Метода за екстракцију обележја из “CDR – Call Detail Records” података

Одговорно лице: Сања Брдар

Аутори: Сања Брдар, Катарина Гаврић, Предраг Лугоња, Дубравко Ћулибрк, Владимир Црнојевић

Развијено: ИИИ 44003 - Интегрисани систем за детекцију и естимацију развоја пожара праћењем критичних параметара у реалном времену (руководилац пројекта Владимир Црнојевић)

Година: 2013

Примена од: 15.05.2013.

Кратак опис

Основу мобилне телекомуникационе мреже чини скуп базних станица. Свака базна станица описана је географским координатама (ширина и дужина). Свакодневно телекомуникациони оператори прикупљају огромну количину података о корисницима. Посебно интересантни су подаци који се чувају у формату *CDR- Call Detailed Records*. *CDR* подаци садрже информације о врсти и трајању обављених телекомуникационих услуга. Слог такође садржи географске локације базних станица преко којих је комуникација остварена. Сваки пут када корисник започне или прими позив или користи неку услугу (*SMS*, *MMS*, итд.) у бази оператера генерише се *CDR* слог. Чувају се само мета подаци о активности (тип, трајање, анонимни идентификатори учесника у комуникацији), док се садржај комуникационе активности не чува. Обрада *CDR* података има велики значај у анализи, разумевању и интерпретацији динамике активности корисника и њихове мобилности.

Техничко решење описује развијену методу за екстаракцију обележја из *CDR* података. Два типа обележја: активност и мобилност корисника се екстракују. Активност корисника се екстракује у облику просечног дневног профила – број позива по сатима или другим временским интервалима. Добијени профили откривају типичне криве активности корисника, али могу указати и на скривене трендове, аномалије или проблеме. Обележја која се односе на мобилност корисника екстракују се из географских позиција. Трајекторије кретања људи показују висок степен временске и просторне регуларности [1], посебно када се посматрају најфреквентније локације (кућа, посао) на којима проводе највише времена. Обележја екстракована из *CDR* података представљају добру апроксимацију кретања. Просторна резолуција одређена је распоредом базних станица у телекомуникационој мрежи.

Информације добијене из *CDR* података могу бити корисне у многим доменима, посебно

за унапређење саобраћаја: развој интелигентног транспортног система, подршка у навигацији, решавање проблема загушења саобраћаја, планирање нове инфраструктуре, оптимизација јавног саобраћаја. Друге значајне области примене су урбано планирање, јавно здравље и економско планирање. Такође се може употребити за побољшање рада хитних служби у реалном времену, јавну безбедност, емитовање обавештења за кориснике у близини најугроженијих места у ванредним ситуацијама.

Техничке карактериситке: Метода за екстракцију обележја имплементирана је у програмском језику *Python* (http://www.python.org), подаци се складиште у *MySql* базу података, фреквентне трајекторије се добијају помоћу софтвера *CommonGIS*. Метода је тестирана на подацима телекомуникационог оператера *France Telecom Orange.*

Реализатор: ФТН – Нови Сад

Корисник: Катедра за Телекомуникације, Факултет техничких наука, Нови Сад

Подтип решења: Нова метода (М85)

Стање у свету

Феномен великих података *„Big Data“* подразумева не само велику количину података него и њихову хетерогеност и стално генерисање нових података. Посебно сложени су просторно-временски подаци, који су окарактерисани како временским одредницама, тако и просторним координатама. Модерни мобилни уређаји чине ове податке све доступнијим за широк спектар апликација. Да би се у потпуности искористио потенцијал тих података потребне су специјализоване методе за њихову обраду. Због тога је ова врло атрактивна област истраживања привукла бројне истраживаче који предлажу методе за анализу и нове области примене.

