Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 234 · 2015

Krzysztof Węcel

Uniwersytet Ekonomiczny w Poznaniu
Katedra Informatyki Ekonomicznej
krzysztof.wecel@ue.poznan.pl

LINKED GEODATA FOR PROFILING OF TELCO USERS

Summary: There is a growing interest in location-based profiling of users de-fined as combining geo-data with anonymous on-line profiles. The profile of an entity usually consists of concepts accompanied by a weight specifying a relative importance of the given concept for making an analysed entity distinct. The proposed profiling method of telco users is a two-step approach. First, profiles of mobile tower stations (BTS) are created based on crowdsourced geographical information. Second, they are used to generalise the behaviour of a calling user, which is determined from Call Detail Records (CRD). The linked data cloud is considered as an additional knowledge source in the user modelling process.

Keywords: linked data, user profiling, linked geodata, call detail record, mo-bile user, telco, cdr, bts, lgd, osm.

Introduction

There is a growing interest in location-based profiling defined as combining geo-data with anonymous on-line profiles. New methods for capturing information on where are the users and how their position changes over time are constantly developed. This information is becoming increasingly more valuable for a growing number of location-based or location-aware services. Some researchers try to estimate value of mobile data information utilising proximity-based advertising valuation (Baccelli, Bolot, 2011). Tourists have been identified as the most rewarding target group.

Call Detail Record (CDR) is the most widely used source of mobile location data in academic research (Song et al., 2010). Presented location is not very precise as it is "rounded" to the co-ordinates of the nearest base transceiver station (BTS). Various granularities of location impact the value of location infor-

mation (Baccelli, Bolot, 2011). CDR has been identified to be sufficient for drawing conclusions at the area level (Qu, Zhang, 2013), hence it is also sufficient for our purposes.

Linked data cloud is very often considered as additional knowledge source in user modelling process. Concepts defined in various ontologies can be used to characterise entities. Profile of the entity consists then of the concept accompanied by a weight specifying relative importance of the given concept for making analyse entity distinct.

In this paper we focus on cell tower granularity of location information and annotate it with geographical ontology derived from OpenStreetMap. The goal of the method is to provide profiles of the users based on profiles of the BTS stations. Therefore, the profiling process has been split into several steps. First, the information about BTS location and its neighbourhood has to be retrieved and analysed. Then, based on this information a summary of BTS profiles is prepared. We propose an improvement in profiling process by leveraging TF-IDF ranking to address the issue of uneven distribution of categories describing mobile tower locations (skewness). In the last step we can characterise users that log-in into specific BTS stations.

Section 2 presents related research. Section 3 explains our general approach to profiling of entities based on geographical context. Section 4 introduces a method for characterisation of BTS stations, from data collection, through analysis, to data aggregation. Section 5 provides a method for profiling of telco users.

1. Related research

In the literature, there are various approaches to profiling of mobile users. Some authors base purely on telco data, i.e. data available for mobile operators. Majority of methods leverage the social media where users manifest their opinions, feelings, reveal location etc. The most sophisticated approaches add mining for generalisation of patterns and classification of users. Having access to anonymised call data, we base our method on this data.

Most methods base on social data like Twitter, Foursquare, Flickr, or Instagram as data is relatively easily retrievable (API available). These services are then widely used and generate large volumes of data. One of the challenges is how to model the use of such social (and mobile) applications by various users. It is essential to understand the semantics of messages posted by them. Two trends are observed here: extraction of meaning and location.

In many studies in order to enrich and disambiguate information gathered from user, semantic technologies are considered. They are particularly useful for providing context, and geographical context is one of the most important. Abel et al. introduced a user modelling framework that utilises semantic background knowledge and use it for point of interest (POI) specification (Abel et al., 2012). Two knowledge sources of linked data are considered: GeoNames and DBpedia. They demonstrate that user modelling quality improves when LOD-based background knowledge is considered. DBpedia is unfortunately too coarse-grained for our purposes.

Instead of analysing separate check-ins, some approaches build activity-travel profile – a spatial trajectory is built from mobile phone call records only. The tricky part is in classification of trajectories, where data mining methods can be applied (Görnerup, 2012). Not only the sequences have to be derived but in order to make sense they have to be classified into typical activity-travel patterns. Their relative frequencies constitute an activity travel profile (Liu et al., 2014). Our method also bases on a number of BTSes visited. However, we do not consider sequences of visits as our experiments have shown that this would not produce meaningful results.

