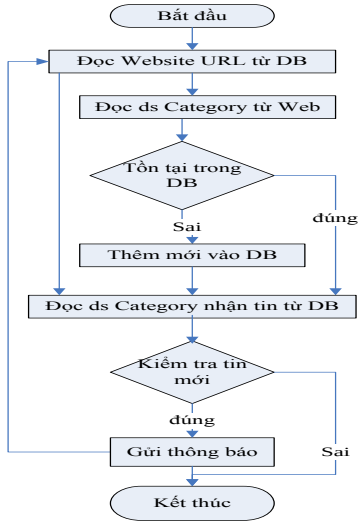
Thuật toán 1: kiểm tra xem một địa chỉ website đã tồn tại trong cơ sở dữ liệu, nếu đã tồn tại thì lấy danh sách các danh mục tin tức có hỗ trợ RSS của web đó lên cho người dùng cập nhật.

Sơ đồ khối giải thuật



Hình 4.12 - Sơ đồ giải thuật kiểm tra sự tồn tại website và danh mục tin tức

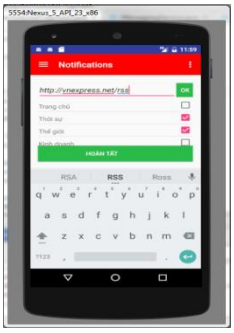
Thuật toán 2: Kiểm tra mỗi khi website cập nhật tin tức mới, báo có tin mới. Sơ đồ giải thuật như sau



Hình 4.13 - Sơ đồ giải thuật kiểm tra sự tồn tại tin tức

4.3. Các chức năng cơ bản của ứng dụng

Chức năng đăng ký: Xem Hình 4.14 Chức năng cho phép người dùng nhập vào địa chỉ của website hoặc mục muốn nhận tin mới.



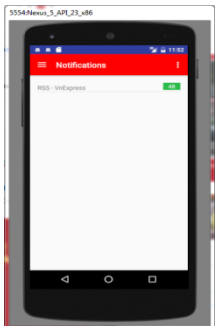
Hình 4.14 -Chức năng đăng ký website nhận tin mới

Để tùy biến cài đặt điều kiện nhận tin mới tại các danh mục tin người dùng kích vào dòng danh mục tin tức, một cửa sổ Popup sẽ được hiển thị cho phép nhận điều kiện tùy biến nhận tin xem Hình 4.15

Hình 4.15 -Chức năngtùy biến cài đặt nhận tin

Sau cùng người dùng nhấn nút hoàn tất để kết thúc chức năng đăng ký nhận tin, lúc này ứng sẽ tiến hành lưu các thông tin về địa chỉ website về mục đánh dấu nhận, về các tin tức liên quan ... xuống database.

Chức năng hiển thị danh sách website đã đăng ký: Hình 4.16 là giao diện chức năng hiển thị danh sách các website đã đăng ký nhận tin.

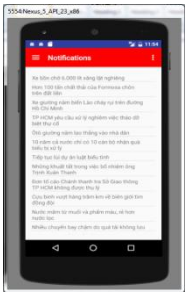


Hình 4.16 -Chức năng danh sách website nhận tin mới

Chức năng hiển thị các danh mục tin tức của 1 website: Tại chức năng này (Xem Hình 4.17) hệ thống sẽ hiển thị tất cả các danh mục tin.

Hình 4.17 -Chức năng danh mục của website nhận tin mới

Chức năng hiển thị tóm tắt nội dung các tin tức:(Xem Hình 4.18) Chức năng này thực hiện việc hiển thị tóm tắt tin tức bao gồm tiêu đề, tóm lược chính...

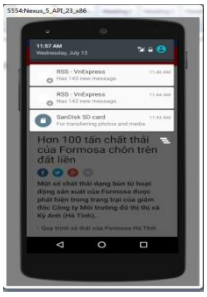


Hình 4.18 -Chức năng danh sách tin mới

Chức năng hiển thị chi tiết tin tức:(Xem Hình 4.19) Chức năng này hiển thị chi tiết nội dung của tin tức thông qua web Browers giúp người dùng có thể đọc toàn bộ tin tức.

