1. Giới thiệu
   1. Mở đầu
   2. Yêu cầu & mục tiêu đề tài
      1. Yêu cầu
      2. Mục tiêu

Hệ thống gồm 3 phần

* Extract
* Match
* Query (Search)

1. Phân tích
2. Dữ liệu cần extract:

Dựa trên yêu cầu của hệ thống: thu thập dữ liệu liên quan đến địa chỉ. Nhóm đã xác định các thuộc tính liên quan đến địa chỉ cần extract

* Tên cơ sở tại địa chỉ đã cho.
* Địa chỉ chính xác của cơ sở (VD: phòng, tầng …)
* E-mail (nếu có)
* Website (nếu có)
* Số điện thoại (nếu có)
* Số fax (nếu có)
* Ngành nghề kinh doanh (nếu có)

1. Extract

* Nguồn dữ liệu của hệ thống : web pages từ các website khác nhau.
* Thuận lợi: nguồn dữ liệu có cấu trúc.
* Page có 2 dạng:

+ List page: chứa 1 hoặc nhiều list, mỗi list gồm nhiều data record được định dạng theo cấu trúc giống nhau.

+ Detail page: page chứa thông tin chi tiết về 1 đối tượng

* mục tiêu là tự động extract dữ liệu từ cả 2 dạng page này.
* Các giải pháp hiện tại:

+ Manual approach

+ Wrapper induction

+ Automatic extraction

* + 1. Manual approach

Người lập trình quan sát webpage và source code của nó để tìm ra pattern, sau đó viết chương trình để lấy dữ liệu, có thể sử dụng các công cụ : xpath… để việc lấy dữ liệu dễ dàng hơn. Chỉ ứng với từng website cụ thể, không thể mở rộng.

* + 1. Wrapper induction

Dựa vào một số page được đánh dấu sẵn vị trí của data record, hệ thống tự động rút ra quy luật và dùng luật này để extract data từ những page có cấu trúc tương tự.

Nhược điểm:

- Tốn công sức và thời gian cho việc đánh dấu vị trí data record bằng tay.

- Chi phí cao khi áp dụng cho nhiều site khác nhau.

* + 1. Automatic extraction

Cho trước một hoặc một số page, hệ thống sẽ tự động tìm ra pattern để extract data. Cách tiếp cận này loại bỏ việc đánh dấu vị trí bằng tay, do đó có thể áp đụng cho 1 số lượng lớn page và site.

Một giải thuật thuộc cách tiếp cận này là giải thuật NET của tác giả Biu Ling. Giải thuật này dựa trên sự so trùng DOM tree để tìm ra pattern cho việc extract data. Giải thuật áp dụng cho list page, tuy nhiên có thể dễ dàng điều chỉnh để áp dụng cho detail page bằng cách tạo một DOM tree từ các detail page này.

Nhược điểm dễ nhận thấy của giải thuật là chưa thể tạo được attribute name cho dữ liệu được extract => cần có cơ chế nhận dạng, gán nhãn cho dữ liệu. Nhóm sử dụng heuristic dựa trên các đặc điểm của dữ liệu để làm việc này.

1. Match
   * 1. Cơ sở dữ liệu hiện có

Cơ sở dữ liệu hiện có là một DB chứa đựng thông tin về các khu đất được đánh địa chỉ tại TpHCM kèm theo tọa độ cụ thể của khu đất.

Chi tiết về cơ sở dữ liệu này được biểu diễn trong ERD dưới đây

ERD trên đã được map thành các bảng như sau:

* + 1. Cách thức map:

Việc map các đối tượng dữ liệu đã được extract vào cơ sở dữ liệu hiện có dựa trên thuộc tính về địa chỉ chính xác của đối tượng dữ liệu.

Quá trình map được thực hiện bằng 2 bước:

- Phân tích chuỗi địa chỉ chính xác để có được số nhà, tên đường, phường, quận.

- Map đối tượng dữ liệu vào 1 địa chỉ trong cơ sở dữ liệu hiện có dựa trên 4 thuộc tính vừa được phân tích.

1. Query

Dựa trên yêu cầu xây dựng 1 website cho phép người dùng tìm kiếm trên dữ liệu địa chỉ vừa extract. Nhóm dự định xây dựng website có các chức năng như sơ đồ use-case dưới đây

Trong đó:

* Chức năng 1:
* Chức năng 2:

1. Giải thuật NET:
2. Tree Matching và giải thuật Simple Tree Matching
3. Multiple Alignment và giải thuật Partial Tree Alignment
4. Giải thuật NET
5. Thiết kế và hiện thực:
   1. Tổng quan hệ thống

Hệ thống gồm 2 hệ thống con Extractor và Query như hình vẽ sau

Database

Match

Query

(Website)

Extractor

Extractor rút trích thông tin địa chỉ từ nguồn dữ liệu về database.

Website sử dụng database để cung cấp thông tin cho người sử dụng.

* 1. Công cụ sử dụng

Python

ASP.NET

Microsoft SQL Server

* 1. Extractor & Match
     1. Thiết kế kiến trúc

Sub-system extractor gồm các module sau

- Matching: hiện thực giải thuật Simple Tree Matching

- Alignment: hiện thực giải thuật Partial Tree Alignment

- Wildcard: chứa các hàm support cho việc hiện thực wildcard trong pattern

- Label: các hàm dùng gán nhãn dữ liệu

- Support: các hàm phụ trợ khác

* + 1. Thiết kế database

ERD

Bảng

* + 1. Hiện thực
  1. Query
     1. Cơ sở dữ liệu

Phần website sử dụng cơ sở dữ liệu phần thiết kế Extractor đã tạo, để tiện trong việc truy vấn, nhóm đã tạo các view sau

- View 1

- View 2

- View 3

* + 1. Hiện thực

- Chức năng 1

- Chức năng 2

1. Sử dụng:
2. Extractor
3. Website
4. Đánh giá:
5. Các kết quả đạt được
6. Các hạn chế của hệ thống
7. Hướng phát triển