Trajectories, or sequences of visited locations, are not very useful unless they are confronted with activities of the users. Lie et al. investigated to what extent the behavioural routines could reveal the activities being performed at mobile phone call locations (Liu et al., 2013). The real value is in annotating locations with activity purposes but for this additional information is required. Although they have devised mathematical models to quantitatively characterize travel patterns, the motivating activities "were still in a less-explored stage" (Liu et al., 2013). Our method provides just another context for reasoning about possible activities of the users, e.g. shopping, playing tennis.

According to (Cano et al., 2013) little has been done in modelling location entities. Therefore, they proposed to profile geographical areas by providing topical categorisation. Cano et al. used additional source – linked data to interlink information from social stream with geographical objects (Cano et al., 2013). They have introduced LinkedPOI ontology, which uses DBpedia categories to profile geographic space and proposed geo-lattice Awareness Stream model as one of the ways to represent location. The process consisted of filtering, enriching, structuring and interlinking microposts from Twitter, Facebook, and TripAdvisor. Our method in fact models location entities but as we do not analyse microposts we do not have to guess the correct location – it is provided in CDR.

Qu proposed a framework and corresponding analytic methods to use User Generated Mobile Location Data (UGMLD) for Trade Area Analysis (Qu, Zhang, 2013). They have defined three key processes: "identifying the activity centre of a mobile user, profiling users based on their location history, and modelling users' preference probability." Application of the method is meant for analysis of customers' visits to business venues. However, they rejected CDR as the data source in their research as it was too coarse. Our method is able to utilise CDR in order to provide profiles of certain locations although we cannot provide profiles of certain venues belonging to bigger chains.

When specific locations are considered, Chen et al. also presented a method for profiling businesses at specific locations that was based on mining information from social media (Chen et al., 2014). They matched geo-tagged tweets against locations from Foursquare to build a profile of mentioned businesses.

Going back to the user, Ostuni et al. presented Cinemappy – a location-based application that computes film recommendations by exploiting contextual information related to current location of the user, leveraging information from DBpedia (Ostuni et al., 2013). Similarly, DBpedia was used by (Cano et al., 2013) who proposed a semantic travel mash-up as possible application. Approach for museums was presented in (Ruotsalo et al., 2013).

One of the obstacles by user modelling is uneven distribution of categories of business, e.g. there are much more visits to a cinema than to second-hand and there are much more shops than theatres. It was first noted by Qu who explained this by social motivation, not necessarily by the differences in the number of various categories (Qu, Zhang, 2013). In their work the categories were very fine grained, for example Foursquare has a hierarchical category structure with 9 top categories and ca. 400 sub-categories. For the clarity of interpretation the categories have been later collapsed to 6 groups. Our methods addresses this issue by applying specific methods from information retrieval domain.

2. Approach to geographical linked data-based profiling

This section describes user profiling based on BTS characteristics derived from the geographical linked data. We follow the idea of location-based user profiling – one of the approaches is geoprofiling, a commonly used method to approximate user characteristics based on neighbourhood demographic data.

In most approaches, there is a venue or a place given and authors are looking for coordinates. This is particularly important when text is analysed, especially in social media, e.g. Twitter. In order to simplify disambiguation, some

portals allow the so called check-ins where users select precise location, e.g. Foursquare, Facebook. We take the reverse process: starting from coordinates we are interested in the objects nearby, thus describing the geographical context, further referred to as location profile.

Our experiments have been carried out on anonymised data, where the only reasonable data for linking was location of BTS towers. There was just one type of information that was widely used and could supplement our records – geographical information. There are several open data sources concerning geographical information that we could use, DBpedia and OpenStreetMap being the most prominent. Taking into account the granularity of available data and the requirement to display results on the map we made a decision to base our method on OpenStreetMap and its triplified counterpart – LinkedGeoData (Auer, Lehmann, Hellmann, 2009). As a crowdsourced data, it is kept relatively up to date and but without breaking fluctuations.

LinkedGeoData (LGD) provides an ontology for classification of locations. There are ca. 1200 categories grouped into ca. 45 top-level categories. Comparing LGD to Foursquare, the latter has ten top-level, 436 second-level, and 266 third-level categories[1]. Foursquare's maps in fact use OpenStreetMap[2]. In our approach more general categories are advantageous as they can make interpretation of generalisation results easier. Sub-categories could be more valuable as they can distinguish users better but the solution is then less stable from the statistical point of view. This is a well-known trade-off of specificity vs. sensitivity (Fawcett, 2006). In order to best characterise the BTS stations, we have restricted our further analysis to some predefined objects (see Fig. 3).

