Department of Computer Science,
Faculty of Mathematics, Physics and Informatics,
Comenius University in Bratislava

Distance oracles for timetable graphs

(Master thesis)

bc. František Hajnovič

Study program: Computer science
Branch of study: 2508 Informatics
Supervisor: doc. RNDr. Rastislav Královič, PhD.

Bratislava 2013

65787891

Comenius University in Bratislava
Faculty of Mathematics, Physics and Informatics

THESIS ASSIGNMENT

Name and Surname: Bc. František Hajnovič
Study programme: Computer Science (Single degree study, master II. deg., full time form)
Field of Study: 9.2.1. Computer Science, Informatics
Type of Thesis: Diploma Thesis
Language of Thesis: English
Secondary language: Slovak

Title: Distance oracles for timetable graphs

Aim: The aim of the thesis is to explore the applicability of results about distance oracles to timetable graphs. It is known that for general graphs no efficient distance oracles exist, however, they can be constructed for many classes of graphs. Graphs defined by timetables of regular transport carriers form a specific class which it is not known to admit efficient distance oracles. The thesis should investigate to which extent the known desirable properties (e.g. small highway dimension) are present int these graphs, and/or identify new ones. Analytical study of graph operations and/or experimental verification on real data form two possible approaches to the topic.

Supervisor: doc. RNDr. Rastislav Královič, PhD.
Department: FMFI.KI - Department of Computer Science
Vedúci katedry: doc. RNDr. Daniel Olejár, PhD.

Assigned: 08.11.2011

Approved: 15.11.2011 prof. RNDr. Branislav Rovan, PhD.
Guarantor of Study Programme

....

Student Supervisor

65787891

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. František Hajnovič
Študijný program: informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: 9.2.1. informatika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: anglický
Sekundárny jazyk: slovenský

Názov: Efektívny výpočet vzdialeností v grafoch spojení lineik.

Cieľ: Cieľom práce je preštudovať možnosti aplikácie výsledkov o distance oracles v grafoch reprezentujúcich dopravné siete na grafy spojení liniek. Otázka, či a aké dôležité vlastnosti ostávajú zachované sa dá riešiť teoreticky pre rôzne triedy grafov a/alebo experimentálne pre reálne dáta.

Vedúci: doc. RNDr. Rastislav Královič, PhD.
Katedra: FMFI.KI - Katedra informatiky
Vedúci katedry: doc. RNDr. Daniel Olejár, PhD.

Dátum zadania: 08.11.2011

Dátum schválenia: 15.11.2011

prof. RNDr. Branislav Rovan, PhD.
garant študijného programu

....
študent

....
vedúci práce

I hereby declare that I wrote this thesis by myself, only with the help of the referenced literature, under the careful supervision of my thesis advisor.

....

Acknowledgements

I would like to thank very much to my supervisor Rastislav Královič for valuable remarks, useful advices and consultations that helped me stay on the right path during my work on this thesis.

I am also grateful for the support of my family during my studies and the work on this thesis.

František Hajnovič

Abstract

Queries for optimal connection in timetables can be answered by running Dijkstra's algorithm on an appropriate graph. However, in certain scenarios this approach is not fast enough. In this thesis we introduce methods with much better query time than that of the efficiently implemented Dijkstra's algorithm.

Our first method called *USP-OR* is based on pre-computing paths, that are worth to follow. This method achieves speed-ups of up to 70, although at the cost of high amount of preprocessed data. Our second algorithm computes a small set of important stations and additional information for optimal travelling between these stations. Named *USP-OR-A*, this method is much less space consuming but still more than 8 times faster than the Dijkstra's algorithm on some of the real-world datasets.

Other contributions of this thesis are

Key words: **optimal connection**, **timetable**, **Dijkstra's algorithm**, **Distance oracles**, **underlying shortest paths**

Abstrakt

V tejto práci sa zaoberáme hľadaním optimálnych spojení v cestovných poriadkoch, na ktorých sme si predpočítali určité informácie. Na základe analýzy reálnych cestovných poriadkov sme vyvinuli exaktné metódy, ktoré na dotaz na optimálne spojenie odpovedajú podstatne rýchlejšie ako časovo závislá implementácia Dijkstrovho algoritmu využívajúca prioritnú frontu na základe Fibonacciho haldy. Presnejšie, náš algoritmus *USP-OR-A* s priestorovou zložitosťou $\mathcal{O}(n^{1.5})$ dosahuje časovú zložitosť odpovede na dotaz $\mathcal{O}(\sqrt{n}\log n)$, prekonávajúc časovo závislý Dijkstrov algoritmus takmer 7 krát v našom najväčšom cestovnom poriadku.

