# Semantic-Based Destination Suggestion in Intelligent Tourism Information Systems

Michelangelo Ceci, Annalisa Appice, and Donato Malerba

Dipartimento di Informatica, Università degli Studi di Bari
via Orabona, 4 - 70126 Bari - Italy
{ceci, appice, malerba}@di.uniba.it

**Abstract.** In recent years, there has been a growing interest in mining trajectories of moving objects. Advances in this data mining task are likely to support the development of new applications such as mobility prediction and service pre-fetching. Approaches reported in the literature consider only spatio-temporal information provided by collected trajectories. However, some applications demand additional sources of information to make correct predictions. In this work, we consider the case of an on-line tourist support service which aims at suggesting places to visit in the nearby. We assume tourist interests depend both on her/his geographical position and on the "semantic" information extracted from geo-referenced documents associated to the visited sites. Therefore, the suggestion is based on both spatio-temporal data as well as on textual data. To deal with tourist's interest drift we apply a time-slice density estimation method. Experimental results are reported for two scenarios.

## 1 Introduction

Tourism has become, in the second part of last century, one of the most important economic activities in the world. According to World Travel and Tourism Council (WTTC), in 2005, about 11% of World Gross Domestic Product (GDP) was generated by the tourism sector and a considerable part (more than 200 million people) of the global workforce is employed [3]. In addition, it is predicted to be one of a few businesses that will continue to grow at an appreciable rate (around 5% per year) and to generate job opportunities in the future. Tourism is nowadays an important vehicle for regional and national developments and, for many countries, it represents the major contributor to the local economy.

More recently, tourism has become an extremely dynamic system [7] and the intensified (web-based) marketing efforts of all tourism organizations have paved the way for new advances in knowledge based technologies applied to the destination management problem [16]. According to the definition provided in [10], a tourism destination may be intended as a geographical area that offers to the tourist the opportunity of exploiting a variety of attractions and services. Recent advances in positioning technology (Global Positioning Systems - GPS) permit to track the tourist position in time and space (trajectories) and, then, to consider the tourist as a moving object into a predefined spatio-temporal space.

Knowledge on the past positions of the tourist can be used both to suggest the next preferred destination and to anticipate or pre-fetch possible services there.

In the machine learning literature, several methods have been proposed to learn location prediction models either from the history of movements of a single object [11, 19] or from the movements of all objects in an area [13]. In both cases, the predictor capitalizes only on the movement histories (trajectories) and its induction is based on the assumption that objects tend to follow common paths.

In this work, we assume each tourist is an independent individual with personal preferences and interests, therefore, the location prediction model is tourist-specific and has to be learned only from the tourist's movement history. This adds complexity to the learning task, and to face the inherent challenges we resort to additional information made available by the tourist during her/his visits. In particular, we consider the documents consulted by the tourist at the visited sites in order to maintain an informative profile of the tourist. This profile, initially empty, is dynamically updated on the basis of both the current spatial position of the tourist and the textual content of the consulted documents.

The main assumption of this work is that a tourist moves towards a close destination which is as much "semantically" consistent with her/his profile as possible. This assumption is supported by the observation of how tourists typically use guides, either paper or electronic (for examples of iPhone & iPod guides see `http://www.phaidon.com/travel`). Indeed, most of tourist guides suggest a set of thematic itineraries and tourists choose and follow one of them according to their specific interests and preferences. Additionally, electronic guides may also provide several documents which describe different aspects of a touristic site. While visiting a site, a tourist may consult related documents which are of her/his interest. In this work, we use information on consulted documents to update the tourist profile and we predict the next destination by minimizing the drift of tourist interests. To deal with tourist's interest drift we apply a time-slice density estimate [1] in order to measure the rate of change of tourist's interests over a time horizon when the tourist moves towards the new destination. In case a tourist consults no documents, only geographical information is used to suggest the next destination. As in many recommendation systems, the limit of the present approach is represented by tourists who perform "random" explorations.

The paper revises and extends the work presented in [4]. The paper is organized as follows. In the next section, some related works are discussed. In Section 3, the proposed method is presented, while empirical results are reported and discussed in Section 4. Conclusions are drawn in Section 5.

## 2 Related Works

Roots of this work are in the research field of moving objects prediction. The pioneering work in this area is presented in [12] where moving object behavior is modeled as repetitions of elementary movement patterns (e.g., linear or circular). The preferential next location is suggested by means of a mobile motion prediction algorithm that is highly sensitive to random movements of the object.

