Towards Geosocial Recommender Systems

Victor de Graaff
Department of Computer Science
University of Twente
Enschede, The Netherlands
v.degraaff@utwente.nl

Maurice van Keulen
Department of Computer Science
University of Twente
Enschede, The Netherlands
m.vankeulen@utwente.nl

Rolf A. de By
Fac. of Geo-Information Sci. & Earth Observation (ITC)
University of Twente
Enschede, The Netherlands
deby@itc.nl

ABSTRACT

The usage of *social networks sites* (SNSs), such as Facebook, and *geosocial networks* (GSNs), such as Foursquare, has increased tremendously over the past years. The willingness of users to share their current locations and experiences facilitate the creation of geographical recommender systems based on *user generated content* (UGC). This idea has been used to create a substantial amount of *geosocial recommender systems* (GRSs), such as Gogobot, TripIt, and Trippy already, but can be applied to more complex scenarios, such as the recommendation of products with a strong binding to their region, such as real estate or vacation destinations.

This extended form of GRS development requires advanced functionality for information collection (from the web, other social media and sensors), information enrichment (such as data quality assessment and advanced data analysis), and personalized recommendations. The creation of a toolset to cope with these challenges is the goal of this research project, for which the outline is presented in this paper.

Keywords

Geosocial networking, recommender systems, user profiles, geoprofiles.

1. INTRODUCTION

Geographic information systems (GIS) allow people to access geographically referenced information, social media allow people to share their experiences, and recommender systems allow people to benefit from the experiences of others. Many applications are combining these system types into one *geosocial recommender system* (GRS). Rudimentary examples are Gogobot, TripIt, and Trippy. A typical GRS (1) is community-based, (2) focuses on geographic content, and (3) allows for the entry of reviews and browsing of recommendations.

The mentioned GRSs only enable users to read or ask for recommendations of *points of interest* (POIs) in the vicinity of a specified location. However, this idea can be extended to recommend locations to the user (as for example in [1]). These location recommendations are based on the *geoprofiles* of the location, which consist of:

1. the POIs within a reasonable distance;
2. the (un)organized events within a reasonable distance;
3. the socio-demographical backgrounds of the region;
4. the geographic properties of the region.

The concept of geoprofiles enables the creation of a GRS for products with a strong binding to their region, such as real estate or vacation destinations. As running example we use the case of a European holiday home broker. This company likes to provide its customers with extra services while choosing a holiday home, as well as during their stay. To provide these services, they are creating an application that gives each user personalized recommendations, based on the knowledge about this customer's preferences. The reviews and other content on POIs around the cottage will be generated by the former tenants, local businesses and the landlords, as depicted in Figure 1. The gray mini-networks on the outside of the figure illustrate that these people may be connected to other cottages as well. On top of this *user generated content* (UGC), data is extracted from the web to collect more information on the region surrounding the cottages, to fill the geoprofile part of the holiday home's product profile, which is shown in Figure 2.

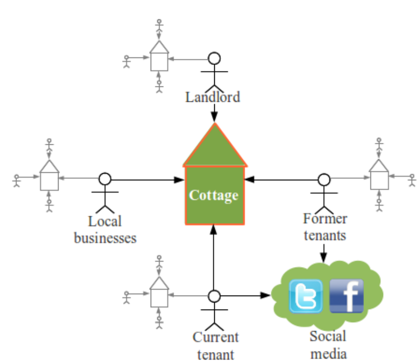

Figure 1: Holiday home domain

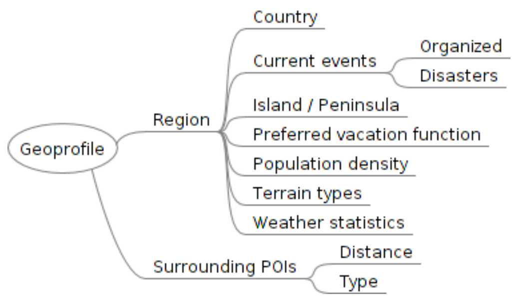

Figure 2: Geoprofile

To collect and increase the quality of this information, GRS development requires advanced functionality for information collection (from the web, other social media and sensors), information enrichment (such as data quality assessment and advanced data analysis), and personalized recommendations. The creation of a toolset to cope with these challenges is the goal of our research project, for which the outline is presented in this paper.

