Zanker M., Fuchs M., Seebacher A., Jessenitschnig M., Stromberger M.: An Automated Approach for Deriving Semantic Annotations of Tourism Products based on Geospatial Information. In: W. Höpken et al. (Hrsg.): Information and Communication Technologies in Tourism, Proceedings ENTER 2009, Springer, 2009, pp. 211-221.

An Automated Approach for Deriving Semantic Annotations of Tourism Products based on Geospatial Information

Markus Zanker[a], Matthias Fuchs[b/c], Alexander Seebacher[a], Markus Jessenitschnig[a]
and Martin Stromberger[d]

[a] Institute for Applied Informatics
University Klagenfurt, Austria
{firstname.lastname}@uni-klu.ac.at

[b] European Tourism Research Institute (ETOUR)
Mid-Sweden University, Sweden
matthias.fuchs@miun.se

[c] e-Tourism Competence Centre Austria, Innsbruck (ECCA), Austria
matthias.fuchs@etourism-austria.at

[d] LOGO Geoinformationssysteme GmbH, Klagenfurt, Austria
stromberger@logo.co.at

Abstract

High quality product data is a necessary prerequisite for supporting efficient browsing and recommendation procedures on e-commerce platforms. This is especially true for the tourism domain where an abundance of information can easily overwhelm users. Although classification data such as to which category (e.g. accommodation, restaurant or sight) a tourism product belongs is usually directly available, qualitative information, such as proximity to a lake or opportunities for dining or shopping, is rarely provided in a structured way. As a consequence, users can not restrict their search on these criteria; rather, it would require costly manual information acquisition efforts. In this paper we propose a knowledge-based approach that automatically associates such qualitative concepts with tourism products based on their geographic coordinates and their spatial proximity. An initial evaluation of the approach that considered automatically generated annotations within different regions suggests that it can be used as an alternative to domain experts.

Keywords: Data extraction, semantic annotation, geospatial information, recommender system

Zanker M., Fuchs M., Seebacher A., Jessenitschnig M., Stromberger M.: An Automated Approach for Deriving Semantic Annotations of Tourism Products based on Geospatial Information. In: W. Höpken et al. (Hrsg.): Information and Communication Technologies in Tourism, Proceedings ENTER 2009, Springer, 2009, pp. 211-221.

1 Introduction

Due to the ever increasing abundance of information available on the Web, users are quickly overwhelmed if they are not sufficiently supported in their decision making processes. Search tools and recommender systems help users to narrow down choices and support the online exploration of large item sets. However, the interaction experience with such tools depends heavily on the quality of the underlying data. For instance, when looking for appropriate accommodation simple categorical information like the hotel class or the price range is usually insufficient for making a decision. Rather, so called *soft* criteria, such as appropriateness for specific tourist types (e.g. families) or specific interests (e.g. art or nightlife) need to be considered (Miles et al., 2000). However, the majority of this qualitative information, although relevant for the user's judgement of tourism products in the consumption decision is not available in a structured representation, thus is unable to be utilized by parametric search tools or knowledge-based recommendation systems. With the advent of the geospatial Web and the wide distribution of GPS devices (Scharl & Tochtermann, 2007) geo-tagging (i.e. adding geospatial context information) has become popular and can now be considered common for tourism products. However, although geo-tags ease the exploration of a tourism destination with the help of GIS like Google Earth, they do not reduce the information overload experienced by users. Thus, the need for adding value to online decision support systems by integrating derived semantic knowledge remains.

Therefore, we propose a computation scheme that exploits the geo-tags of different tourism service providers, general POIs (points of interest) and user-generated content to automatically derive semantic annotations. The approach builds on the rather obvious assumption that spatial proximity transfers semantic meaning from one object to the other. For instance, given two hotels where one is located closer to the town centre, then the closer one will be considered *ceteris paribus* as possessing more of a fuzzy concept like the *downtown factor* than the one farther away. Though the approach appears quite simple at first glance, to the best of our knowledge it has neither been proposed nor put into practice in the tourism context until now. However, online users could profit enormously by narrowing down the product space by using automatically derived, semantically enriched information such as *neighbouring shopping*, *sunbathing* or *recreational facilities*, respectively. Currently, qualitative product information like "a hotel recommendable for those who like to go shopping" can only be derived if multi-dimensional community ratings are available as discussed in the section on related work.