Kľúčové slová: **optimálne spojenie**, **cestovný poriadok**, **Dijkstrov algoritmus**, **Dištančné orákulá**, **podkladové najkratšie cesty**

Contents

1 Introduction

2 Preliminaries

3 Related work

4 Data & analysis

In this section we would like to introduce the timetable datasets we were working with and provide the analysis of their properties. The main reason for this analysis is that it gives some insight into the characteristics of the timetables and so may contribute to develop an oracle based method with better qualities.

4.1 Data

We have obtained timetable datasets from numerous sources, in varying formats and of different types. Some of them were freely available on the Internet while others were provided by companies upon demand. Let us provide their brief description.

The dataset *air01* contains schedules of **domestic flights in United States** for the January of 2008. It is not comprehensive in the sense that it contains entries only for flights of some of the major airports in US. However it is large enough for our purposes (almost 300 airports). This dataset is just a fraction of the data that are freely available at the pages of American Statistical Association [1] in CSV format.

Timetable *cpsk* represent the **regional bus** schedules from the areas of **Ružomberok and Žilina, Slovakia**. The data were provided by the company in charge of the *cp.sk* portal - Inprop s.r.o. . The timetable contains about 1900 bus stops and came in a JDF 1.9 format [2]. Apart from the actual schedules, the data in JDF contain numerous other information which were not relevant for our purposes. From both timetables we have extracted subsets with a time range of one day.

The *gb-coach* and *gb-train* timetables are freely available from National Public Transport Data Repository (NPTDR) [3] in an ATCO-CIF format. These are not actually timetables but rather weekly snapshots of national public transport journeys made by **coach and train in Great Britain** (during certain week in year 2011). The datasets contain about 2500 stations each.

The *montr* dataset is part of a public feed for **Greater Montreal public transportation**, available at Google Transit Feeds [4]. The data are in a GTFS format (defines relations between CSV files listing stations, routes, stop-times...) and were made available by Montreal's Agence métropolitaine de transport. Our timetable *montr* corresponds to daily schedules of the Chambly-Richelieu-Carignan bus services (more than 200 bus stops).

Also in GTFS format come the data of **French railways** operated by company SNCF, publicly available at their website [5]. The schedules are weekly and there were two of them: one for intercity trains and one for TER trains (regional trains). Thus the three timetables *sncf-inter* (366 stations), *sncf-ter* (2637 stations) and their union *sncf* (2646 stations).

Finally, one more country-wide railway timetable was provided by ŽSR, the company in charge of the **Slovak national railways**. This timetable was exported in a MERITS format and its time range is for one year. The number of stations in *zsr* dataset is 233.

With the help of Python and Bash scripts, we converted each of these datasets to our timetable format (described in appendix A). This timetables were then loaded by our application TTBlazer,

[1] `http://stat-computing.org/dataexpo/2009/the-data.html`
[2] Jednotný dátový formát (JDF).
[3] `http://data.gov.uk`
[4] `http://code.google.com/p/googletransitdatafeed/wiki/PublicFeeds`
[5] `http://test.data-sncf.com/index.php/ter.html`

which can further generate sub-timetables (with less stations or smaller time range), underlying graphs and TE and TD graphs.

For a summary of the used timetables' descriptions, see table ?? and for their main properties, refer to table ??.

Name	Description	Format	Provided by	Publicly available
air01	domestic flights (US)	CSV	American Stat. Assoc.	✔
cpsk	regional bus (Ružomberok & Žilina, SVK)	JDF 1.9	Inprop s.r.o.	✗
gb-coach	country-wide buses (GB)	ATCO-CIF	NPTDR	✔
gb-train	country-wide rails (GB)	ATCO-CIF	NPTDR	✔
montr	public transport (Montreal, CA)	GTFS	Montreal AMT	✔
sncf	country-wide rails (FRA)	GTFS	SNCF	✔
zsr	country-wide rails (SVK)	MERITS	ŽSR	✗

Table 4.1: Datasets descriptions.

Name	El. conns.	Cities	UG arcs	Time range	Height
air01	601489	287	4668	1 month	24374
cpsk	97916	1905	5093	1 day	370
gb-coach	260710	2448	5793	1 week	3140
gb-train	1714535	2555	8335	1 week	7978
montr	7153	217	349	1 day	363
sncf	416302	2646	7994	1 week	2679
sncf-inter	22750	366	901	1 week	1052
sncf-ter	393587	2637	7647	1 week	2646
zsr	932052	233	588	1 year	60308

Table 4.2: Main properties of the timetables. The value of time range is approximate.

