Секція 7. Бази даних і знань та побудова інтелектуальних систем на їх основі

UDC 004.6

ON THE PROBLEM OF MINING SEMANTICS OF USER BEHAVIOR IN LOCATION-BASED SOCIAL NETWORKS.

Vedernikov O.O.
National Chiao Tung University, student

Introduction and motivation

According to increasing number of usage GPS devices by people, problem of mining trajectories of users appears consequently. The number of popular websites Thus, finding communities include people who apply the same routes at same cities, but not similar sensible behavior. Contemporary location-based social networks also provide the ability of assignment specific semantic tags to geographical locations, so this goal can be achieved.

Search of groups of people with the same behavior can be used, for example, in different applications of advertisement targeting. For example, these problems may be of interest to institutions that are interested in what other establishment also attended by their customers. Users also can be informed about goings-on of another people with resembling predilections. Sociologists can explore the hidden persuasive of users with various distinctive features (f.e. gender, nationality) through the establishment of relationships between the locations they tend to visit. User communities can be also used in friends suggestion systems, if the web-service includes opportunity of friendship between users.

Having initial dataset C = {<u, p, t>}, where u is element of set of users, p is element of set of location tags, t is time-stamp, some places t_i may be annotated with the set of semantic tags S_i. Thus, we can examine the data as set of pairs of <u,s>, count how many times current user has visited places with this semantic tag in given time interval and to build a matrix of users' preferences.

Related works

Existing works on finding communities are related to pure GPS data and deal with raw coordinates. [1] A significant number of works consider mining hot regions of users' trajectories and use grid-based approaches or density-based approaches of areas.[2]

There is less number of works dedicated to research of check-in data with semantic tags. Those include annotation of missing tag places [3], finding time similarity between tags [4], analysis of tags in OpenStreetMap system [5], finding significant semantic locations from GPS data [6] but there are no works which are related to finding similarity between users through semantic and check-in time similarity of tags.

Proposed solution

The following steps are included into proposed algorithm:

- Finding semantic similarity between tags.
- Finding check-in time similarity between tags.
- Finding importance of semantic tags for every user.
- Making Semantic Relation Graph (SRG) of users' tag preferences, where nodes describe importance of each tag to user, and edges describe transition of users.
- Computing user's similarity as graph's similarity.
- Finding communities.

Check-in semantic analysis is based on Singular Value Decomposition of User-Tag matrix, where each column is a tag, each row represents user, and the value of their intersection describes how many times user checked-in in locations with given tag. Proposed solution in thesis applies operation of Singular Value Decomposition to the stated problem. Operation of singular value decomposition is very useful in area of describing a matrix. It reduces the effects of similar tags, merging the dimensions associated with the tags that have similar behavior. If some place is visited by a relatively large number of every people, it will have less effect on resulting matrix.