The following sections introduce a motivating example and formalize the technical approach. We then report on our experiences from a preliminary evaluation. Finally, we explore related work and present our conclusions.

2 Motivating example

To illustrate our approach, we consider a motivating example that ranks different accommodation offers based on their proximity to restaurants and bars. Let's assume a hedonistic couple that wants to spend a few days on vacation may choose between three different hotels of the same category and with comparable service characteristics. The preferred leisure activities of the two are dining out, going to bars and enjoying the nightlife. Therefore, an additional characteristic that quantifies the aptness of each hotel for those that like to go out and enjoy the nightlife, i.e. the *nightlife factor*, would be of great help. However, such qualitative information is rarely available on tourism online platforms. One possibility would be to compute such semantic annotations based on the geo-coding of the hotels and those objects related to the nightlife factor. Table 1 lists the available geographic information with Cartesian coordinates.

Name	Type	Coordinates (x/y)
1	Hotel	(0/0)
2	Hotel	(2/2)
3	Hotel	(4/-1)
A	Restaurant	(-1/0)
B	Bar	(1/-1)
C	Bar	(0/2)
D	Restaurant	(-2/-2)

Table 1: Product catalogue

The goal is to compute a utility score for each of the three hotels based on their proximity to restaurants and bars. In addition, a maximum Euklidean distance[1] is assumed that restricts which items are considered as being in the neighbourhood of an item (Chajed et al., 1993). The setting of such a limit depends on the concept under consideration and what is generally considered to be acceptable in this respect. For instance, dining out or having a drink is obviously more sensitive to distance than visiting different cultural sights as in the first case one would probably prefer to take a

[1] Note that an offline pre-computation of real distances between objects based on road maps and a route planning algorithm is recommended in practice.

taxi or walk instead of driving a car. In practice, this distance parameter could be either chosen based on expert opinions, empirically researched or dynamically set by online users themselves. Figure 1 presents the different items in a two dimensional space where hotels are depicted as rectangles and restaurants and bars as triangles. The circles denote the neighbourhood of each hotel with an assumed maximum distance of 2.5 units.

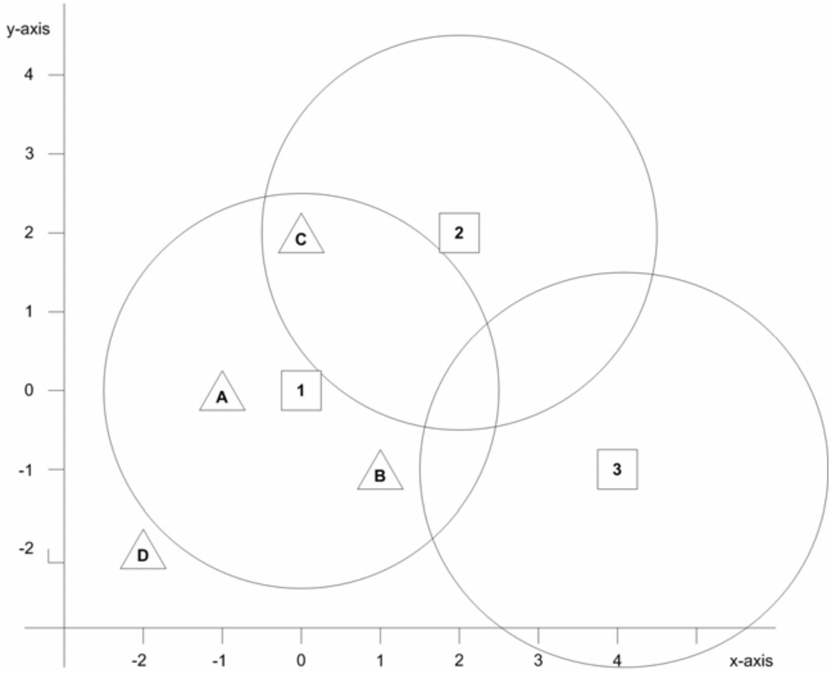

Figure 1: Motivating example

Without any further computation it can be quickly observed that Hotel 1 is the most ideally located for going out as three of the restaurants and bars are in its neighbourhood, while only one bar is in the neighbourhood of Hotel 2 and none is close to Hotel 3. In light of this example, we will present a more generic computation algorithm in the following section.