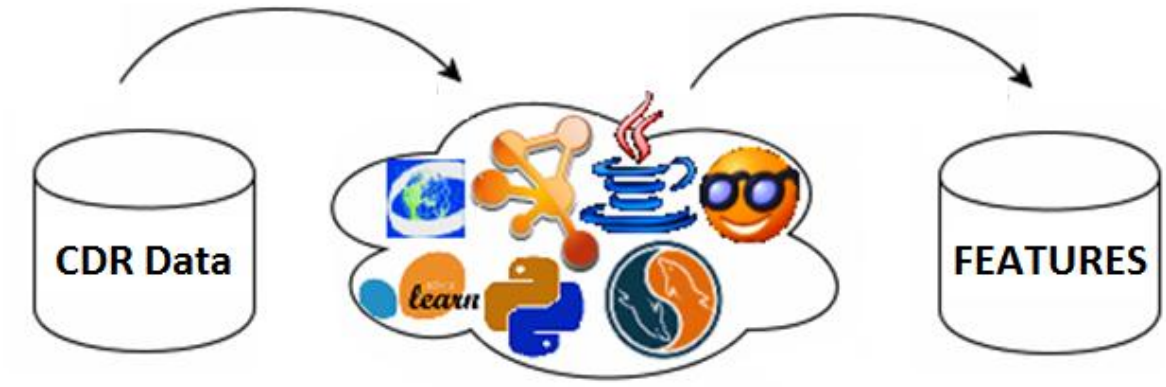
Истраживачка група *SENSEable City* са *МИТ*-а извршила је географско мапирање употребе мобилног телефона у различито доба дана на подручју града Милана [2]. Добијени резултати омогућавају графички приказ интензитета урбаних активности и њихове еволуције кроз простор и време. Анализа геопросторних информација се такође показала изузетно корисном за хуманитарне апликације. Истраживачи са *Karolinska* института су анализирали *CDR* податке пре и после разорног земљотреса на Хаитију са циљем да се процени обим и трендови кретања након катастрофе великих размера [3]. Утврдили су да се радијус и просечна дневна растојања која људи прелазе увећава, али да остају регуларности што доприноси предвидљивости у кретњу. Тако праћење кретања становништва и њиховог расељавања након земљотреса даје могућност усмеравања хуманитарних активности. У *IBM* центру развијена је метода за естимацију путања које људи прелазе у граду помоћу информација садржаних у *CDR* подацима [4]. Први пут су ови подаци коришћени за потребе оптимизације градске саобраћајне мреже. Њихов систем естимира места полазишта и одредишта и број људи које иду добијеним трајекторијама, затим се врши оптимизацаија и даје процена које нове руте би највише побољшале постојећи систем (смањити време чекања и дужину путовања).

Метода коју описујемо овим техничким решењем издваја просторно-временска обележја из *CDR* података и може бити примењена за било коју од претходно наведених области. Методом су дефинисане сложене трансформације података које од иницијално неструктуираних података издвајају корисне информације о активности и мобилности корисника.

Конструкција и принцип рада

На слици 1. приказана је логичка шема предложене методе за екстракцију. Метода обухвата следеће модуле:

- Учитавање података у трансакциону базу података
- Агрегација података и статистичка анализа - откривање скривених структура, објашњење и сумирање кључних карактеристика у подацима.
- Екстракција обележја из података (излаз може бити у облику фајла или графички)



Сл. 1. Шема

CDR подаци се смештају у *MySql* базу. Иницијални подаци у *CSV* форми се *SQL* скриптама пребацују у припремљену шему базе података. Затим се *SQL* скриптама подаци агрегирају у посебне табеле. Агрегација се врши на основу временских и просторних одредница. На пример сумирају се или упросечавају активности корисника (број или трајање успостављених позива) по различитим данима: викенд или радни дан; по различитим деловима дана: радни часови, слободно време, ноћни сати. Просторни параметри по којима се агрегирају подаци су задати региони, административне јединице или локације базних станица. Екстракција обележја врши се *Python* скриптама. Део *Python* скрипте за екстракцију обележја издвојен је на слици 2. За успостављање конекције са базом користи се *Orange* модул. Након конекције могуће је извршавати селекцију података из базе дефинисаним *SQL* упитима. На примеру је дато једноставно селектовање свих корисника из базе. Пример такође обухвата екстракцију једног просторног обележја кретања – *gyration*, који за сваког корисника на основу свих просторних тачака не којима је био