Subsequently, Markov chain models have been studied in order to estimate the probability of an object's movement from one region or cell to another at the next time period. Ishikawa et al. [9] propose to derive transition probabilities between cells over the space from indexed trajectories. Histograms are used to predict the next cell in which the object would probably move in the future.

Other approaches use sequential patterns in order to model trajectories in terms of ordered sequences of time-stamped locations [8, 18]. Most of these approaches try to predict the movement of an object on the basis of the assumption that people typically follow the crowd. Morzy [14] proposes to periodically mine offline historical data of other objects moving on the same area and discover frequent trajectories of objects representing popular movement routes. The unknown location of a moving object can be predicted by ranking on-line trajectories which match past history of the object, according to support and confidence, and using the selected trajectory to predict next destination. More recently, Monreale et al. [13] propose *WhereNext* which extracts trajectory patterns as a concise representation of behavior of moving objects, that is, sequences of regions frequently visited within a travel time. A decision tree is then learned from trajectory patterns that insist a certain area and it is used to predict the next location of a new trajectory by finding the best matching path in the tree.

Methods described above suggest the destination of a moving object based upon the movement history of either the object itself or the objects which move in the surrounding area. Although, space and time are considered, none of these methods takes into account semantic information which may be descriptive of the object profile and be bearing of information on the next destination. The concept of trajectories enriched with semantic information was originally formalized in [2] where authors consider, as semantic information, the name of the geographic layer (e.g., hotels, museums) associated to each site in the trajectories (called stops). In a semantic based moving environment, as that considered in this paper, we do not consider only the information on the geographical layer, but all the semantic information which can be automatically extracted from textual documents possibly geo-referenced with the trajectory sites. This semantic information is intended to express the interests, preferences and needs of the object. As in a stream, each time the object moves towards a new site, semantic information geo-referenced with the site contribute to dynamically construct/update the object profile. Hence, it is reasonable to assume that an object moves towards a site that slightly changes the profile or, in other terms, that is semantically close to the object profile. Following this idea, our point of view is that suggesting a semantic based next destination can be intended as an application of the change diagnosis in evolving data streams. In this area, the seminal work is that of Aggarwal [1], which firstly proposes to capture the change of spatially referenced characteristics over time with the concept of velocity density. The idea of velocity density is that of measuring the rate of change of data concentration at a given spatial location over a user-defined time horizon. Our assumption is that the destination which is spatially close to the current one and which minimizes the rate of change in profile is the most probable next destination.

## 3 ITiS (Intelligent Tourism information System)

The applicative task we address is that of suggesting the next destination of a tourist which moves on a map given: (1) a spatial referencing system that permits to uniquely define a spatial position on the map (e.g., latitude and longitude); (2) the set of destinations, each of which has a spatial position with respect to the spatial referencing system and geo-references a set of textual documents; (3) the current tourist position with respect to the spatial referencing system; (4) the trajectory already followed by the tourist and its associated profile which is updated on the basis of the semantic information extracted from textual documents consulted by the tourist over a user-defined time horizon.

ITiS addresses the suggestion task by updating the tourist profile each time the tourist visits a new site. The profile update operations take into account three basic assumptions. First, a tourist is not asked to consult documents before the visit starts, but documents geo-referenced at a site can be consulted only when the tourist is visiting the site itself and the profile is updated according to the new set of consulted documents. Second, a tourist consults a document if document content is interesting for him/her. Third, the content of documents recently consulted is more interesting for the tourist than that of documents consulted in the past. Based on the tourist profile, a time-slice density estimation is then used to suggest the preferential next destination of the tourist.

Before presenting how the suggestion is performed, we introduce some preliminary definitions, describe how to extract the semantic information from the consulted textual documents in order to update the profile, how to use the time-slice density estimation in order to suggest next destination.

### 3.1 Preliminary Concepts

Let $P = \{p_i = \langle x_{p_i}, y_{p_i}\rangle | i = 1 \ldots n\}$ 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 $D = \{d_j | j = 1 \ldots, N\}$ 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 function $\delta$ is not injective, that is, the same textual document can be geo-referenced to two 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:

$$v_{u_j}(t) = \big(\langle p_{j_1}, t'_{j_1}, t''_{j_1}, D_{j_1}\rangle, \ldots, \langle p_{j_s}, t'_{j_s}, t''_{j_s}, D_{j_s}\rangle\big) \tag{1}$$

where $p_{j_k} \in P$ represents the $k$-th ($k = 1, \ldots, s$) 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:

$$\begin{cases} t''_{j_k} \leq t'_{j_{k+1}} & iff \quad k \leq s-1 \\ t''_{j_k} \leq t & iff \quad k = s \end{cases}$$

