2011 International Conference on Computer and Software Modeling
IPCSIT vol.14 (2011) © (2011) IACSIT Press, Singapore

A Vietnamese Query Processing Based Bus Routes Searching System

Dang Tuan Nguyen, An Quoc Truong, Hung Le Truong+

Natural Language and Knowledge Engineering Research Group
University of Information Technology (VNU-HCM)
Ho Chi Minh City, Vietnam

Abstract. In this paper, we introduce our initial results in the aim of building a Vietnamese query processing based bus routes searching system in Ho Chi Minh City. We apply a simple, empirical and intuitional searching method to look for bus routes with two objectives: the itinerary must have the shortest moving time among itineraries whose times of route change are relatively few. We have built a bus routes searching system based on Vietnamese query processing for the bus network in HCMC. The experimental results indicate that the system can respond to the practical requires of testing. This paper focuses on the overall architecture of bus itinerary searching system, the modeling of the bus system in HCMC, the bus routes searching method and the approach in processing of Vietnamese queries.

Keywords: Natural Language Processing, Search, Vietnamese

1. Introduction

In this paper, we introduce our initial results in the aim of building a bus routes searching system in Ho Chi Minh City (HCMC). A particular feature of this searching system is to support passengers to use simple Vietnamese questions to search. With the searching system, passengers do not have to use keywords to search for bus itineraries on fixed data fields of a database search engine. The system will suggest passengers reasonable itineraries which are ranked by priority rates based on predefined criteria.

In principle, the bus routes system, bus stops and terminals of bus network of HCMC can be modeled as a graph. However, the operation of the bus system of HCMC obeys its own rules. For example, many bus routes are able to run in the same road, but some of the bus stops in that road cannot be used in the itinerary of some bus routes.

Assume that the bus network is modeled as a directed and weighted graph, the application of traditional algorithms, such as Dijkstra, just allows identifying the shortest path from a start bus stop to an end bus stop. Passengers do not know which bus route is the shortest one. In addition, passengers may have to change bus routes many times on the shortest one, and that is inconvenient for passengers.

With the above considerations, we observe that traditional graph search algorithms cannot satisfy all objectives of searching bus routes. Thus, we apply a simple, empirical and intuitional searching method to look for bus routes with two objectives: the itinerary must have the shortest moving time among itineraries whose times of route change are relatively few.

This paper focuses on the overall architecture of bus itinerary searching system, the modeling of the bus system in HCMC, the bus routes searching method and the approach in processing of Vietnamese queries.

+ Email: {ntdang, tqan, tlhung}@nlke-group.net

2. Modelling The Bus Network

The bus network consists of bus routes. Each of bus routes always has an outward trip and a return trip. With the same bus route, the outward trip and the return trip possibly have the same, common or totally different bus stops. Besides, the different bus routes may also intersect each other at some common bus stop. The bus system also contains bus stops, bus stations. Each of bus routes has many buses. Buses, one by one, leave the station with a predetermined time-table. The period of time between a bus and the next one leaving the station is called time of stretch.

Based on the data and information resources provided by the Public Passenger Transport Management and Control Center of HCMC [7], we have built a database to organize and store data/information of the bus network in HCMC. The stored information includes:

- Route: Route consists of a start station, streets or special locations where buses passing and an end station.
- Street: There can be one or many bus stops in a street that a bus itinerary gets through.
- Bus stop: The related information of the bus stop is:
  - Location: location is identified by a specific house numbering or a specific place.
  - District.
  - Code of other buses passing this bus stop (passengers can change route here).

The bus network of HCMC can be modeled as a directed, weighted graph G = (V, E). In this model, bus stops, bus terminal, and bus stations are vertexes of set V in graph G. The lines link two consecutive bus stops together through a special location (called neighbor bus stops of a special place, passengers can walk between them) are edges of set E in graph G.