The reasoning behind our profiling approach is presented in the following user story. A user often visits sport amenities. They are in the scope of some BTS stations. Profiles of such stations contain sport amenities with higher frequency than an average station. This is further reflected in a user profile where sport amenities gain higher weight when user trajectory is aggregated. Some visited venues are additionally annotated with a kind of sport, for example tennis[3]. Looking into calendars of sport events we can even reason further in which kind of sport user is interested or whom the user is supporting (whether team or individual).

[1] https://developer.foursquare.com/categorytree.
[2] https://foursquare.com/about/osm.
[3] This depends on data availability in OpenStreetMap.

3. Characteristics of BTS

3.1. Retrieval

In our experiments we have used BTS towers located in Poland, with ca. 8000 unique locations, stored in MySQL. At the beginning the information about BTS locations has been retrieved. Using a Python script, for each location a SPARQL query was prepared to retrieve list of objects in the neighbourhood along with their categories. As a source of data for our queries we have used LinkedGeoData which is a derivative of OpenStreetMap. Two main categories of objects are distinguished therein: nodes (just a point according to GIS terminology) and ways (lines or polygons). Separate queries for nodes and ways had to be prepared because the Virtuoso's built-in distance function has different behaviour for nodes and ways. As there were ca. 8000 locations, two kinds of objects, two means for object capturing (bounding box and circle) and 3 various distances, we had to post ca. 80 thousand queries.

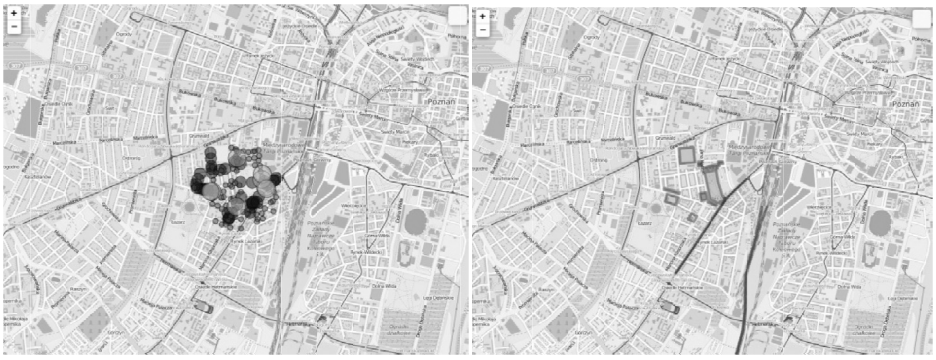
Below sample SPARQL query is presented:

```
PREFIX lgdm:<http://linkedgeodata.org/meta/>
PREFIX geom:<http://geovocab.org/geometry#>
PREFIX ogc:<http://www.opengis.net/ont/geosparql#>
SELECT distinct ?class ?way
WHERE { ?way a lgdm:Way .
?way a ?class .
?way geom:geometry [ogc:asWKT ?geo ] .
filter(bif:st_within( ?geo,bif:st_point(%f,%f),%f ))
}
```

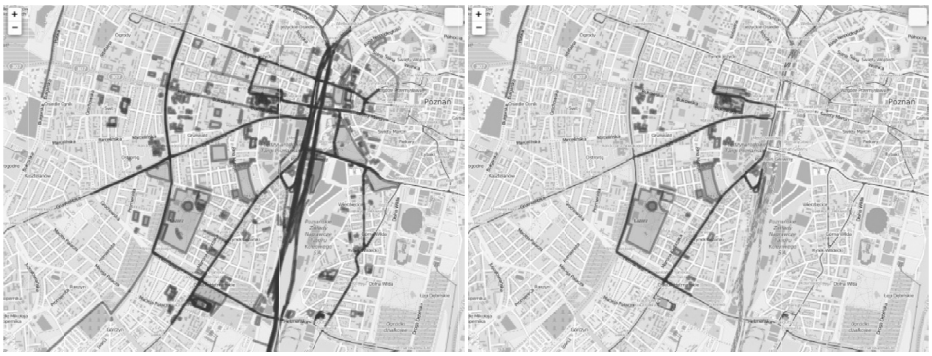
Listing 1. Sample SPARQL query using geospatial functions

We have decided to use LGD's endpoint instead of OSM's API as it was possible to use Virtuoso built-in SPARQL functions for spatial queries. For the retrieval of nodes, it was possible to provide detailed query and the radius was in fact expressed in kilometres. For example, Figure 1a presents various venues located in a circle of 1 km diameter.