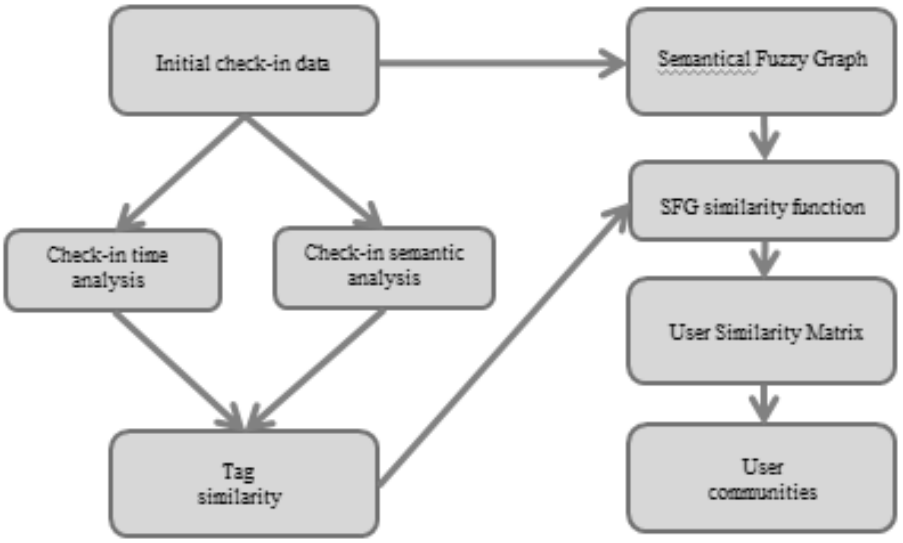

Figure 1 - Workflow of process

The SVD is a very powerful and useful matrix decomposition, particularly in the context of text analysis, dimension reducing transformations of images, earning human-like knowledge, image compression, reconstruction of gene regulatory networks etc. It is based on the fact that any real $m \times n$ matrix A can be spreaded as $\mathbf{A} = \mathbf{S}\Sigma\mathbf{C}^T$, where S is $m \times m$, C is $n \times n$, and Σ is $m \times n$ and has only singular values on its diagonal in decreasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_l \geq 0$, where $l = \min\{m, n\}$. Because the diagonal elements $\Lambda_{kk} = 0 \quad \text{for } k = n+1, \ \ldots, m$, the eigenvectors (singular vectors) $\mathbf{s}_{n+1}, \ \cdots, \mathbf{s}_m$ are of no importance. As a result we define a new $m \times n$ matrix $\hat{\mathbf{S}}$ (it is $\mathbf{S}$ with the last $m-n$ columns deleted) and a new $n \times n$ diagonal matrix $\hat{\Sigma}$ (whose diagonal elements are $\sigma_1, \ \ldots, \sigma_n$) and write the thin SVD (or reduced SVD) of A as $\mathbf{A} = \hat{\mathbf{S}}\hat{\Sigma}\mathbf{C}^T$. After this operation, the correlation between columns of A will be more representative [7], and we can state is as $TagSemSim(i,j)$.

Finding check-in time similarity between tags is described in [4], so we can measure difference between tags as difference between their probability distribution:

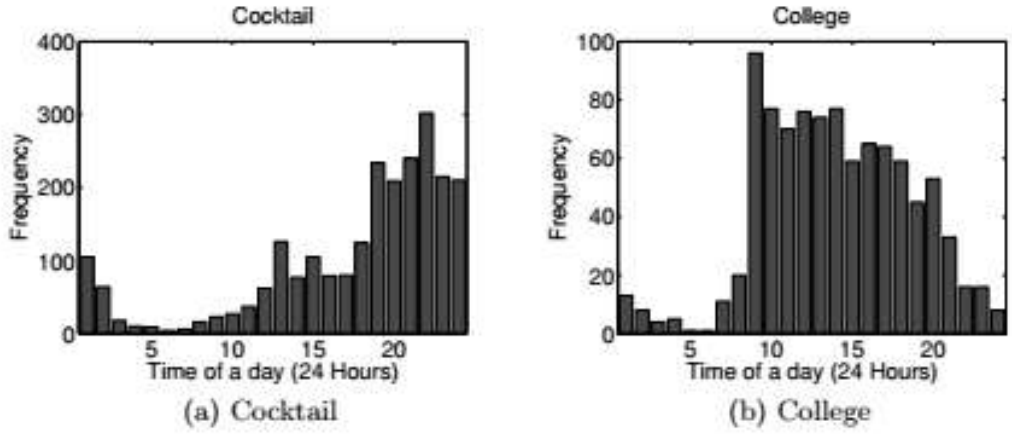

Figure 2 - Two tags time-distribution

Compute time similarity as similarity between probability distributions

$$TagTimeSim(i,j) = 1 - \frac{1}{2}\sum_{k=1}^{24}(p_{i,k} - p_{j,k})$$

$$TagSim(i,j) = \rho TagSemSim(i,j) + (1-\rho)TagTimeSim(i.j)$$

Weight ρ we have to find empirically, and proposed value is 0.35.