Because of having no information about the distances between bus stops in bus network, the weight of each edge of set E in graph G is identified by the time of moving between two consecutive bus stops of a route. To set up the moving time between two consecutive bus stops of a route, we rely on the following assumptions:

- *Assumption 1: On the same route, assume that the average time of moving between two consecutive bus stops of a bus route is equal. Therefore, the time of moving between two consecutive bus stops is: the total time of a route ÷ (the number of bus stops of the route – 1).*
- *Assumption 2: With the stations connected together by a special location, the weight of these edges are identified manually depending on the location.*

In addition, the system also has a function that allows the administrator of the system to adjust the time of moving between two consecutive bus stops.

3. Architecture of Bus Routes Searching System

The architecture of bus routes searching system based on Vietnamese query processing is composed of the following main modules:

- Module 1: Vietnamese Query Processing.
- Module 2: Bus Itinerary Searching.
- Module 3: Bus Information Querying.
- Module 4: Bus Itinerary Guiding.

Besides the above main modules, there is a database storing information of bus network in the architecture.

The bus routes searching system operates in these principles:

- Receiving a user's Vietnamese question.
- Processing Vietnamese questions: the system analyzes the syntactic of Vietnamese question, identifies the meanning of the question.
- If the question requests to provide a detailed information of bus network, the system relies on the semantics of the question to query the database

- If the question requests to search a bus itinerary, the system relies on the semantics of the question to look for the itinerary with two separated criteria: the itinerary must have the shortest moving time among itineraries whose times of route change are relatively few.
- Itinerary guiding: the system selects, normalizes and ranks found itineraries, then represents the itinerary guides in texts.

4. Searching Bus Itineraries

The bus itinerary searching module accepts inputs that are two list of candidate bus stops: the former is start bus stops, and the latter is end bus stops. This module has functions as looking for possible bus itineraries, ranking found itineraries by predefined criteria, selecting the best itineraries and returning the results.

Based on bus network model, we use a simple, empirical and intuitional searching method to search bus itineraries. This searching method is appropriate for the characteristics of the bus system in HCMC.

4.1. Searching Assumptions

The bus routes searching strategy is based on these assumptions:

- *Assumptions 3: Only searching itineraries that have the maximum times of changing bus routes are 2 times, exclusive of choosing the start route.*
- *Assumptions 4: The best route is a route that satisfies assumption 3 and has the shortest time of moving.*

4.2. An Empirical and Intuitional Searching Method

Based on identified searching strategy, we have built a searching method with three steps as following:

- Step 1: Searching itineraries which require passengers only to select start route and do not change the route.

  For each start bus stop, considering all routes possibly starting from this, if which route can arrive at one of the end bus stops in the list, enregistering the itinerary and compute the time of moving.
- Step 2: Searching itineraries which require passengers to select start route with one time of route change.

  For each end bus stop, considering all routes possibly arriving at this, enregister the itinerary if which one intersects start routes from start bus stops and compute the time of moving.
- Step 3: Searching itineraries which require passengers to select start route with two times of route change.

Considering all bus routes possibly starting from start bus stops (group 1), all bus routes possibly arriving at end bus stops (group 2) and all intermediate bus routes joining one route in group 1 and one route in group 2. Enregister these itineraries and compute the time of moving for each itinerary.

In each step, we select three itineraries having the shortest time of moving. At the end, we select the best of these itineraries. In case we do not find any itineraries after doing three steps above, it means that there is possible no itinerary which lets a bus move from one of the start bus stops to one of the end bus stops, or there are itineraries having more than 2 times of route change.