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

### 3.2 Document Representation

A document is pre-processed in order to remove *stopwords*, such as articles, adverbs, prepositions and other frequent words and determine equivalent stems (*stemming*) by means of Porter's algorithm for English texts [15]. Pre-processed documents are subsequently 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 $d \in D$ (*Term Frequency*),
- $TF_C(w) = max_{d \in C} TF_d(w)$ the maximum value of $TF_d(w)$,
- $DF_C(w) = \frac{|\{d \in C| \quad w \ occurs \ in \ d\}|}{|C|}$ the percentage of documents in $C$ in which $w$ occurs (*Document Frequency*),
- $CF_{C', C'', \dots C^{(s)}}(w)$ is the number of sets of documents where the token $w$ occurs. In this formulation, sets of documents are denoted as $C', C'', \dots C^{(s)}$ with $C^{(i)} \subseteq D$ (*Category Frequency*).

Then the following measure [5] permits to associate a token $w_i$ with its score $v_i$ and to select relevant tokens for the representation of documents in $D$:

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

Tokens in $d_{consulted}(v_{u_j}(t)$ that minimize $v_i$ ($maxTF \times DF^2 \times ICF$ criterion[1]) are penalized since they are used in 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 reasonably represent documents in $D$. The best $n_{dict}$ tokens form the *dictionary* $Dict(v_{u_j}(t))$ of the tourist $u_j$ at the time $t$.

Once $Dict(v_{u_j}(t))$ is determined, it is used to index the $N$ documents in $D$ according to the normalized $TF \times idf$ measure [17]. In the matrix representation:

$$\omega(v_{u_j}(t)) = \begin{bmatrix} \omega_{1,1} & \omega_{1,2} & \dots \\ \vdots & \ddots & \dots \\ \vdots & \vdots & \omega_{N,n_{dict}} \end{bmatrix} \tag{3}$$

[1] ICF stands for Inverse Category Frequency

where the $TF \times idf$ measure is computed as $\omega_{p,q} = \frac{TF_{d_p}(w_q) \times ln \frac{N}{1+N \times DF_D(w_q)}}{\|\omega(v_{u_j}(t))\|_1}$ with $d_p \in D$ and $w_q \in Dict(v_{u_j}(t))$. It is noteworthy that $\omega_{p,q} \in [0, 1]$.

### 3.3 Time-Slice Density Based Profile

We define the profile of the tourist $u_j$ at the time $t$ as the triple $\langle x_{u_j}(t), y_{u_j}(t), X(v_{u_j}(t))\rangle$, where $(x_{u_j}(t), y_{u_j}(t))$ represents the geographical position of the tourist, while $X(v_{u_j}(t))$ represents 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 a discretization of the interval $[0, 1]$ according to a supervised discretization function $\psi : [0, 1] \rightarrow \Phi$, where $\Phi$ is a finite set of values whose cardinality $\beta$ is apriori defined by the user. This way, the continuous space $[0, 1]^{n_{dict}}$ is mapped into a discrete space $\Phi^{n_{dict}}$. In ITiS, $\psi$ is based on the equal-width discretization algorithm [6] that associates $x$ with its nearest value in $\Phi = \{0, \frac{1}{\beta}, \frac{2}{\beta}, \dots, \frac{\beta-1}{\beta}, 1\}$.

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,

$$X(v_{u_j}(t)) = \underset{X \in \Phi^{n_{dict}}}{\operatorname{argmax}} F(X, t, h_t, u_j) \tag{4}$$

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, $S$ is defined as: $S = \langle d_1, t_1\rangle, \dots, \langle d_{|S|}, t_{|S|}\rangle$, where $\forall \langle d_i, t_i\rangle \in S, \quad \exists \langle p_{j_k}, t'_{j_k}, t''_{j_k}, D_{j_k}\rangle \in v_{u_j}(t)$ 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 to provide us a continuous estimate of the density $F(X, t, h_t, u_j)$ as sum of smoothed values of kernel functions $K_{h_t,u_j}(X, t)$.

$$F(X, t, h_t, u_j) = C_F \times \sum_{\langle d_i, t_i\rangle \in S} K_{h_t,u_j}(X - \omega_{d_i}, t - t_i) \tag{5}$$

In this Eq., $\omega_{d_i} = [\omega_{d_i,1}, \dots, \omega_{d_i,n_{dict}}]$ is the vector representation of the document $d_i \in D$ (see Eq. 3), $C_F$ is a constant 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:

$$K_{h_t,u_j}(\Delta X, \Delta t) = \left(1 - \frac{\Delta t}{h_t}\right) K'(\Delta X) \tag{6}$$

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

### 3.4 Preferential Next Destination Suggestion

In order to suggest the preferential next destination, ITiS assumes that a tourist moves towards a site spatially close to her/his current position and is not interested to visit that same site more than once. According to these assumptions, the set of candidate next destinations is defined as:

$$P_r(v_{u_j}(t)) = \{p \in p_{notVisited}(v_{u_j}(t)) | EuclideanDistance(p, (x_{u_j}(t), y_{u_j}(t))) \leq r\} \tag{7}$$

where $r$ is the maximum spatial distance that the tourist is willing to cover.