User's SRG has to represent user's behavior, so it has to include

- Weight of node = importance of tag for given user (normalized by TF-IDF alue)
- Weight of edge = normalized transition between tags check-in time. Weight function:

$$tdc(t) = \begin{cases} 1, t \le t_{min} \\ \dfrac{t_{max} - t}{t_{max} - t_{min}}, t_{min} < t \le t_{max} \\ 0, t > t_{max} \end{cases}$$

For given Whrrl dataset, possible values of t_{min}=1 hour, t_{max}=8 hours.

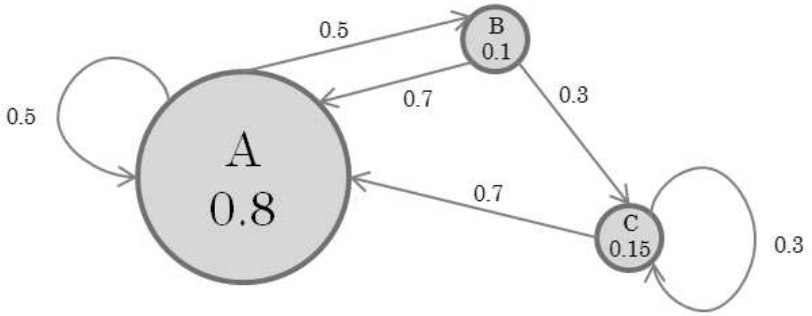

Figure 3 - Semantic Relation Graph

Measuring of user's similarity has to be represented as users' SRG similarity. Proposed approach includes not only equality of tags, but also their similarity and how transition probability between tags is equal. $SFGSim(P,Q) = \lambda NodeSim(P,Q) + (1-\lambda)EdgeSim(P,Q)$ (after experiments $\lambda = 0.6$ is most justified value), where $NodeSim(P,Q) = \|V(P_{nodes}) - V(Q_{nodes})\|$, *where vector* $V(G) = \sum_{i=1}^{|G|} Ts(i) * P(i)$, where $TS(i) - similarity\ vector\ of\ tag\ i$, $P(i)$ $- weight\ of\ tag\ i\ in\ current\ SFG$, $G = P_{nodes} \cup Q_{nodes}$, and $EdgeSim(P,Q)$ is calculated in the same way.

Thus, finding all the user similarity measure, we can find communities using any fuzzy clustering algorithm (f.e., c-means). Experiments showed that most distinguish tags are not necessarily often (such as "restaurant" or "shopping"), but that one who separate users into disjoint groups. For example, maximum feature importance have such tags as "church" and "night club", because some people appear often, some do not at all. Proposed experiments stated that this approach can be used in different fields – analysis of social groups behavior, user recommendation etc.

References

1. Ling-Yin Wei, Yu Zheng, and Wen-Chih Peng. Constructing Popular Routes from Uncertain Trajectories. KDD 2012
2. Wen-Yuan Zhu, Wen-Chih Peng, Ling-Jyh Chen. Mining Movement-based Communities from Trajectories. 2013, in advance
3. Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin and Krzysztof Janowicz. On the Semantic Annotation of Places in Location-Based Social Networks. KDD'11, USA
4. Mao Ye, Krzysztof Janowicz, Christoph Mülligann, Wang-Chien Lee: What you are is when you are: the temporal dimension of feature types in location-based social networks. GIS 2011: 102-111
5. Mülligann, C., Janowicz, K., Ye, M., and Lee, W.-C. (2011): Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. COSIT 2011, Springer LNCS 6899, pages 350-370, 2011.
6. Xin Cao, Gao Cong, Christian S. Jensen: Mining Significant Semantic Locations From GPS Data. PVLDB 3(1): 1009-1020 (2010).
7. Hansen, P. C. (1987). "The truncated SVD as a method for regularization". BIT 27: 534–553