Example 1: An illustration of the result of searching route from a start location to a destination location.

```
Start Location: Bến xe miền đông (640 641 642 647 649 3321 3322 3328 3330 3833 3836 3856 3857 P21 )
Destination Location: Trường Đại học Sài Gòn (73 76 77 3038 3060 3088 3089 3092 3094 )
Total itineraries: 2
Guided itinerary: about 51 minutes
        Đi đến trạm 75 Nguyễn Xí, Quận Bình Thạnh
        Bắt tuyến số 14
        Xuống ở trạm 466 Nguyễn Thị Minh Khai, Quận 3
        Bắt tuyến số 38 hoặc 6
        Xuống ở trạm 227 Nguyễn Văn Cừ, Quận 5
```

5. Processing Vietnamese Queries

Based on the approach of existing research results ([2], [3]) in processing Vietnamese queries in an open coursewares retrieval system, we process Vietnamese queries in the bus routes searching system following these steps:

- Step 1: Building a restricted parser based on ANTLR [5]
  + Identifying Vietnamese question forms that our system can process.
  + Defining syntactic rules follows CFG (Context-Free Grammar) [1]; these rules are described by EBNF (Extended Backus–Naur Form) [6].
  + Using ANTLR [5] to build a restricted parser.

- Step 2: Syntactic analysis of Vietnamese questions
  + Using the restricted parser to analyze a Vietnamese question.
  + Building principles to transform the syntactic tree of the Vietnamese question to corresponding database queries in SQL Language (in the case of querying information of the bus network), or defining the set of start bus stop and the set of end bus stop (in the case of searching bus itinerary)

Example 2: A đi B?

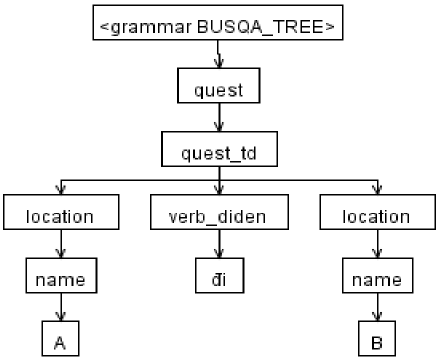

Fig. 1: The syntactic tree of the Vietnamese question in example 2

In this paper, we only introduce our approach in processing Vietnamese questions. The details of syntactic and semantic analysis of the Vietnamese queries, as well as the method for generating SQL queries from the semantic structures, will be represented in the next paper.

6. Conclusions

We have built a bus routes searching system based on Vietnamese query processing for the bus network in HCMC. This system applies an empirical and effective method for searching bus itineraries. This searching method is implemented in a bus network model in which we do not use the distances between two bus stops as weight of corresponding edges of a directed and weighted graph. The efficiency of this method depends on initial assumptions we have supposed.

We have done an experiment of 100 tests with this system. The results are evaluated manually. The experimental results indicate that the system can respond to the practical requires of testing. The system can search itineraries which the times of route change are not more than 2, exclusive of the start route, then it ranks these itineraries by time of moving criteria to select the best route. However, the system should be improved in processing more complex and various Vietnamese questions in future.

7. References

[1] N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*. 1956, 2(3): 113–124.

[2] D. T. Nguyen, H. Q.-T. Luong. Document searching system based on natural language query processing for Vietnam Open Courseware library. *International Journal of Computer Science Issues (IJCSI).* November 2009, 6(2): 7-13.

[3] Lương Quý Tịnh Hà. *Xây dựng công cụ tìm kiếm tài liệu học tập bằng các truy vấn ngôn ngữ tự nhiên trên kho học liệu mở tiếng Việt*. M.Sc. Thesis in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2009.

[4] Trương Quốc An, Trương Lê Hưng. *Mô hình tìm kiếm lộ trình xe bus bằng truy vấn tiếng Việt với tính năng chỉ dẫn trực quan trên bản đồ*. B.Sc. Thesis in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2011.

[5] ANTLR. [Online]. http://www.antlr.org/

[6] J. A. Farrell. August 1995. [Online]. http://www.cs.man.ac.uk/~pjj/farrell/comp2.html

[7] Trung Tâm Quản lý và Điều hành Vận tải Hành khách Công cộng. [Online]. http://www.buyttphcm.com.vn/