To see better the differences in the properties of different timetable types (train, flight, bus...), we made sub-timetables with 200 cities and with the upper bound on time range being 1 day and 6 hours [6] ($thigh_T <$ 1 day and 6 hours) from each of our dataset. We name these datasets by appending to the original name "*-200d*" [7]. See table ?? for details.

Name	El. conns.	Cities	UG arcs	Height
air01-200d	19010	200	3973	772
cpsk-200d	14747	200	592	370
gb-coach-200d	2760	200	564	498
gb-train-200d	24323	200	792	957
montr-200d	6841	200	320	355
sncf-200d	4192	200	611	269
sncf-inter-200d	2172	200	493	128
sncf-ter-200d	8469	200	600	419
zsr-200d	2031	200	454	133

Table 4.3: 200-station sub-timetables with the time range of one day.

Also, to further justify our choice of using TD graphs instead of TE graphs in this thesis, we provide

[6] We took all elementary connections that were within our time range. From this timetable, we made an UG and its (random) sub-graph of 200 cities. Finally we selected only those elementary connections, that were on top of this sub-graph to form a timetable with 200 cities and the desired (maximal) time range.

[7] Similarly, throughout this thesis, suffix "*-d*" would mean "with daily time range", "*-w*" "weekly time range" and suffix "*-#*" would mean sub-timetable with # stations.

their space consumption comparison in table **??**.

	TD graph			TE graph		
Name	**Nodes**	**Arcs**	**Size (MB)**	**Nodes**	**Arcs**	**Size (MB)**
air01	287	4668	27	715211	1307432	72
cpsk	1905	5093	5	95601	189205	11
gb-coach	2448	5793	12	259589	512862	32
gb-train	2555	8335	79	2042316	3745751	263
montr	217	349	0.4	7182	13992	0.9
sncf	2646	7994	19	758867	1166646	85
sncf-inter	366	901	1.1	39765	60602	4.6
sncf-ter	2637	7647	18	720651	1107301	81
zsr	233	588	42	1706077	2637896	173

Table 4.4: Space consumption of time-dependent vs. time-expanded model. The number of nodes and arcs for TD graph is the same as for the corresponding underlying graph.

4.2 Analysis of properties

First we will take a look at the optimal connection *sizes* (size is the number of elementary connections) in the timetables. For a given timetable T, we will denote the average optimal connection size as $\boldsymbol{\gamma_T}$ and will call it the **optimal connection diameter** (OC diameter). We computed an approximate OC diameter for each of our datasets by measuring an average connection size of sufficiently many OCs. The results in table **??** indicate that the average OC size generally falls under $\sqrt{n}$.

Next we would like to get an idea of the sparsity of the underlying graphs. We see from the table **??** that the graphs are pretty sparse (with the exception of *air01*), but we would like to make sure that the sparsity is uniform. More specifically, we will be interested in the δ-density:

Definition 4.1. ***δ-density***
A graph G of n vertices and m arcs is δ-dense $\iff \forall G' \subseteq G, n' \geq \sqrt[4]{n} : \frac{m'}{n'} \leq \delta$

- *For a timetable T, we will denote its* **density** *parameter* [8] *as* $\boldsymbol{\delta_T} = min\{\delta |\ ug_T \text{ is } \delta\text{-dense}\}$

To find out at least approximate δ_T values for our timetables, we have randomly sampled their UGs for (connected) sub-graphs of various sizes (starting from $\sqrt[4]{n}$ [9]). In table **??** you can see the maximal density found during the sampling.

The density is related to the **average degree** deg_{avg} in the UG, since in oriented graphs:

$$deg_{avg} = \frac{m}{n}$$

So the average degree is a lower bound on the graph's density. Table **??** lists the average and maximal degrees in the underlying graphs.

We would also assume, that the underlying graphs of each timetable will be **connected** (and even strongly connected), or at least that the largest connected component spans almost the whole graph.

[8]Note that this has nothing to do with the frequency of elementary connections, only with the density of the underlying graph.

[9]The choice of $\sqrt[4]{n}$ will be justified later, during the analysis of the algorithms.