Retrieval of ways was more complicated. Function `bif:st_within` was not returning what it was expected for Way objects when the other parameter was a point. Several methods to get satisfactory results have been tested, including generation of boxes to simulate containment (see Fig. 1b and 1c). Finally, the circle overlap after toleration parameter tuning to 0.01 has been chosen as a method to query for neighbouring ways (Fig. 1d).



(a) nodes, circle, 1km diameter (b) ways, box, 1km



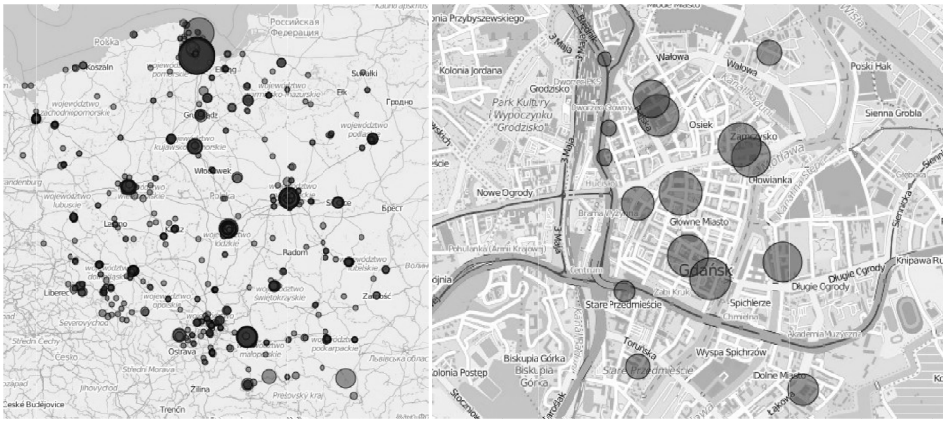
(c) ways, box, 5km (d) ways, circle, distance 0.01

Fig. 1. Various venues selected based on object type and distance, near Poznań International Fair

3.2. Analysis

Various aspects of location characteristics can be analysed, both qualitative and quantitative. For example Fig. 2a shows BTS stations that have hotels in their neighbourhood. The size of the circle specifies the number of hotels. It is interesting to observe that the city most packed with hotels in Poland is Gdansk. The Fig. 2b shows a closer look.

Another aspect analysed is number of venues of given category per BTS location. General findings are as follows: on average there are 4.0 shops per location and 1.2 restaurants; on the other end are universities (0.04) and cinemas (0.06). Such asymmetry needs to be addressed when building the location and user profiles.



(a) Poland (b) Gdańsk

Fig. 2. Number of hotels located within BTS stations

Although nodes and ways were distinguished due to technical limitations, results provided by these two types of objects have to be merged to provide a comprehensive location profile. There is a strong preference between category and GIS type and the OpenStreetMap community has adapted certain patterns. Objects like parking or leisure areas are better represented by showing the area, hence they are more popular as ways. For example, parking is represented as a way in almost 110 thousands cases; only 3.5 thousands parking areas are marked as a node. On the other end are ATMs – over 6500 entities are represented as nodes and only 4 as a way. The same applies for tram stops. Shops are the most balanced entity: 24899 ways vs. 28871 nodes.

After experiments we concluded that it is useful to prepare characteristics of a typical BTS location, including input from both nodes and ways. Combined profile is presented in the Fig. 3. On average, 22.1% of objects within BTS are parkings and 15.3% leisure-related.