3 Semantic Annotation based on Geospatial Information

The task of associating semantic concepts with tourism products builds on the following prerequisites:

- A set of uniquely identifiable products *P* (e.g., see Table 1)
- An initial taxonomy *T* that allows the differentiation of *P* into different product types (i.e. *type(1) = 'Hotel'*)
- Geographic coordinates (i.e. *coord(1) = 0/0*[2])
- Semantic concepts *C* including domain knowledge to be able to define each concept on the basis of its proximity to different product types in *T*.

The goal is to compute for each item *p* in *P* and concept *c* in *C* a degree that tells the confidence for *c* to be a characterising property of *p*.

First, we define the function *nh(p)* that returns all items that are in the neighbourhood of item *p*:

$$nh(p) = \{q \in P \mid type(q) \in types(c) \wedge dist(p,q) \leq \max dist_c \quad (1)$$

$type(q)$ … type of a tourism product

$types(c)$ … function returning the set of product types that when in the neighbourhood support a concept *c*

$dist(p,q)$ … Euklidean distance between two items *p* and *q*

$\max dist_c$ … constant that sets the maximum distance for a concept *c*

Example: Let concept *c* be the *nightlife factor* and the maximum distance be 2.5. Furthermore, the concepts supporting *c* are consequently: *types(c) = {restaurant, bar}.* As a result *nh(1) = {A, B, C}*, *nh(2) = {C}* and *nh(3)={}.*

[2] Note that we used simple Cartesian coordinates (2D) from which distances can be easily computed. In practice, geographic coordinates are typically given in a geographic frame of reference, such as GPS or Lambert. However, these latitude and longitude angles can always be transformed into a planar space with some error. For further information on spherical trigonometry and transformation functions the reader is referred to the Mathworld Encyclopedia (Weisstein, 2008).

In a next step, the proximity between two items is defined:

$$proximity_c(p,q) = 1 - \frac{dist(p,q)}{\max dist_c} \qquad (2)$$

Example: *proximity(1,A)= 1–1/2.5= 0.6, proximity(1,B)= 0.43, proximity(1,C)= 0.2*

Equation (2) uses the inverse of the linear distance and normalizes it on the interval [0..1]. However, for different concepts different proximity functions might be sensible such as, for instance, penalizing distance on a logarithmic or an exponential scale.

Finally, the confidence for a *tuple (p,c)* can be computed as given in (3). In (4) it is normalized relative to the maximum confidence of any product *r* in *P* for concept *c*.

$$confidence(p,c) = \sum_{q \in nh(p)} proximity_c(p,q) \times w_q \qquad (3)$$

w_q … optional weighting factor for product *q*

$$confidence_{norm}(p,c) = \frac{confidence(p,c)}{Max_{r \in P} confidence(r,c)} \qquad (4)$$

Example: We assign uniform weights to all products in the neighbourhood of a hotel. However, based on domain expertise it could be decided to assign higher weights to bars than to restaurants as they might contribute more to the *nightlife factor* in general or to let the users parameterize on their own. Thus, the following table contains the resulting confidence values:

	Confidence	**normalized confidence**
Hotel 1	1.23	1.00
Hotel 2	0.20	0.16
Hotel 3	0.00	0.00

Table 2: Confidences for *nightlife factor*

Again, alternate implementations of this confidence function would be permissible as long as they support the partial ordering of the product base with respect to a concept *c*. Although we have not yet evaluated different designs of the proximity and confidence functions, this base approach was validated as a first proof of concept as outlined in the next section.