Hình 4.19 -Chức năng danh nội dung tin tức mới

Chức năng tự động thông báo có tin tức mới:Hệ thống sử dụng một services chạy ngầm phía dưới tự động quét file RSS của các website từ đó lọc, phát hiện ra các tin tức mới. Sự dụng công nghệ push notification của Android để thực hiện báo cho người dùng biết. (Xem Hình 4.20) là giao diện đẩy tin thông báo, người dùng muốn đọc các tin tức mới.



Hình 4.20-Chức năng tự động thông báo tin mới

4.4. Kết quả thực nghiệm và nhận xét

- **Kết quả thực nghiệm:** Ứng dụng phân tích và lấy tin tự động từ website đã áp dụng thành công các kỹ thuật phân tích DOM, SAX cấu trúc file XML, RSS, HTML. Cụ thể như sau:
  - Áp dụng thành công công cụ Jsoup với kỹ thuật phân tích DOM để phân tích mã nguồn HTML của website để lấy được danh mục các mục tin có cung cấp file RSS.
  - Áp dụng thành công kỹ thuật phân tích SAX với XMLPullParser trong Android để phân tích cú pháp và đọc dữ liệu của RSS 2.0.
  - Phát triển được tính năng tự động quét phát hiện các tin tức mới sau đó tự động gửi thông báo cho người dùng.
  - Ứng dụng có thể phân tích và lấy tin tốt với hầu hết các website có cung cấp RSS
- **Nhận xét:** Đối với mỗi bài toán phân tích lấy tin tự động từ website cụ thể cần phải cân nhắc xem xét thực trạng bài toán, đánh giá được qui mô, độ phức tạp, các yêu cầu về chức năng hệ thống để có thể đưa ra lựa chọn giải pháp, lựa chọn công nghệ cho phù hợp.

Trong trường hợp cụ thể thực nghiệm của luận văn này với các website có file nguồn mã Html, Xml vừa và nhỏ sử dụng kỹ thuật phân tích DOM là rất tốt vì toàn bộ mã nguồn sẽ được tải vào bộ nhớ để bộ phân tích DOM thực hiện một lần duy nhất tạo lên cây DOM, việc truy cập ngẫu nhiên tới các các phần tử của cây DOM là thuận tiện, nhanh chóng, hiệu suất xử lý dữ liệu cao.

Ngược lại các website thường có tuần suất cung cấp tin mới ở file RSS cao, số lượng tin mới nhiều nên việc áp dụng kỹ thuật phân tích SAX để xử lý một nhiệm vụ hiện hành cần thực thi tại một thời điểm trên bộ nhớ là rất tốt và cho hiệu quả cao, giảm việc yêu cầu tài nguyên của thiết bị phần cứng.

Việc áp dụng các công cụ Jsoup và thư viện XmlPullParser trong Android với trường hợp cụ thể này để phù hợp là thực sự cần thiết.

KẾT LUẬN

Luận văn tìm hiểu và trình bàycác kỹ thuật phân tích và lấy tin tự động từ website với các kết quả chính đạt được là:

- ***Hệ thống lại các kỹ thuật phân tích cấu trúc của website*** để có thể lấy được các tin tức một cách tự động từ các website đó.
- ***Phát triển và xây dựng một ứng dụng di động*** trên nền Android có khả năng tự động cập nhật và thông báo đến cho người dùng các tin tức mới tùy theo cài đặt.

Các kỹ thuật trình bày trong luận văn gồm:

