ECAI 2010
H. Coelho et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-606-5-1107

Time-Slice Density Estimation for Semantic-Based Tourist Destination Suggestion

Michelangelo Ceci and **Annalisa Appice** and **Donato Malerba** [1]

1 INTRODUCTION

In recent years, a growing interest has been given to trajectory data mining applications that permit to support mobility prediction with the aim of anticipating or pre-fetching possible services [3]. Proposed approaches typically consider only spatio-temporal information provided by collected trajectories. However, in some scenarios, such as that of tourist supporting, semantic information which express needs and interest of the user (tourist) should be taken into account. This semantic information can be extracted from textual documents already consulted by the tourists. In this paper, we present the application of the time-slice density estimation [1] that permits to suggest/predict the next destination of the tourist. In particular, time-slice density estimation permits to measure the rate of change of tourist's interests at a given geographical position over a user-defined time horizon. Tourist interests depend both on the geographical position of the tourist with respect to a reference system and on semantic information provided by geo-referenced documents associated to the visited sites. Our basic assumption is that a tourist moves towards a close destination which is semantically consistent with her/his current profile as much as possible. Destinations that minimize the profile drift are suggested/predicted as next destinations.

2 ITiS

The applicative task we address is that of suggesting/predicting the next destination of a tourist which moves on a map given: *i*) a spatial referencing system; *ii*) the set of destinations which geo-references a set of textual documents; *iii*) the current geographical tourist position; *iv*) the trajectory followed by the tourist and its associated profile. ITiS addresses this task by automatically updating the profile each time the tourist visits a new site on the map. The profile takes into account the following assumptions: *i*) a tourist consults a document if the document content is interesting for him/her; *ii*) content of documents recently consulted is more interesting for the tourist than that of documents consulted in the past.

Preliminary Concepts. Let $P = \{p_i = \langle x_{p_i}, y_{p_i}\rangle | i = 1...n\}$ be the set of candidate destinations on a map towards a tourist can move, such that x_{p_i} and y_{p_i} represent the spatial coordinates of p_i, and n is the cardinality of P. Let $D = \{d_j | j = 1...N\}$ be a set of textual documents. One or more documents in D are geo-referenced to a destination p_i according to the function $\delta : P \rightarrow 2^D$ such that $\delta(p_i) = \{d_j \in D | d_j$ is geo-referenced to $p_i\}$. The same textual document can be geo-referenced to one or more destinations.

Given U be the set of tourists, it is also possible to define the set of visits of the tourist $u_j \in U$, that is, the movement history of the tourist, as: $v_{u_j}(t) = \big(\langle p_{j_1}, t'_{j_1}, t''_{j_1}, D_{j_1}\rangle ... \langle p_{j_s}, t'_{j_s}, t''_{j_s}, D_{j_s}\rangle\big)$ where $p_{j_k} \in P$ represents the k-th ($k = 1...s$) destination the tourist u_j visited, $D_{j_k} \subseteq \delta(p_{j_k})$ represents the set of consulted documents geo-referenced to p_{j_k} and $[t'_{j_k}, t''_{j_k}]$ represents the time interval (starting time and ending time) of the k-th visit such that $t'_{j_k} \leq t''_{j_k}$ and $t''_{j_k} \leq t'_{j_{k+1}}$ iff $k \leq s-1$; $t''_{j_k} \leq t$ iff $k = s$.

The set of consulted documents at the time t by the tourist u_j is defined as: $d_{consulted}(v_{u_j}(t)) = \bigcup_{k=1...s} D_{j_k}$. Analogously, the set of documents which are still not consulted at the time t is defined as: $d_{notConsulted}(v_{u_j}(t)) = D - d_{consulted}(v_{u_j}(t))$.

The set of visited destinations at the time t by the tourist u_j is defined as: $p_{visited}(v_{u_j}(t)) = \bigcup_{k=1...s}\{p_{j_k}\}$. Finally, the set of destinations which are still not visited at the time t is defined as: $p_{notVisited}(v_{u_j}(t)) = P - p_{visited}(v_{u_j}(t))$.

Document Representation. A document is pre-processed in order to remove *stopwords* and determine equivalent stems (*stemming*) by means of Porter's algorithm for English texts. Pre-processed documents are represented by means of a feature set which is determined on the basis of some statistics whose formalization is reported below.

Let C be a set of documents, with $C \subseteq D$, and w be a token of a stemmed (non-stop) word which occurs in a document of D, it is possible to define: $TF_d(w)$ as the *relative* frequency of w in a document $d \in D$; $TF_C(w) = max_{d \in C} TF_d(w)$ the maximum value of $TF_d(w)$ on all documents $d \in C$; $DF_C(w) = \frac{|\{d \in C | \quad w \ occurs \ in \ d\}|}{|C|}$ the percentage of documents in C in which w occurs; $CF_{C',C'',...C^{(s)}}(w)$ is the number of sets of documents where the token w occurs. In this formulation, sets of documents are denoted as $C', C'', \ldots C^{(s)}$ with $C^{(i)} \subseteq D$.