4 Evaluation

The proposed approach was applied to the product catalogue of an Austrian tourism destination that includes approximately 9.500 different accommodation service providers that are structured into 16 regions in the winter season and 17 regions during summer season. In addition, we utilized several thousand geo-tagged points-of-interest (POIs) and tourism service providers that are classified according to a three-level taxonomy that extends the Thesaurus on Tourism and Leisure Activities of the World Tourism Organization (World Tourism Organisation, 2002). As a result, 10 different concepts were defined based on this classification scheme (see Table 3).

Nr.	**Concept**	**Description**	***types(c)***	***maxdist***$_c$
1	Art & Culture	For art lovers and the culturally aware.	Ruin or tower, church, archeological site, music festival, exhibition, architectural house, atelier, museum,…	35km
2	Downtown factor	For those who want to stay close to the centre.	Municipal office, city hall, market town	1km
3	Nightlife factor	For those who like to go out and party.	Bar, disco, wine bar, local scene, night spot, live music, pubs	7km
4	Fine food	For aficionados of fine food and the savoir vivre.	Gourmet restaurant, steakhouse, excellent cuisine á la carte, ethnic cuisine, fish specialities,…	15km
5	Golf	For golf players.	Golf course, driving range	15km
6	Shopping	Addresses shopping enthusiasts.	Shopping centre, fashion boutique, major town centre	15km
7	Sights/ Leisure activities	Encompasses all types of sights and leisure activities	Movie theatre, museum, zoo, amusement park, castle, waterfall, gorge, national park, scenic road,…	35km
8	Summer sports and activities	All types of sports and activities carried out in the summer season.	All water sports, biking, trekking, climbing, walking, high trails, sky diving,…	15km
9	Sunbathing, swimming and water sports	For those wanting to relax and enjoy the water.	Open air bath, beach, hot springs, thermal bath, nudist area, boat renting, waterski, sailing, canyoning, …	35 km
10	Winter sports and activities	All types of sports and activities carried out in the winter season.	Skating, skiing, ski tours, cross country skiing, iceclimbing, icehockey, horse slide rides,…	15km

Table 3: Definition of concepts

For each accommodation service provider and each concept a normalized confidence value was computed and plotted on a map, where higher confidence values appear as points in a darker shade of grey. Figure 2 visualizes the concept *sunbathing, swimming and water sports* and allows a first plausibility test of the approach.

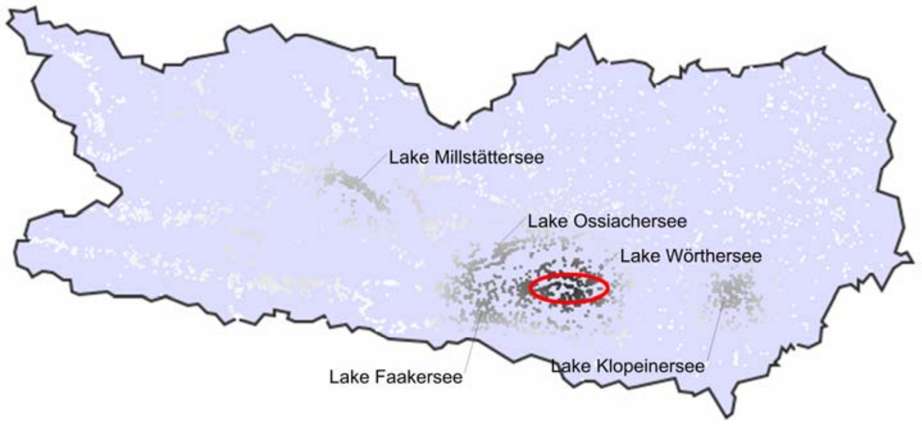

Figure 2: Confidence for the concept s*unbathing, swimming and water sports*

Carinthia is often referred to as Austria's Riviera and, therefore, it is not astounding that the map is well covered with accommodation providers supporting water-related activities (see Figure 2). However, from the map it quickly becomes clear that the most well-known lake of the country, Lake Wörthersee, is the 'centre of gravity' of this concept and the service providers around the lake reach highest confidence values. Furthermore, other lakes with significant water sport and fun infrastructure can be spotted in Figure 2.