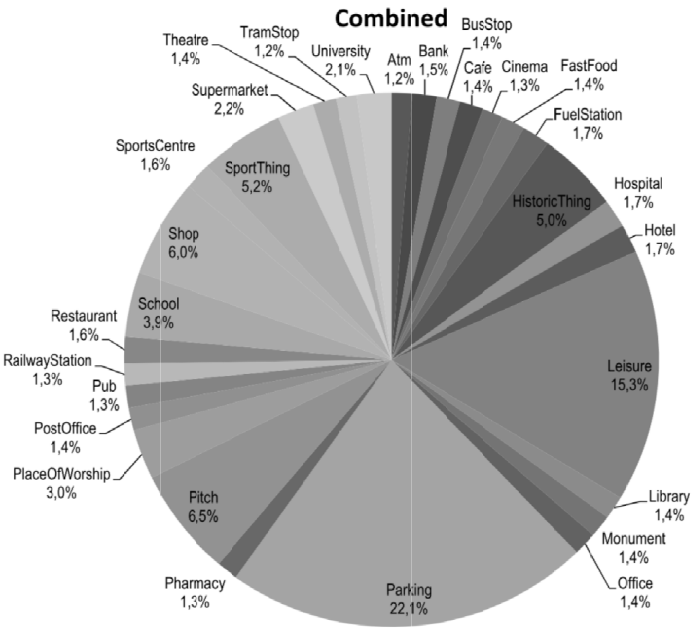


Fig. 3. Average distribution of objects among predefined 30 categories of objects

4. Profiles of locations and users

4.1. TF-IDF-inspired method for location profiling

Simple aggregation of geographical categories assigned to places is not sufficient to correctly profile locations. Relative values are more important than absolute values. For example, as there are much more shops than libraries the results would be biased if we had not included this correction in the characteristics of locations. In fact, we are interested in information, if given users visit some kind of objects more often than an average user.

In order to alleviate the effect of uneven distribution of categories we propose to use TF-IDF weighting schema known from information retrieval. TF-IDF is actually a product of two statistics: term frequency (TF) and inverse document frequency (IDF).

TF can be calculated in different ways, e.g. boolean or raw frequency. We have decided to use relative frequency (which is normalised to 1.0), expressed as follows:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

where $n_{i,j}$ is number of times that the term t_i occurred in document d_j.

IDF is a measure of term specificity – less frequent terms have bigger discrimination power as they can better characterise a document. It is calculated as follows:

$$IDF_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

where $|D|$ is a number of documents in corpus, and denominator contains number of documents containing the term t_i.

An entry to IDF calculation is presented in Fig. 4. TF-IDF factor in the case of the most popular category – Shop – is 0.766 and for least popular category – Pitch – is 4.234. Conclusion: shops, present in almost half of the locations, are not very useful to distinguish locations. Sample results obtained from the above TF-IDF method are given in the Table 1. They have been built based on the restricted list of 30 objects.

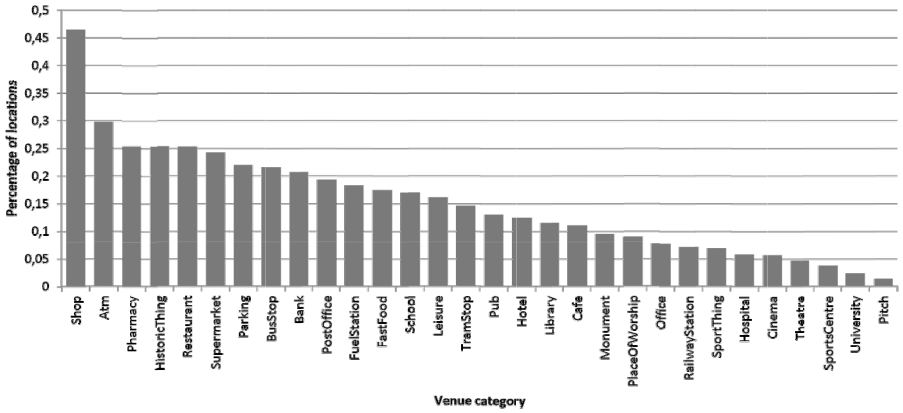


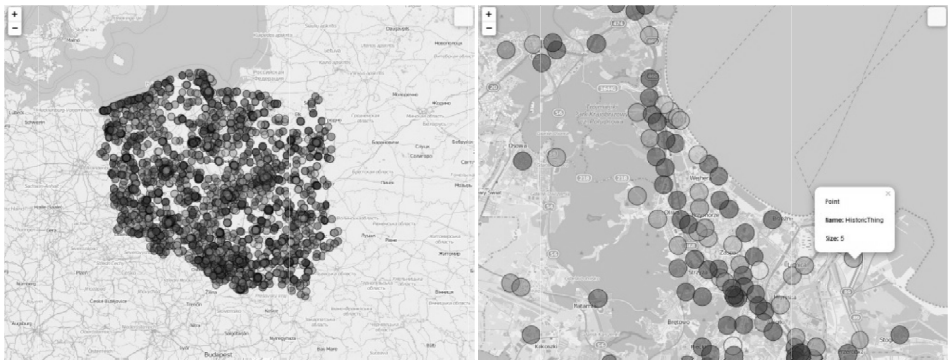
Fig. 4. Percentage of locations containing venues (nodes) of given category