Then the following measure, used in text categorization [2], permits to associate a token w_i with its score v_i and to select relevant tokens for the representation of documents in D:

$$v_i = \frac{TF_{d_{consulted}(v_{u_j}(t))}(w_i) \times \big(DF_{d_{consulted}(v_{u_j}(t))}(w_i)\big)^2}{CF_{d_{consulted}(v_{u_j}(t)), d_{notConsulted}(v_{u_j}(t))}(w_i)}$$

Tokens that minimize v_i are penalized since they are commonly used in documents of both $d_{consulted}(v_{u_j}(t))$ and $d_{notConsulted}(v_{u_j}(t))$ and do not permit to discriminate between the two sets. Differently, tokens that maximize v_i can be used to represent documents in D. In particular, the set of the best n_{dict} tokens forms the *dictionary*, denoted as $Dict(v_{u_j}(t))$, of the tourist u_j at the time t. $Dict(v_{u_j}(t))$ is used to index the set of documents in D according to the classical normalized $TF \times idf$ measure. In the matrix representation: $\omega(v_{u_j}(t)) = [\omega_{p,q}]_{p=1..N, q=1..n_{dict}}$, where $\omega_{p,q} = \frac{TF_{d_p}(w_q) \times ln \, 1+N \times DF_D(w_q)^N}{\|\omega(v_{u_j}(t))\|_1}$ and $d_p \in D$ and $w_q \in Dict(v_{u_j}(t))$.

Time-Slice Density Based Profile. The profile of the tourist u_j

[1] University of Bari A. Moro, Italy, email: {ceci,appice,malerba}@di.uniba.it

at the time t is defined as the triple $\langle x_{u_j}(t), y_{u_j}(t), X(v_{u_j}(t))\rangle$, where $(x_{u_j}(t), y_{u_j}(t))$ is the geographical position of the tourist, $X(v_{u_j}(t))$ is the semantic position of the tourist over the space $[0,1]^{n_{dict}}$. Since it would be computationally impractical to represent and search this continuous space, ITiS uses a discrete version of the same space. The discrete space is defined by resorting to the discretization function $\psi : [0,1] \rightarrow \Phi$. Φ is a finite set of values whose cardinality β is defined by the user. This way, the continuous space $[0,1]^{n_{dict}}$ is transformed into the discrete space $\Phi^{n_{dict}}$. In ITiS, ψ is based on the equal-width discretization algorithm that associates x with its nearest value in $\Phi = \{0, \frac{1}{\beta}, \frac{2}{\beta}, \ldots, \frac{\beta-1}{\beta}, 1\}$.

The semantic position $X(v_{u_j}(t))$ is computed by a forward time-slice density estimator $F(X, t, h_t, u_j)$ that is obtained by adapting the forward density estimator presented in [1] to our scenario. Formally, $X(v_{u_j}(t)) = \underset{X \in \Phi^{n_{dict}}}{\operatorname{argmax}} F(X, t, h_t, u_j)$ where the density function $F(X, t, h_t, u_j)$, that is measured for all possible semantic positions $X \in \Phi^{n_{dict}}$ of the tourist u_j at the time t, is maximized. The value of density at a given semantic position X is forward estimated on the basis of the sequence S of time-stamped textual documents which belong to $d_{consulted}(v_{u_j}(t))$ and have been consulted during the visits of the tourist in the time slice $[t-h_t, t]$. Formally: $S = \langle d_1, t_1\rangle, \ldots, \langle d_{|S|}, t_{|S|}\rangle$ where $\forall \langle d_i, t_i\rangle \in S$, $\exists \langle p_{j_k}, t'_{j_k}, t''_{j_k}, D_{j_k}\rangle \in v_{u_j}(t)$ such that: *i)* $d_i \in D_{j_k}$, *ii)* $t'_{j_k} \leq t_i \leq t''_{j_k}$ and *iii)* $t - h_t \leq t_i \leq t$.

A kernel density estimation is used in order to provide us a continuous estimate of the density $F(X, t, h_t, u_j)$ as sum of smoothed values of kernel functions $K_{h_t,u_j}(X, t)$:

$$F(X, t, h_t, u_j) = C_F \times \sum_{\langle d_i, t_i\rangle \in S} K_{h_t,u_j}(X - \omega_{d_i}, t - t_i).$$

where $\omega_{d_i} = [\omega_{d_i,1}, \ldots, \omega_{d_i,n_{dict}}]$ is the vector representation of the document $d_i \in D$, C_F is a constant value that makes $\sum_{X \in \Phi^{n_{dict}}} F(X, t, h_t, u_j) = 1$ and $K_{h_t,u_j}(X - \omega_{d_i}, t - t_i)$ is a semantic-temporal kernel function that uses a time fading factor to give more importance to recently consulted documents: $K_{h_t,u_j}(\Delta X, \Delta t) = \left(1 - \frac{\Delta t}{h_t}\right) K'(\Delta X)$.