For evaluation purposes, regional tourism managers were asked via a questionnaire to name those regions that are in their opinion recommendable to tourists interested in each of the specific concepts. In addition, the system computed the average confidence for all accommodation providers for each concept and for each specific region. Subsequently, we compared the regions recommended for each concept by the experts with those regions that were ranked above average by the system (i.e. the Top 7 regions). If the experts' recommendation was also top-ranked by the system we considered this to be a *hit* and a *failure* otherwise. The Recall of the system was then computed as the ratio between hits and the total number of expert recommendations (Herlocker et al., 2004).

Recall	Concepts
100%	5, 6, 10
at least 66%	1, 3, 7, 9
at least 50%	8

Table 4 Evaluation results

As can be seen from Table 4, 7 concepts had a Recall of at least 66%, one was only above 50% and two concepts could not be validated due to missing answers from the experts. As only three experts have filled out their questionnaires so far[3], alternative evaluation scenarios will have to be developed as part of future study work. Furthermore, the concept annotations for POIs will be reviewed in cases where the system's predictions differed from the expert recommendations, to detect inconsistencies in the data. As most expert recommendations included an additional argument as to why they considered the region to be suitable for tourists with specific interests, the concept definitions need to weight higher those POI types that are considered more influential on the tourist's decision. For instance, a large ski-resort deserves more weight than a local skating ring with respect to the concept *winter sports and activities*.

5 Related Work

In the field of geographic information retrieval, a variety of techniques can be used to extract the geographic position of Web resources, such as from the IP address of the webserver or from the content itself (geo-tagging). For instance, Wang et al. (2005) developed an approach that analyzes the content as well as link information (i.e. web structure). Dickinger et al. (2008) gave an overview of geo-tagging research within the scope of e-tourism. By contrast, the work presented in this paper assumes the existence of already geo-tagged informational resources and analyzes their neighbourhood to extract additional semantic knowledge. Reeve and Han (2005) give an overview on different platforms that automatically extract semantic information.

As a second step the extracted knowledge could be exploited by intelligent systems to support tourists in their decision making process (Werthner, 2003). Knowledge-based recommender systems (Burke, 2000, Felfernig et al., 2006) build on explicit domain

[3] Regional tourism managers that did not return the questionnaire argued that all regions do equally well fit the needs of all tourists and that such weaknesses and strengths profiles for regions are not sensible.

and product knowledge to relate users' specific preferences to suitable product items. Examples of commercial recommender systems that have been successfully fielded in the e-tourism domain include DIETORECS, a European project that researched the requirements for efficient decision support for tourists and proposed several different modes of interaction (Fesenmaier et al., 2003), Trip@dvice (Venturini & Ricci, 2006) and ADVISOR SUITE (Jannach et al., 2007).

In contrast to the aforementioned knowledge-based systems, collaborative filtering-based recommenders do not require such explicit knowledge as the extracted semantic knowledge could also be derived from collected user opinions. Adomavicius et al. (2005) worked on context-aware recommender systems that collect multidimensional user ratings. In addition to giving their opinions on a destination, they also disclose what they were actually looking for. As a result, considering the example of the *nightlife factor*, if most users that wanted to have fun and to go out during their vacations liked a specific region, it is probable that the *nightlife factor* for this region is relatively high. However, in reality such context-aware systems suffer from cold-start problems. Thus, a sufficient number of users must provide their feedback to the system before it can return sensible recommendations. Furthermore, the dimensionality of ratings must be low enough to prevent user confusion.

Future work by the authors will further develop recommendation algorithms to exploit the semantic annotations of geo-tagged objects for personalizing the interaction with maps based on (Zanker, 2008).

6 Conclusions

This paper presented a computation scheme for automatically deriving semantic knowledge for tourism products based on their geographic neighbourhood. A first preliminary evaluation showed that the knowledge gained could be used as an alternative to expert opinions that are usually quite expensive to acquire. This additional information can be exploited for bootstrapping a kowledge-based recommender system which is on our agenda for future work. Furthermore, different learning strategies will be explored in order to automatically improve the concept definitions themselves.