Table 1. Sample profiles of BTS locations calculated with TF-IDF

loc_id	class	count	tf	idf	tf-idf
32	Leisure	1	0,25	1,823	**0,456**
32	Restaurant	1	0,25	1,374	0,343
32	HistoricThing	1	0,25	1,371	0,343
32	Shop	1	0,25	0,766	0,192
34	BusStop	2	0,50	1,533	**0,766**
34	PlaceOfWorship	1	0,25	2,408	0,602
34	FuelStation	1	0,25	1,696	0,424
36	School	1	1,00	1,771	**1,771**
37	FuelStation	1	0,50	1,696	**0,848**
37	HistoricThing	1	0,50	1,371	0,686
41	Bank	8	0,16	1,574	**0,257**
41	Monument	5	0,10	2,351	0,240
41	Shop	13	0,27	0,766	0,203
41	HistoricThing	5	0,10	1,371	0,140
41	Hospital	2	0,04	2,839	0,116
41	University	1	0,02	3,721	0,076

They profiles in the Table 1 should be interpreted as follows: in the neighbourhood of the location id 34 there are 4 geographical objects, all of them are nodes (points), including 2 bus stops, 1 place of worship, and 1 fuel station. Some locations have just one object (e.g. id 36), other have many more (e.g. id 41). In bold we have marked the category with the highest measure, thus being the most characteristic for a given location. When many categories are equally frequent (location id 32), the one with the highest IDF makes top of the ranking. It is interesting to observe that 13 shops in location id 41 are less important than 5 monuments and 8 banks. Weights can be compared between locations, but then those categories with smaller number of categories get higher weights (case id 36). This is to some extent justified – such categories clearly profile the location.

Fig. 5 presents a visualisation of top classes ranked according to TF-IDF measure. Each BTS location is annotated with the top-ranked category in the neighbourhood. The most important objects are coded as follows: red – shops, blue – historical objects, orange – schools, green – public transport. Figures for the whole Poland (Fig. 5a) and Gdańsk (Fig. 5b) are attached.



(a) Poland (b) Gdańsk

Fig. 5. Most popular annotations of BTS locations

4.2. User profiling

For evaluation, we have used the database with a sampled 10.000 users containing data for 3 months. From this database we have queried users with at least 10 BTS locations and then randomly selected users[4]. The calculation of the char-

[4] Please note that this is illustration of the method and some decisions are arbitrary. Any other period can be considered, any other set of object types can be used.

acteristics of users is rather straightforward. The profile of a user is prepared as a weighted sum of profiles of visited BTS locations, where the weight is the number of connections initiated with a given BTS. Similarly to locations, we can visualise profiles of users as a pie chart. Fig. 6 presents profiles for sample users.