Among the candidate destinations in $P_r(v_{u_j}(t))$, ITiS suggests the tourist to move toward the destination which geo-references the set of documents whose consultation will lead to minimize her/his profile drift, that is:

$$p_{next}(v_{u_j}(t)) = \underset{p \in P_r(v_{u_j}(t))}{\operatorname{argmin}} \; drift(X(v_{u_j}(t)) \; , \; \langle p, t, t, \delta(p) \rangle) \tag{8}$$

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 $drift(\cdot, \cdot)$ in Eq. 8 permit to rank destinations in order of preference. It can be computed by resorting to three alternative ways, that is:

- by computing the cosine similarity between the semantic position of the tourist profile (see Eq. (4)) at the time $t$ and the set of textual documents $\delta(p)$ which are geo-referenced to the candidate next destination $p$;
- by measuring the variation of the semantic position of the tourist profile (see Eq. (4)) due to the simulation of a visit to the candidate next destination;
- by measuring the variation of the density function of the tourist (see Eq. (5)) due to the simulation of a visit to the candidate next destination.

By computing the cosine similarity, $drift(\cdot, \cdot)$ is obtained as:

$$drift(X(v_{u_j}(t)), \langle p, t, t, \delta(p) \rangle) = \frac{1}{|\delta(p)|} \cdot \sum_{d \in \delta(p)} \frac{X(v_{u_j}(t)) \cdot \omega_d}{\|X(v_{u_j}(t))\| \|\omega_d\|} \tag{9}$$

Alternatively, by measuring the variation of the semantic position of the tourist profile due the visit, $drift(\cdot, \cdot)$ is obtained as:

$$drift(X(v_{u_j}(t)), \langle p, t, t, \delta(p) \rangle) = \|X(v_{u_j}(t)) - X(v_{u_j}(t), \langle p, t, t, \delta(p) \rangle)\|_2 \tag{10}$$

Finally, by measuring the variation of the density function of the tourist due to the simulation of a visit to the candidate next destination, $drift(\cdot, \cdot)$ is:

$$\begin{aligned} drift( \; & X(v_{u_j}(t)), \langle p, t, t, \delta(p) \rangle) = \\ & F(X(v_{u_j}(t)), t, h_t, u_j) - F((X(v_{u_j}(t), \langle p, t, t, \delta(p) \rangle), t, h_t, u_j) \end{aligned} \tag{11}$$

## 4 Experiments

In this Section we present two applications where we use ITiS to suggest the next destination of a tourist on the basis of the time-slide density estimation of her/his semantic-base profile. We consider two touristic areas, that is, Stockport (United Kingdom) and Paris (France).

Due to difficulty in obtaining real data, we asked sixteen users to perform virtual thematic tours over either Stockport or 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 stored on the server have been selected by a tourism expert.

In the experiments, ITiS is run with the following parameter values: $n_{dict} = 5$, $\sigma = 0.5$, $\beta = 20, 30$. Additionally, $h_t$ is appropriately set in order to temporally consider, for each tourist, the entire set of stored visits. $r$ is set to 3 Kms in Stockport experiment and it is set to 32 Kms in Paris experiment. This choice of $r$ permits to consider all sites in the corresponding maps as candidate destinations to be suggested. As to $\sigma$, if it is chosen too small then spurious fine structure becomes visible, while if $\sigma$ is too large then the bimodal nature of the distribution is obscured. As to $n_{dict}$, it is mainly related to the number of different themes a tourist can be interested in (we have only 8-9 categories, see Tables 1 and 5).