Acknowledgements

This work was partly supported by FFG grant MAPREC nr. 814294 – "Map-based recommendation services for e-tourism applications".

References

Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005), "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach." *IEEE Transactions on Information Systems*, vol. 23(1), pp. 103-145.

Adomavicius, G., and Tuzhilin, A. (2005), "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *IEEE Transactions on Knowledge and Data Engineering*, vol. 17(6), pp. 734-749.

Burke, R. (2000), "Knowledge-based Recommender Systems." *Encyclopedia of Library Information Systems*, vol. 69(2), pp. 180-200.

Chajed, D.R., Francis, L. and Lowe, T.J. (1993), "Contributions of Operations Research to Location Analysis", Location Science, 1(4), pp. 263-287.

Dickinger, A., Scharl, A., Stern, H., Weichselbraun, A., and Wöber, K. (2008), "Acquisition and Relevance of Geotagged Information in Tourism," *Proceedings of the Information and Communication Technologies in Tourism* (ENTER), Innsbruck, pp. 545-555.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., and Zien, J.Y. (2003), "SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation." *Proceedings of the Twelfth International World Wide Web Conference* (WWW), Budapest, Hungary, ACM, pp. 178-186.

Felfernig, A., Friedrich, G., Jannach, D., and Zanker, M. (2006). „An Integrated Environment for the Development of Knowledge-Based Recommender Applications." *International Journal of Electronic Commerce* (IJEC), Special Issue on Recommender Systems, 11(2), pp. 11-34.

Fesenmaier, D.R., Ricci, F., Schaumlechner, E., Wöber, K., and Zanella, C. (2003). "DIETORECS: Travel Advisory for Multiple Decision Styles." *Proceedings of the Information and Communication Technologies in Tourism* (ENTER), Helsinki.

Herlocker, J., Konstan J. A., Terveen, L.G., and Riedl, J. (2004), „Evaluating collaborative recommender systems." *ACM Transactions on Information Systems*, vol. 22(1), pp.5-53.

Jannach, D., Zanker, M., Jessenitschnig, M., and Seidler, O. (2007). Developing a Conversational Travel Advisor with ADVISOR SUITE, *Proceedings of the Information and Communication Technologies in Tourism* (ENTER), Ljubljana.

Miles, G. E., Howes, A. and Davies, A. (2000), "A framework for understanding human factors in web-based electronic commerce," *International Journal on Human-Computer Studies*, vol. 52, pp. 131-161.

Reeve, L., and Han, H. (2005), "Survey of Semantic Annotation Platforms." *Proceedings of the 20th ACM Symposium on Applied Computing*, Santa Fe, New Mexico, ACM, pp. 1634-1638.

Scharl, A., and Tochtermann, K. (2007), "The Geospatial Web - How Geobrowsers, Social Social Software and the Web 2.0 are Shaping the Network Society." London: Springer.

Venturini, A., and Ricci, F. (2006), “Applying trip@advice recommendation technology to www.visiteurope.com.” *In Proceedings of the 17th European Conference on Artificial Intelligence,* Amsterdam, IOS Press, pp. 607-611.

Wang, C., Xie, X., Wang, L., Lu, Y., and Ma, W.-Y. (2005), "Detecting Geographic Locations from Web Resources," *Proceedings of the 13th ACM International Symposium on Advances in Geographic Information Systems* (GIR), Bremen, Germany.

Weisstein, E.W. (2008), “Spherical Trigonometry.” From MathWorld – A Wolfram Web Ressource. Available from http://mathworld.wolfram.com/SphericalTrigonometry.html.

Werthner, H. (2003), “Intelligent Systems in Travel and Tourism.” *Proceeding of the 18th International Joint Conference on Artificial Intelligence* (IJCAI), Acapulco, Mexico.

World Tourism Organisation (2002), “Thesaurus on tourism and leisure activities” of the World Tourism Organization. Available from http://www.unwto.org.

Zanker, M. (2008), “A collaborative constraint-based meta-level recommender.” *Proceedings of the 2nd International ACM Conference on Recommender Systems* (RecSys), Lausanne, Switzerland, pp. 139-146.