активан израчуна стандардну девијацију од просечне локације. За мерење растојања међу тачкама коришћен је модул *geopy*, а за рачунање потребних статистика модул *numpy*.

```
# Database connection through Orange python module
from Orange.data.sql import *
r = SQLReader()
r.connect('mysql://root@localhost/CDR')
…
# Extract users from system
r.execute("SELECT distinct(user_id) FROM cdr;")
data = r.data()
users = [int(user[0].value) for user in data]
…
# Extract gyration feature, Input: lists of latitudes and longitudes - lat_list and lon_list

import numpy as np
from geopy import distance
from geopy import Point


latc = np.average(lat_list)
lonc = np.average(lon_list)
pc = Point(str(latc) + ';' + str(lonc))
dist_list = []
for i in range(len(lat_list)):
    p = Point(str(lat_list[i]) + ';' + str(lon_list[i]))
    dist_list.append(distance.distance(p, pc).km)
gyration = np.std(dist_list)
```

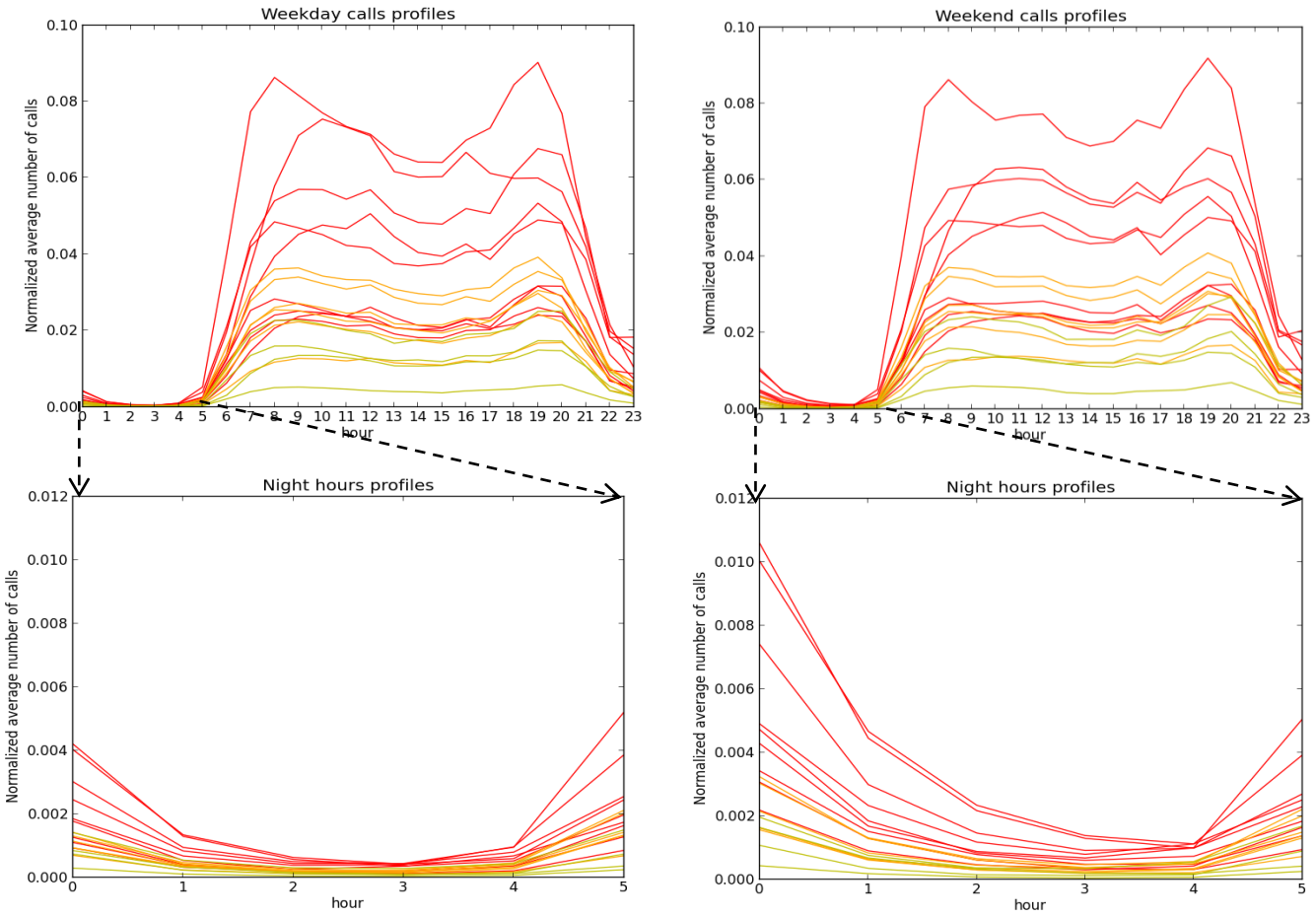
Сл. 2. Део *Python* скрипте за екстракцију обележја