To evaluate how much a suggested destination $p$ matches the interest of the tourist $u$, we compute the following score:

$$score(u, p) = \begin{cases} 1 \text{ if } u \text{ accepts to move toward } p \\ 0 \text{ otherwise.} \end{cases} \tag{12}$$

By considering that $p_{next}(v_{u_j}(t))$ (see Eq. 8) may suggest a set of (equivalent) destinations, denoted as $P_{next}$, then:

$$score(u, P_{next}) = \sum_{p_i \in P_{next}} score(u, p_i)/|P_{next}| \tag{13}$$

### 4.1 Stockport

Stockport is a large town in Greater Manchester and, in this study, we consider eight tourists who visited Stockport by moving from one site to another. We consider thirty candidate destination sites which, for descriptive purposes, are classified into eight categories, namely, transport net, monument and museum, park, restaurant and hotel, school and university, shopping, sport and entertainment and church (see Table 1). Each site has a geographic position (latitude and longitude) over the map of Stockport (see Figure 1) and it geo-references a set of textual documents (including Wikipedia pages) which are descriptive of the site attractiveness. In all, we consider a total of sixty-four textual documents. Additionally, we have tracked the moving trajectory and the consulted documents of the tourists. A brief description of tourists' trajectories is reported in Table 2.

The destinations suggested by ITiS by considering either the cosine drift or the semantic drift or the density drift are reported in Table 3 for $\beta = 20$ and

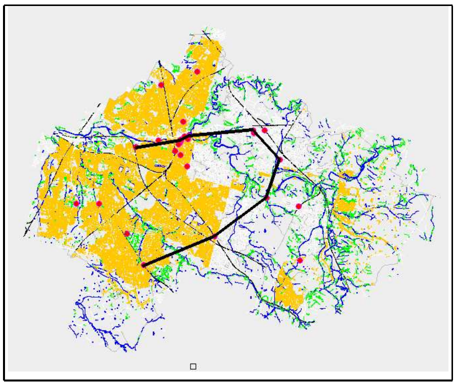

**Fig. 1.** A trajectory in Stockport

in Table 4 for $\beta = 30$. The score shows that the cosine drift measure generally outperforms both the semantic drift measure and the density drift measure. In particular, the semantic drift measure often leads to suggest additional non-interesting destinations which result in decreasing the score value. Anyway, we observe that increasing $\beta$ (from 20 to 30) semantic drift measure generally reduces the number of suggested destinations. This is mainly due to the fact that semantic drift measure is affected by the discretization function more than the cosine drift measure. Differently, we observe that the cosine drift measure and the density drift measure are not very sensitive to $\beta$.

### 4.2 Paris Dataset

In this experiment, we consider fifty-one candidate destinations located over the map of Paris. Destinations are classified into seven categories, namely, transport net, monument, museum, park, school and university, sport and entertainment and church (see Table 5). In all, we consider ninety-two textual documents and results are collected on eight tourists.

The destinations suggested by ITiS are reported in Tables 7-8. By analyzing the average score, we observe a significant improvement with respect to the results obtained with Stockport data. This is motivated by the fact that in this experiment, tourists followed trajectories where it is possible to recognize well defined thematic interests of the tourist (e.g., a thematic interest for the impressionist art). This depends on the fact that Paris offers a wide spectrum of touristic attractions which may match distinct thematic interests of a possible tourist. By comparing the score obtained with the drift measures, we observe that results confirm the main considerations drawn from the analysis of Stockport data (i.e., cosine drift measure outperforms both semantic drift measure and density drift measure). Also in this experiment, score results do not appear to be very sensitive to the $\beta$ value.

| Site | Category |
|---|---|
| Hazel Grove Rail Station, Bramhall Rail Station, Rose Hill Marple Rail Station, Stockport Rail Station, Bredbury Rail Station | TRANSPORT NET |
| The Co-Op Bank Pyramid, Wellington Mill, Stockport Viaduct, Staircase House, Stockport Town Hall, Air Raid Shelters Museum | MONUMENT AND MUSEUM |
| Vernon Park | PARK |
| The Bowling Green, Duke Of York, The Hare & Hounds, The Romper Inn, Bredbury Hall Hotel | RESTAURANT AND HOTEL |
| Stockport College (Town Centre), Stockport College (Heaton Moore) | SCHOOL AND UNIVERSITY |
| TK Maxx, Merseyway, Debenhams, John Lewis Cheadle | SHOPPING |
| Stockport Plaza, Bramhall Park Golf Club | SPORT AND ENTERTAINMENT |
| Unitarian Church, Salvation Army Church, Salvation Army Church (Chesire), St. Barnabas Parish Church, St. Elizabeth Church | CHURCH |

**Table 1.** Candidate destinations description for Stockport data.

## 5 Conclusion

In this paper, we have presented a forward time-slice density estimation that is tailored for suggesting the next destination where a tourist reasonably would move towards. The forward time-slice density estimation approach is used to measure the drift of the tourist's interests by taking into account the current geographical position of the tourist and the thematic history of her/his visited sites over a user-defined time horizon. For each visited site, we consider the set of textual documents geo-referenced to the site that the tourist has consulted during the visit time. Results on two datasets show both effectiveness and accuracy of the proposed approach. By comparing destinations suggested by three distinct measures, that is, cosine drift, semantic drift and density drift, we have observed that cosine drift measure outperforms other measures.