Name	γ_T	Max. OC size found	$\sqrt{n}$
air01	2.4	8	16.9
cpsk	40.8	162	43.6
gb-coach	25.2	128	49.5
gb-train	25.6	111	50.5
montr	21.1	63	14.7
sncf	36.8	111	51.4
sncf-inter	17.1	58	19.1
sncf-ter	48.0	167	51.3
zsr	15.0	57	15.3

Table 4.5: With one exception, OC diameter is less then $\sqrt{n}$ (this was expected, as *montr* is the only timetable with "geographically one dimension long" - all other timetables span areas with more uniform shape). Note extremely low value for airline timetable - this is due to the fact that UGs of airline timetables have small-world characteristics [?]. Another thing we may notice is that regional timetables (*cpsk*, *sncf-ter*) have higher OC diameter then country-wide and inter-city timetables. We also point out that the inter-city trains in French railways decrease the average optimal connection size by one about third.

Name	Maximal δ_T found
air01	34.5
cpsk	4.1
gb-coach	5.0
gb-train	5.8
montr	1.9
sncf	5.0
sncf-inter	3.0
sncf-ter	4.8
zsr	3.2

Table 4.6: Approximate density of the underlying graphs.

Name	Avg. degree	Max. degree
air01	16.3	166
cpsk	2.7	27
gb-coach	2.4	103
gb-train	3.3	30
montr	1.6	5
sncf	3.0	27
sncf-inter	2.5	12
sncf-ter	2.9	27
zsr	2.5	12

Table 4.7: Average and maximal degree in the underlying graphs.

From the table **??** we may see that this assumption holds.

		Connectivity		Strong connectivity	
Name	n	Connected	Largest comp.	Connected	Largest comp.
air01	287	✔	287	✗	286
cpsk	1905	✔	1905	✗	1903
gb-coach	2448	✗	2374	✗	2332
gb-train	2555	✔	2555	✔	2555
montr	217	✗	211	✗	209
sncf	2646	✔	2646	✗	2594
sncf-inter	366	✗	328	✗	316
sncf-ter	2637	✔	2637	✗	2583
zsr	233	✔	233	✗	225

Table 4.8: Connectivity of underlying graphs.

In the previous section 3 we have mentioned the highway dimension [?] as a parameter which, when being low, guarantees low query times for certain route-planning methods. Here we were interested in the highway dimension of our underlying graphs.

Definition 4.2. ***Highway dimension***
Highway dimension $\boldsymbol{HD(G)}$ *for a directed, edge-weighted graph* $G = (V, E)$ *is the smallest integer* h*, such that:*

$$\forall r \in R^{+},\ \forall u \in V,\ \exists S \subseteq B_{u,2r},\ |S| \leq h,\ \forall v, w \in B_{u,2r}:$$
$$\textit{if } r < |P(v,w)| \leq 2r \textit{ and } P(v,w) \subseteq B_{u,2r} \textit{ then } P(v,w) \cap S \neq \emptyset$$

where:

- $\boldsymbol{P(v,w)}$ *is the* ***shortest path*** *between v and w*
- $\boldsymbol{B_{u,r}} = \{v \in V |\ |P(u,v)| \leq r$ *or* $|P(v,u)| \leq r\}$ *and is called* ***ball*** *of radius r centred at u.*

Intuitively, a graph has a low HD, if for any r we have a *sparse* set of vertices S_r, such that every shortest path longer then r includes a vertex from S_r. By the set being sparse, we mean that every ball of radius $\mathcal{O}(r)$ contains just a few elements of S_r.

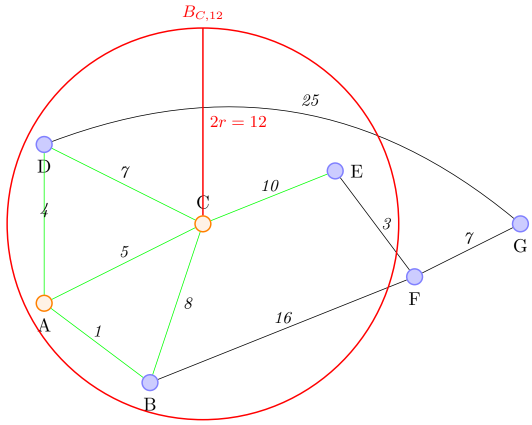

Figure 4.1: Demonstration of a definition of HD. We chose some r ($r = 6$) and some vertex v ($v = C$) to root the ball $B_{v,2r}$. All the shortest paths *longer* than r *inside* the ball have to contain a vertex from S (orange vertices C and A in our case). The upper bound on $|S|$, considering any ball with any radius, is the required highway dimension. Note: in our case, we had to choose also A to set S, since a shortest path from B to D does not include C.

5 Underlying shortest paths

6 Neural network approach

7 Application TTBlazer

8 Conclusion

Appendices

A File formats