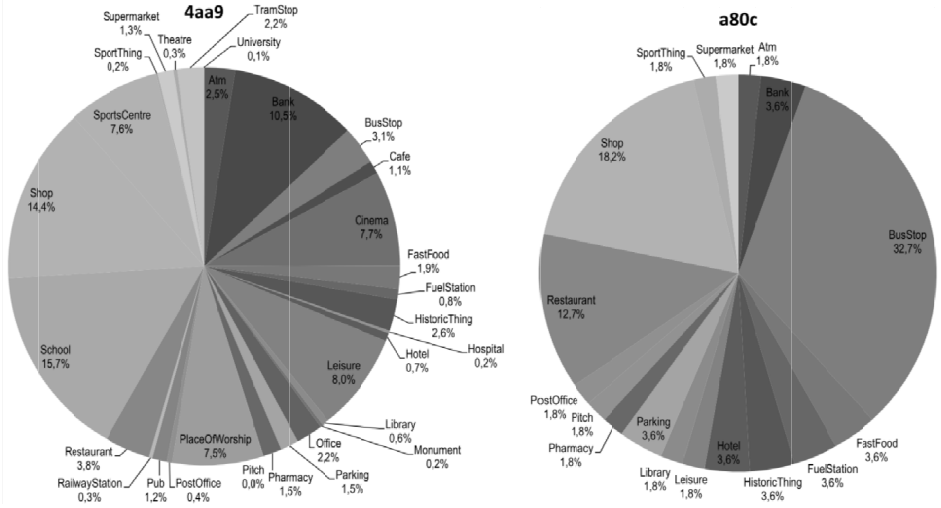


Fig. 6. Geographical Linked Data-Based profile of the sample users

For comparison of user profiles much more suitable is a radar chart that can better emphasize differences and similarities between users. One of the problems with this kind of charts is that axes should have the same measure. Therefore, we have normalised the values in charts in such a way, that user with the highest value of given object type in profile has value 1.0. The same users are compared in Fig. 7. Grouping of certain categories or even reducing number of categories should further improve the readability of the chart.

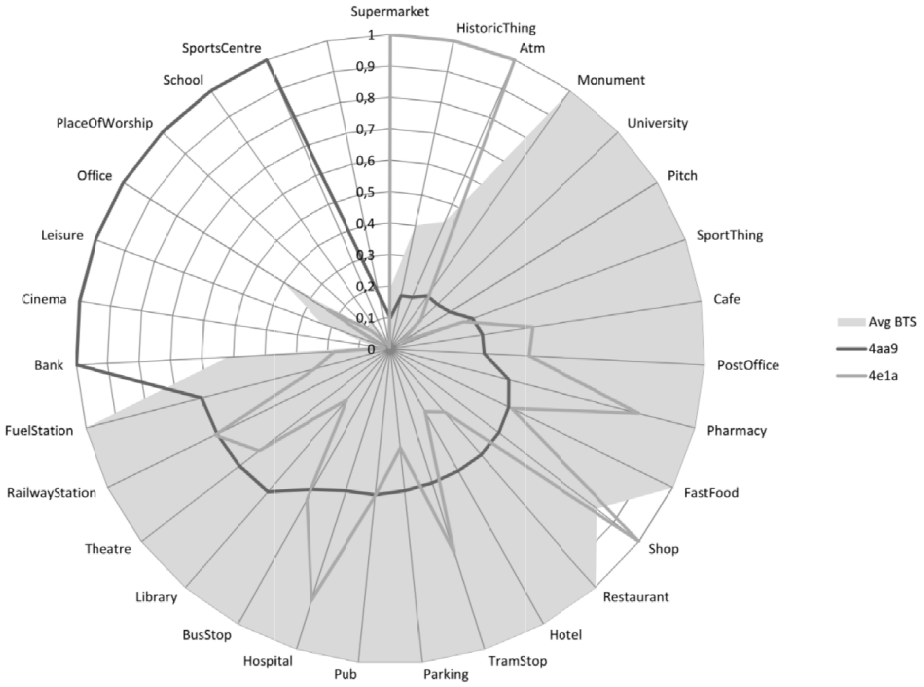


Fig. 7. Comparison of user profiles on radar charts – restricted to two users

Conclusions and future work

In this paper we have contributed the method for exploitation of LinkedGeoData and also the TF-IDF-based method for profiling of locations and users. The method has been applied in the particular settings of mobile telco operators. Nevertheless, the elaborated method can be used universally for profiling any business requiring profile of neighbourhood, e.g. for marketing purposes or revenue estimation.

So far we have applied a simple approach for aggregation to obtain user profile. In the future we plan to use Latent Dirichlet Allocation (LDA) (Blei, Ng, Jordan, 2012). It is a generative probabilistic model where each item of a collection is modelled as a finite mixture over an underlying set of topics. This resembles our approach for profiling: users are modelled as a mixture of categories that are provided by locations. What we need to determine are weights allowing to prepare appropriate mixture. The method can also be improved by considering the hierarchy of categories. As of now LinkedGeoData contains all intermediate categories when annotating specific venue (e.g. both shop and amenity). There is an extension of the method to hierarchical LDA (hLDA).

References

Abel F., Hauff C., Houben G.-J., Tao K. (2012), *Leveraging User Modeling on the Social Web with Linked Sata* [in:] *Web Engineering*, Springer-Verlag, Berlin-Heidelberg, pp. 378-385.

Auer S., Lehmann J., Hellmann S. (2009), *LinkedGeoData: Adding a Spatial Dimension to the Web of Data*, ISWC 2009, Vol. 5823, Springer, Heidelberg, pp. 731-746.