As future work, we would extend experiments by considering a higher number of thematic tourist trajectories, although data privacy and novelty of the considered application scenario make difficult to obtain a large base of touristic data. Additionally, we also intend to take into account constraints in the suggestion step. This way, for example, the system can avoid to suggest specific destinations during closing times or out of the available budget. Finally, we intend to perform new experiments by using a spatio-temporal kernel function.

## Acknowledgment

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

| Tourist | Visited sites | No. of consulted docs |
|---|---|---|
| T1 | Stockport Rail Station | 4 |
| T2 | Salvation Army Church | 1 |
| T3 | Salvation Army Church | 1 |
| | Salvation Army Chesire | 1 |
| | St. Elizabeth Church | 3 |
| | Unitarian Church | 2 |
| T4 | The Bowling Green | 1 |
| | Duke Of York | 2 |
| | The Hare & Hounds | 2 |
| T5 | The Co-Op Bank Pyramid | 1 |
| | Stockport Viaduct | 2 |
| | The Hare & Hounds | 1 |
| | Vernon Park | 1 |
| | Stockport Plaza | 3 |
| T6 | Wellington Mill | 2 |
| | Staircase House | 3 |
| T7 | TK Maxx | 2 |
| | Merseyway | 2 |
| | Debenhams | 1 |
| T8 | Stockport Plaza | 2 |
| | Stockport College (Town Centre) | 1 |
| | John Lewis Cheadle | 1 |

**Table 2.** A description of both the tourist trajectories and the corresponding number of consulted documents in Stockport data.

## References


1. C. C. Aggarwal. On change diagnosis in evolving data streams. *IEEE Trans. Knowl. Data Eng.*, 17(5):587–600, 2005.
2. L. O. Alvares, V. Bogorny, B. Kuijpers, J. A. F. de Macêdo, B. Moelans, and A. A. Vaisman. A model for enriching trajectories with semantic geographical information. In H. Samet, C. Shahabi, and M. Schneider, editors, *GIS*, page 22. ACM, 2007.
3. R. Baggio and L. Caporarello. Decision support systems in a tourism destination: literature survey and model building . In *itAIS 2005 - 2nd Conference of the Italian Chapter of AIS: Organising Information Systems: Evolution of Studies in the Field*, 2005.
4. M. Ceci, A. Appice, and D. Malerba. Time-slice density estimation for semantic-based tourist destination suggestion. In *ECAI-poster session.* to appear, 2010.
5. M. Ceci and D. Malerba. Classifying web documents in a hierarchy of categories: a comprehensive study. *J. Intell. Inf. Syst.*, 28(1):37–78, 2007.
6. J. Dougherty, R. Kohavi, and M. Sahami. Supervised and unsupervised discretization of continuous features. In *ICML*, pages 194–202, 1995.
7. B. H. Farrell and L. Twining-Ward. Reconceptualizing tourism. *Annals of Tourism Research*, 31(2):274–295, 2004.
8. F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi. Trajectory pattern mining. In P. Berkhin, R. Caruana, and X. Wu, editors, *KDD*, pages 330–339. ACM, 2007.

| Tourist | Next Destination Suggestion | | | | | |
|---|---|---|---|---|---|---|
| | **CosineDrift** | **Score** | **SemanticDrift** | **Score** | **DensityDrift** | **Score** |
| T1 | Rose Hill Marple Station | 1 | Rose Hill Marple Station | 1 | Bramhall Rail Station | 1 |
| T2 | Salvation Army Church (Chesire) | 1 | Salvation Army Church (Chesire) | 1 | St. Elizabeth Church | 1 |
| T3 | St. Barnabas Parish Church | 1 | *more than one* | 1/3 | Stockport Town Hall | 0 |
| T4 | The Romper Inn | 1 | *more than one* | 3/5 | The Romper Inn | 1 |
| T5 | Stockport College (Town Centre) | 0 | *more than one* | 1/7 | Bredbury Hall Hotel | 0 |
| T6 | Vernon Park | 1 | *more than one* | 1/4 | St. Elizabeth Church | 0 |
| T7 | John Lewis Cheadle | 1 | John Lewis Cheadle | 1 | John Lewis Cheadle | 1 |
| T8 | Stockport Town Hall | 0 | Salvation Army Church | 0 | Vernon Park | 1 |
| **Avg.** | | **0.75** | | **0.54** | | **0.625** |