Из података се могу израчунати и други параметри кретања: укупно растојање, радијус, дијаметар [5]. Поред тога естимирају се обележја везана за миграције (краткорочне и дугорочне; долазне и одлазне). За посматрани регион се прво одреди број корисника који у њему живи (сви корисници чија је најфреквентнија локација припада региону). Потом се издвајају корисници који из свог региона одлазе у друге – њихов број описује одлазне миграције. Долазне миграције квантификују се издвајањем корисника који посећују посматрани регион. Додатни временски параметар одређује миграције од интереса: задржавање дуже од 1, 5, 10 или 30 дана.

Примена

Резултат описане методе су вектори обележја за дефинисане просторне јединце - регионе. Иницијална просторна целина (држава, област, град) се подели на регионе од интереса и за сваки се из *CDR* података екстаркује вектор обележја који носи информације колико и када су људи у региону активни, на који начин се крећу, колике су долазне и одлазне миграције.

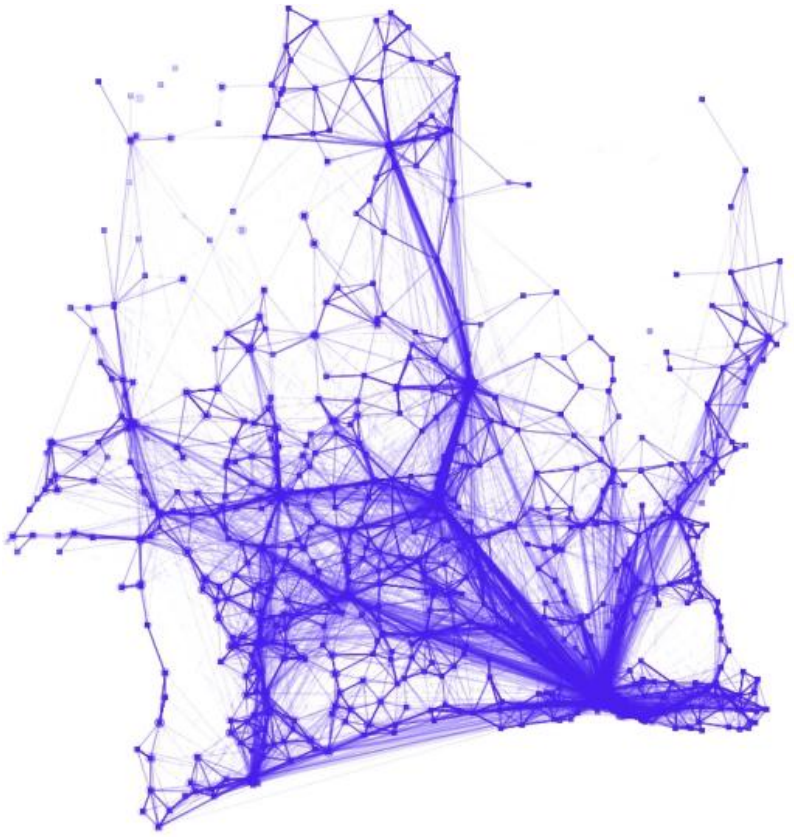
Добијени скуп вектора обележја погодан је за даљу статистичку анализу и повезивање са другим релевантним подацима (економски, саобраћајни, јавно здравље, итд.)
На слици 3. дат је пример графичког приказа 19 профила просечне активности корисника по дефинисаним регионима. Профили су нормализовани са бројем корисника по региону. За сваки регион сумирају се по сатима сви *CDR* слогови чије је настајање иницирано корисником из региона. Добијени профили се могу графички приказати. У оквиру техничког решења развијен је део за приказ профила у форми *Python* скрипте коришћењем *Matplotlib* библиотеке. Профили региона могу бити приказани различитим бојама према задатом параметру. На пример, боја може бити повезана са економском развијеношћу, саобраћајном гужвом, учесталости оболевања током одређене епидемије по регионима.