Specifically, $K'(\Delta X)$ is the product of n_{dict} identical Gaussian kernel functions: $K'(\Delta X) = \prod_{q=1}^{n_{dict}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{\Delta X_q^2}{2\sigma^2}}$ where σ is a user defined smoothing parameter.

Next Destination Suggestion/Prediction. In order to suggest the next possible destination, ITiS assumes that: (i) a tourist moves towards a site spatially close to her/his current position, and (ii) he/she is not interested to visit that same site more than once. This way, the set of candidate destinations is defined as: $P_r(v_{u_j}(t)) = \{p \in p_{notVisited}(v_{u_j}(t)) | EuclideanDist(p, (x_{u_j}(t), y_{u_j}(t))) \leq r\}$ where r is the maximum spatial distance that the tourist is willing to cover. Among destinations in $P_r(v_{u_j}(t))$, ITiS suggests the tourist to move toward the destination which geo-references the documents whose consultation will lead to minimize the profile drift, that is:

$$p_{next}(v_{u_j}(t)) = \underset{p \in P_r(v_{u_j}(t))}{\operatorname{argmin}} drift(X(v_{u_j}(t)), \langle p, t, t, \delta(p)\rangle)$$

If several destinations minimize the drift measure, then ITiS suggests all of them ordered according to the Euclidean distance from the current geographical position of the tourist. Function $drift(\cdot, \cdot)$ can be computed by resorting to two alternative ways:

• By computing the cosine similarity between the semantic position of the tourist profile at the time t and the set of textual documents $\delta(p)$ which are geo-referenced to the candidate next destination p, $drift(\cdot, \cdot)$ is computed as: $drift(X(v_{u_j}(t)), \langle p, t, t, \delta(p)\rangle) = \frac{1}{|\delta(p)|} * \sum_{d \in \delta(p)} \frac{X(v_{u_j}(t)) \cdot \omega_d}{\|X(v_{u_j}(t))\| \|\omega_d\|}$.

• By measuring the variation of the semantic position of the tourist profile due the visit, $drift(\cdot, \cdot)$ is computed as: $drift(X(v_{u_j}(t)), \langle p, t, t, \delta(p)\rangle) = \|X(v_{u_j}(t)) - X(v_{u_j}(t), \langle p, t, t, \delta(p)\rangle)\|_2$.

3 EXPERIMENTS

We present an application where ITiS is employed to suggest the possible next destination of a tourist which visits the touristic area of Paris (France). Due to difficulty in obtaining real data, we asked eight users to perform virtual thematic tours in Paris. The basic hypothesis is that the tourist has a Java enabled mobile device with GPS and remotely access geo-referenced textual documents stored in the server. Documents have been selected by a tourism expert. In the experiment, we consider fifty-one candidate destinations located over the map of Paris and ninety-two textual documents. ITiS is run with $n_{dict} = 5$, $\sigma = 0.5$, $\beta = 20$. h_t is appropriately set in order to temporally consider, for each tourist, the entire set of stored visits. r is set to 32 Kms in order to consider all sites over the map as candidate destinations to be suggested. To evaluate how much a suggested destination p matches the interest of the tourist u, we compute the following score: $score(u, p) = 1$ if u accepts to move toward p; 0 otherwise. By considering that $p_{next}(v_{u_j}(t))$ may suggest a set of (equivalent) destinations (denoted as P_{next}), we define:

$$score(u, P_{next}) = \frac{1}{|P_{next}|} \sum_{p_i \in P_{next}} score(u, p_i).$$

The destinations suggested by ITiS for each tourist with both the cosine similarity measure and the semantic variation measure have been analyzed. Due to space limitations, we report only the average score computed for the destination predicted/suggested for the eight tourists. The average score obtained with cosine similarity (semantic variation) measure is 0.88 (0.69) These results enlighten that cosine similarity measure significantly outperforms semantic variation measure. In particular, the destination suggested by ITiS, using the cosine similarity measure, is approved by the tourist in the 88% of cases.

4 CONCLUSION

In this paper, we have proposed the use of forward time-slice density estimation approach in order to measure the drift of the tourist's interests by taking into account the geographical position of the tourist and the thematic history of her/his visited sites. Drift of the tourist's interests due to the movement toward a suggested destination is measured by resorting to either the cosine similarity measure or the semantic variation measure. Results on a real-world dataset evaluate effectiveness and accuracy of the proposed approach.

Acknowledgments. This work is partial fulfillment of the objective of ATENEO-2009 project "Estrazione, Rappresentazione e Analisi di Dati Complessi".

REFERENCES

[1] Charu C. Aggarwal, 'On change diagnosis in evolving data streams', *IEEE Trans. Knowl. Data Eng.*, **17**(5), 587–600, (2005).

[2] Michelangelo Ceci and Donato Malerba, 'Classifying web documents in a hierarchy of categories: a comprehensive study', *J. Intell. Inf. Syst.*, **28**(1), 37–78, (2007).

[3] Anna Monreale, Fabio Pinelli, Roberto Trasarti, and Fosca Giannotti, 'Wherenext: a location predictor on trajectory pattern mining', in *KDD*, pp. 637–646. ACM, (2009).