**Table 3.** Score computed over the destinations suggested by ITiS for the Stockport data ($\beta = 20$).


9. Y. Ishikawa, Y. Tsukamoto, and H. Kitagawa. Extracting mobility statistics from indexed spatio-temporal datasets. In J. Sander and M. A. Nascimento, editors, *STDBM*, pages 9–16, 2004.
10. J. Jafari. *Encyclopedia of Tourism.* Routledge, London, 2000.
11. H. Jeung, Q. Liu, H. T. Shen, and X. Zhou. A hybrid prediction model for moving objects. In *ICDE*, pages 70–79. IEEE, 2008.
12. G. Y. Liu and G. Q. M. Jr. A predictive mobility management algorithm for wireless mobile computing and communications. *Proceedings of the Fourth IEEE International Conference on Universal Personal Communications*, page 268272, 1995.
13. A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti. Wherenext: a location predictor on trajectory pattern mining. In J. F. E. IV, F. Fogelman-Soulié, P. A. Flach, and M. J. Zaki, editors, *KDD*, pages 637–646. ACM, 2009.
14. M. Morzy. Mining frequent trajectories of moving objects for location prediction. In P. Perner, editor, *MLDM*, volume 4571 of *Lecture Notes in Computer Science*, pages 667–680. Springer, 2007.
15. M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.
16. J. R. B. Ritchie and G. I. Crouch. *The competitive destination : a sustainable tourism perspective.* CABI, Oxon, UK, 2003.
17. F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
18. J. Yang and M. Hu. Trajpattern: Mining sequential patterns from imprecise trajectories of mobile objects. In Y. I. et al., editor, *EDBT*, volume 3896 of *Lecture Notes in Computer Science*, pages 664–681. Springer, 2006.
19. G. Yavas, D. Katsaros, Ö. Ulusoy, and Y. Manolopoulos. A data mining approach for location prediction in mobile environments. *Data Knowl. Eng.*, 54(2):121–146, 2005.

| Tourist | Next Destination Suggestion | | | | | |
|---|---|---|---|---|---|---|
| | **CosineDrift** | **Score** | **SemanticDrift** | **Score** | **DensityDrift** | **Score** |
| T1 | Bredbury Rail Station | 1 | Bredbury Rail Station | 1 | Bredbury Rail Station | 1 |
| T2 | Salvation Army Church (Chesire) | 1 | Salvation Army Church (Chesire) | 1 | Salvation Army Church (Chesire) | 1 |
| T3 | St. Barnabas Parish Church | 1 | *more than one* | 1/3 | Vernon Park | 0 |
| T4 | The Romper Inn | 1 | The Romper Inn | 1 | The Romper Inn | 1 |
| T5 | Stockport College (Town Centre) | 0 | Bramhall Rail Station | 0 | Bredbury Hall Hotel | 0 |
| T6 | Vernon Park | 1 | Bredbury Rail Station | 0 | St. Elizabeth Church | 0 |
| T7 | John Lewis Cheadle | 1 | John Lewis Cheadle | 1 | John Lewis Cheadle | 1 |
| T8 | Stockport Town Hall | 0 | Bredbury Rail Station | 0 | The Hare & Hounds | 0 |
| **Avg.** | | **0.75** | | **0.54** | | **0.5** |

**Table 4.** Score computed over the destinations suggested by ITiS for the Stockport data ($\beta = 30$).

| **Site** | **Category** |
|---|---|
| Gare de Lyon, Gare de l'Est | TRANSPORT NET |
| Palais-Royal, Place Vendôme, Conciergerie, Place des Victoires, Place des Vosges, Luxembourg Palace, Tour Eiffel, Place de la Concorde, Arc de Triomphe de l'Étoile, Église de la Madeleine, l'Opéra National de Paris, Place de la Bastille, Palais de l'Elysée, la défense, Montmartre, Avenue des Champs-Élysées' | MONUMENT |
| Musée Louvre, Musée National Picasso, Centre Pompidou, Musée Cluny, Musee Hôtel National des Invalides, Musée Orsay, Musée Rodin, Musée de Orangerie, Musée Jacquemart-André, Grévin, Musée des Gobelins, Musée de La Poste, Musée dArt Moderne de la Ville de Paris, Musée Marmottan Monet, Fragonard Musee du Parfum, Musée de Les Égouts de Paris | MUSEUM |
| Bois de Boulogne, Parc des Buttes Chaumont, La Villette, Canal Saint-Martin | PARK |
| Sorbonne | SCHOOL AND UNIVERSITY |
| Aquarium du trocadero, EuroDisney, Park Asterix, Moulin Rouge, Hard Rock café, Folies Bergère | SPORT AND ENTERTAINMENT |
| Basilique du Sacre Coeur, Sainte Chapelle, Saint Eustache, Notre Dame, Saint Marrie, Le Panthéon, Saint Étienne du Mont | CHURCH |