Baccelli F., Bolot J. (2011), *Modeling the Economic Value of the Location Data of Mobile Users*, INFOCOM, IEEE, pp. 1467-1475.

Blei D.M., Ng A.Y., Jordan M.I. (2012), *Latent Dirichlet Allocation*, "Journal of Machine Learning Research", Vol. 3(4-5), pp. 993-1022, doi:10.1162/jmlr.2003.3.4-5.993.

Cano A.E., Dadzie A.-S., Burel G., Ciravegna F. (2013), *Topica-Profiling Locations through Social Streams*. *Semantic Technology*, Springer-Verlag, Berlin-Heidelberg, pp. 290-305.

Chen F., Joshi D., Miura Y., Ohkuma T. (2014), *Social Media-based Profiling of Business Locations*, Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia, Orlando, FL, pp. 1-6.

Fawcett T. (2006), *An Introduction to ROC Analysis*, "Pattern Recognition Letters", Vol. 27(8), pp. 861-874, doi:10.1016/j.patrec.2005.10.010.

Görnerup O. (2012), *Scalable Mining of Common Routes in Mobile Communication Network Traffic Data* [in:] J. Kay, P. Lukowicz, H. Tokuda, P. Olivier, A. Krüger (eds.), "Pervasive Computing", Vol. 7319, Springer-Verlag London, pp. 99-106, doi:10.1007/978-3-642-31205-2_7.

Liu F., Janssens D., Cui J., Wang Y., Wets G., Cools M. (2014), *Building a Validation Measure for Activity-based Transportation Models Based on Mobile Phone Data*, "Expert Systems with Applications", Vol. 41(14), pp. 6174-6189, doi: 10.1016/ j.eswa.2014.03.054.

Liu F., Janssens D., Wets G., Cools M. (2013), *Annotating Mobile Phone Location Data with Activity Purposes Using Machine Learning Algorithms*, "Expert Systems with Applications", Vol. 40(8), pp. 3299-3311. doi:10.1016/j.eswa.2012.12.100.

Ostuni V.C., Gentile G., Di Noia T., Mirizzi R., Romito D., Di Sciascio E. (2013), *Mobile Movie Recommendations with Linked Data* [in:] *Availability, Reliability, and Security in Information Systems and HCI*, Springer, Berlin-Heidelberg, pp. 400-415.

Qu Y., Zhang J. (2013), *Trade Area Analysis Using User Generated Mobile Location Data*, Proceedings of the 22nd International Conference on World Wide Web, Republic and Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee, pp. 1053-1064, http://dl.acm.org/citation.cfm?id= 2488388.2488480 (accessed: 30.08.2015).

Ruotsalo T., Haav K., Stoyanov A., Roche S., Fani E., Deliai R., Mäkelä E., Kauppinen T., Hyvönen E. (2013), *SMARTMUSEUM: A Mobile Recommender System for the Web of Data*. "Web Semantics: Science, Services and Agents on the World Wide Web", Vol. 20(0), pp. 50-67, doi:10.1016/j.websem.2013.03.001.

Song C., Qu Z., Blumm N., Barabási A.-L. (2010), *Limits of Predictability in Human Mobility*, "Science", Vol. 327(5968), pp. 1018-1021.

POWIĄZANE GEODANE DLA PROFILOWANIA UŻYTKOWNIKÓW TELCO

Streszczenie: Obserwuje się rosnące zainteresowanie geograficznym profilowaniem użytkowników, rozumianym jako łączenie danych geograficznych z anonimowymi profilami użytkowników. Profil jednostki zazwyczaj składa się z pojęć geograficznych oznaczonych wagami, odzwierciedlającymi względną ważność poszczególnych pojęć dla odróżniania użytkowników. Proponowana metoda profilowania użytkowników sieci komórkowych jest dwuetapowa. W pierwszej kolejności tworzone są profile stacji przekaźnikowych (BTS) na podstawie społecznie dostarczonych informacji geograficznych. Następnie te profile są wykorzystywane do uogólnienia zachowania użytkownika, wynikającego z analizy logów jego połączeń (CDR). Chmura danych powiązanych (linked data) jest wykorzystywana jako dodatkowe źródło wiedzy w procesie modelowania użytkownika.

Słowa kluczowe: dane powiązane, profilowanie użytkownika, powiązane geodane, logi połączeń, użytkownik mobilny, telco, cdr, bts, lgd, osm.