- ***XML*** là ngôn ngữ đánh dấu với mục đích dùng chung, có khả năng mô tả nhiều loại dữ liệu khác nhau.
- ***RSS 2.0*** là dịch vụ cung cấp thông tin thực sự đơn giản dùng trong việc chia sẻ tin tức web. RSS là một danh sách các đối tượng mẩu tin được miêu tả gồm có: tiêu đề, nội dung tóm lược, một liên kết đến trang chính, ngày tháng, tác giả...
- ***DOM*** là kỹ thuật tiếp cận dữ liệu XML mà trong đó tất cả các thành phần tạo nên XML được xem là đối tượng, được xem là các "node". Khi parser đọc XML thành đối tượng DOM trên bộ nhớ chúng ta sẽ có một cấu trúc cây với các đối tượng là các node, giữa các node này có mối quan hệ phân tầng cha-con.
- ***SAX*** là một kỹ thuật phân tích dựa vào sự kiện trong quá trình đọc file XML từ trên xuống dưới. SAX không ghi nhớ XML như một cấu trúc cây trên bộ nhớ mà chỉ xây dựng cấu trúc của nhiệm vụ đang thi hành trên bộ nhớ dó đó SAX sẽ thực hiện nhanh hơn và ít tốn tài nguyên hơn so với DOM.
- ***XmlPullParser*** cho phép mã ứng dụng "kéo" hoặc tìm kiếm các sự kiện từ trình phân tích, trái ngược với SAX tự động đẩy các sự kiện cho trình xử lý. XmlPullParser cũng thể hiện được ưu điểm đơn giản trong sử dụng, ít tốn bộ nhớ và xử lý nhanh hơn DOM và SAX.
- *Jsoup* là một thư viện được sử dụng để phân tích tài liệu HTML, được cung cấp các API dùng để lấy dữ liệu và thao tác dữ liệu từ URL hoặc từ file HTML. Sử dụng các phương pháp DOM, CSS, JQuery để lấy dữ liệu và thao thác với dữ liệu.

Các ứng dụng di động ngày nay được xây dựng rộng khắp bởi sự phổ biến của các thiết bị di động cầm tay và Internet ngày nay. Áp dụng các kỹ thuật phân tích cấu trúc của website để xây dựng các ứng dụng cập nhật tin tức một cách tùy biến theo yêu cầu của người dùng sẽ được phát triển rộng trong thời gian tới.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. *DOM và SAX*, http://bodua.vn/knowledge/La%CC%80m-vie%CC%A3c-vo%CC%81i-XML/DOM-va%CC%80-SAX_72.html.

[2]. *NewsFeed*, https://play.google.com/store/apps/details?id=com.aaravmedia.newsfeed.

[3]. *Phương pháp Interface DOM phân tích RSS 2.0*,http://o7planning.org/vi/10109/phan-tich-xml-bang-cach-su-dung-mo-hinh-dom-trong-java#a6435.

[4]. *Phương pháp Interface SAX phân tích RSS2.0,* http://www.bodua.vn/knowledge/Java-Server-Pages/Du%CC%80ng-SAX-do%CC%A3c-XML_54.html.

[5]. *Phương pháp phân tích HTML Jsoup*, http://o7planning.org/vi/10399/huong-dan-su-dung-java-jsoup-phan-tich-html.

[6]. *RSS 2.0,* https://www.ibm.com/developerworks/vn/library/webservices/201301/x-rss20/.

[7]. *Ứng dụng đọc báo mới*, https://play.google.com/store/apps/details?id=com.epi&hl=vi.

[8]. *Ứng dụng đọc báo Tinmoi24h*, https://play.google.com/store/apps/details?id=mobi.fiveplay.tinmoi24h&hl=vi.

[9]. *Ứng dụng đọc tin tự động Việt Báo*, http://vietbao.vn/Vi-tinh-Vien-thong/RSS-Ung-dung-tuyet-voi-nhat/55135693/217/.

[10]. *XML*,https://www.ibm.com/developerworks/vn/edu/xmlintro/.

[11]. *XML structure*, https://www.ibm.com/developerworks/vn/library/12/x-androidxml/#listing13.

Tiếng Anh

[12]. *Development Org.XmlPull.V1.XmlPullParser NameSpace*, https://developer.xamarin.com/api/type/Org.XmlPull.V1.XmlPullParser/.

[13]. Gabe Beged-Dov, *JFinity Systems LLC... RDF Site Summary (RSS 1.0)*, http://web.resource.org/rss/1.0/spec#s9.

[14]. *Json*, http://www.w3schools.com/js/js_json_intro.asp.

[15]. *RSS 2.0 Specification*, http://cyber.law.harvard.edu/rss/rss.html#ltcategorygtSubelementOfLtitemgt.

[16]. *RSS2.0 XML Pull Parsing*,http://www.extreme.indiana.edu/xmlpull-website/index.shtml.

[17]. *Techniques for feedback*, http://www2.le.ac.uk/offices/red/rd/career-development/research-staff/mentoring/feedback.

[18]. *Techniques for Search engine*, http://www.htmlgoodies.com/beyond/seo/article.php/3837311/Search-Engine-Optimization-SEO-Advanced-Techniques.htm.