**Table 5.** Description of the candidate destinations in the Paris data.

| Tourist | Trajectory | Visited sites | No. of consulted docs |
|---|---|---|---|
| T1 | Traditional itinerary | Tour Eiffel - Champs Elisée - Arc de Triomphe de l'Étoile - Place de la concorde - Louvre - Notre dame | 9 |
| T2 | Museum itinerary | Musée du Louvre - Musée Orsay - Musée National Picasso - Musée Orangerie - Musée Jacquemart-André - Centre Pompidou | 8 |
| T3 | Church itinerary | Notre Dame - Sacre Cour - Pantheon - Madleine | 8 |
| T4 | Impressionism itinerary | Monmatre - Musée Orsay - Musée Orangerie - Musée Monet | 6 |
| T5 | Historical itinerary | Arc de Triomphe de l'Étoile - Place de la Bastille - Place de la Concorde - Concergerie - Place vendome - Place de josges - Egouts de Paris | 16 |
| T6 | Historical-political itinerary | Palais Royal - Luxembourg Palace - Palais de l'Elysée - La défense - Sorbonné - Musée National des Invalides - Place Vendome | 14 |
| T7 | Entertainment itinerary | Champs élisée - Acquarium du Trocadero - Musée du Parfum - Folies Bergére - Disneyland - Moulin Rouge - Park Asterix - Hard Rock Café | 12 |
| T8 | Nature itinerary | Canal SaintMartine - Parc des Buttes Chaumont - Bois de Boulogne - LaVillette | 8 |

**Table 6.** A description of the tourist trajectories and the corresponding number of consulted documents in Paris data.

| **Tourist** | **Next Destination Suggestion** | | | | | |
|---|---|---|---|---|---|---|
| | **CosineDrift** | **Score** | **SemanticDrift** | **Score** | **DensityDrift** | **Score** |
| T1 | Musée Orsay | 1 | *more than one* | 2/9 | Musée Orsay | 1 |
| T2 | Musée de La Poste | 1 | *more than one* | 3/3 | Musée de Les Égouts de Paris | 1 |
| T3 | Monmatre | 1 | *more than one* | 3/11 | Sainte Chapelle | 1 |
| T4 | Musée Rodin | 1 | *more than one* | 2/3 | Musée Jacquemart-André | 1 |
| T5 | Champs Elisée | 1 | *more than one* | 2/3 | Aquarium du trocadero | 0 |
| T6 | Park Asterix | 0 | *more than one* | 9/13 | Monmatre | 1 |
| T7 | Musée de La Poste | 1 | *more than one* | 4/4 | Gare de l'Est | 0 |
| T8 | Park Asterix | 1 | EuroDisney | 1 | EuroDisney | 1 |
| **Avg.** | | **0.87** | | **0.69** | | **0.75** |

**Table 7.** Score computed over the destinations suggested by ITiS for the Paris data ($\beta = 20$).

| **Tourist** | **Next Destination Suggestion** | | | | | |
|---|---|---|---|---|---|---|
| | **CosineDrift** | **Score** | **SemanticDrift** | **Score** | **DensityDrift** | **Score** |
| T1 | Musée Orsay | 1 | Musée Orsay | 1 | Musée Orsay | 1 |
| T2 | Grévin | 1 | *more than one* | 3/3 | Musée de Les Égouts de Paris | 1 |
| T3 | Saint Étienne du Mont | 1 | *more than one* | 2/5 | Sainte Chapelle | 1 |
| T4 | Musée Rodin | 1 | *more than one* | 2/3 | Musée Jacquemart-André | 1 |
| T5 | Champs Elisée | 1 | Musée de La Poste | 1 | Aquarium du trocadero | 0 |
| T6 | Park Asterix | 0 | Place de la Bastille | 1 | Monmatre | 1 |
| T7 | Musée Jacquemart-André | 0 | *more than one* | 1/2 | Gare de l'Est | 0 |
| T8 | Park Asterix | 1 | Luxembourg Palace | 0 | EuroDisney | 1 |
| **Avg.** | | **0.75** | | **0.70** | | **0.75** |

**Table 8.** Score computed over the destinations suggested by ITiS for the Paris data ($\beta = 30$).