Сл. 3. Комуникациони профили по регионима

Техничко решење такође обухвата део за издвајање фреквентних трајекторија. За сваког корисника се издваја временски уређени низ локација које он посети током посматраног периода. Добијене трајекторије апроксимирају кретање корисника и оне се агрегирају да би се детектовале најфреквентније путање. Агрегација се врши алгоритмом за кластеровање који проналази сличне трајекторије и генерише репрезентативне. Помоћу софтвера *CommonGIS* могуће је графички приказати идентификаоване фреквентне трајекторије креатања. На слици 4. дат је пример графичког приказа фреквентних

трајекторија. Визуелизација омогућава јасно уочавање главних путних праваца и раскршћа. Информације о фреквентним трајекторијама посебно су значајне за саобраћај: надзор, управљање, и побољшање протока саобраћаја, отклањање саобраћајних гужви, откривање набоље путање између било које две тачке на мапи.



Сл. 3. Фреквентне трајекторије

Неке од могућности које пружа описани метод:

- Унос података у базу
- Селекција и агрегација података *SQL* упитима
- Могућност екстракција обележја о активности и мобилности за дефинисане регионе
- Приказ активности корисника, агрегираних по регионима
- Приказ фреквентних трајекторија

Наведене могућности дају велику подршку разумевању динамике активности корисника у временској и просторној димензији. Представљено техничко решење доприноси унапређењу анализе *CDR* података и добијању квалитетних информација за потребе истраживања и подршке одлучивању у системима где знање о активности и мобилности корисника има велики значај.

Техничке карактеристике

- ***MySql*** база података
- ***Python*** скрипте за екстракцију обележја
- Библиотеке:
    - **Geopy** *(Geocoding Toolbox for Python):* **https://code.google.com/p/geopy/**
    - ***Numpy*** *(Package for scientific computing in Python)* **http://www.scipy.org**
    - ***Orange*** *(Data Mining Toolbox for Python)* **http://orange.biolab.si/**
    - ***Matplotlib****(python 2D plotting library)* **http://matplotlib.org**
- ***CommonGIS*** **http://www.gisig.it/common-gis/** за визуелизацију фреквентних трајекторија

ЛИТЕРАТУРА

[1] M. González, C. Hidalgo, and A Barabási,"Understanding Individual Human Mobility Patterns", *Nature*, vol. 453, pp.779–782, 2008.

[2] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli, "Mobile landscapes: using location data from cell phones for urban analysis", *Environment and Planning B: Planning and Design*, 33(5):727, 2006.

[3] L. Bengtsson, X. Lu, A.Thorson, R. Garfield, and J. von Schreeb, "Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti", vol. 8, *PLoS medicine*, 2011.

[4] Berlingerio, Michele and Calabrese, Francesco and Di Lorenzo, Giusy and Nair, Rahul and Pinelli, Fabio and Sbodio, Marco Luca, "AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data", *Machine Learning and Knowledge Discovery in Databases*, 663—666, 2013.

[5] B. Csaji, A. Browet, V.A. Traag, J.C. Delvenne, E. Huens, V. Dooren, Z. Smoreda, and V.D. Blondel, Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, 2012.