This paper is further structured as follows: an architecture is proposed in Section 2, information collection is discussed in Section 3, information enrichment is discussed in Section 4, the selection of recommendations is discussed in Section 5, related work is discussed in Section 6, and Section 7 contains the conclusion.

2. ARCHITECTURE

The architecture for a GRS consists of three main components, through which the information flows before a recommendation is made: *information collection*, *information enrichment*, and *recommendation selection*, as depicted in Figure 3. The five arrows on the left represent the sources of data: the system's product database, authoritative data, web content, external social media (e.g. Facebook) content, and user generated content (e.g. reviews, GPS trajectories). For each of the phases of the information flow several complementary example approaches are proposed.

For the information collection phase, a flexible and maintainable *data harvesting* approach with a focus on the collection of geographic content is necessary. *Social media connectors* need to be created for the detection of the user's social media accounts and collection of their content. The output of the information collection phase are filled *user profiles* and *product profiles*.

Examples of information enrichment approaches that are relevant for the running example include *user trail analysis* to convert the raw GPS trajectories to semantically meaningful content, *quality assessment* to filter out untrusted or outdated content, *knowledge base accessing* to retrieve more detailed or related content (e.g. geocoding of addresses), and *social graph analysis* to uncover hidden relationships from the *social graph*. The output of the information enrichment phase consists of enriched profiles.

Example approaches for the recommendation selection phase are *filtering* and *profile matching*.

3. INFORMATION COLLECTION

Information is collected from four sources, based on the contents of the fifth source, the product database. The geographic locations and types of content in the product database determine which information is relevant from the other sources. The information types that are most challenging to collect are the *web content* and *social media activities*.

3.1 Data harvesting

Web content can be collected using data harvesting techniques. Online information sources such as the yellow pages have categorized information on many different POI categories. However, this information is often limited to a name, address, phone number and category. More detailed information (such as opening hours, prices, etc.) can be found on domain-specific websites. Since there are many different POI categories, separate data harvesters for each website are time-consuming to build and difficult to maintain. To overcome this problem, there is a need for adaptive wrappers, for example using search result detection algorithms as in [2].

3.2 Social media connectors

Social media connectors are required to collect the data from the social media accounts. People sharing their appreciation for a certain POI, for example by 'liking' its page on Facebook, are reviewing their visit in a minimalistic way. But even though the information in such a review is minimal, this reviewer finds it interesting enough to share this information with his friends. Collecting this information creates a small set of reviews for each new user, without any additional input in the GRS, which can be used to fill the geoprofiles of nearby products, as well as the user's profile.

A more advanced way of collecting reviews from a social media account, is to analyze the user's messages on his or other's bulletin board (such as Facebook's wall). This involves information retrieval techniques, to extract POIs from these messages, such as described in [3], as well as the detection of the corresponding sentiment, similar to [4].

To make use of the social media data, it is compulsory to determine which social network accounts belong to the user. Requiring the user to log in is the straightforward approach and will provide the most detailed information on the user. However, users may be reluctant to do so. For those users, it may be beneficial to detect their accounts automatically, to still use the public content of the account. In a GRS where the name of the user is known (e.g. from previous purchases) this information, combined with his geolocation, can be used to retrieve his social media accounts.

4. INFORMATION ENRICHMENT

Information enrichment approaches can be based on two concepts: analysis of the available data and collection of additional data. Examples of analysis approaches are user trail analysis, quality assessment, and social graph analysis. Collection of additional data is performed by accessing knowledge bases.

4.1 User trail analysis

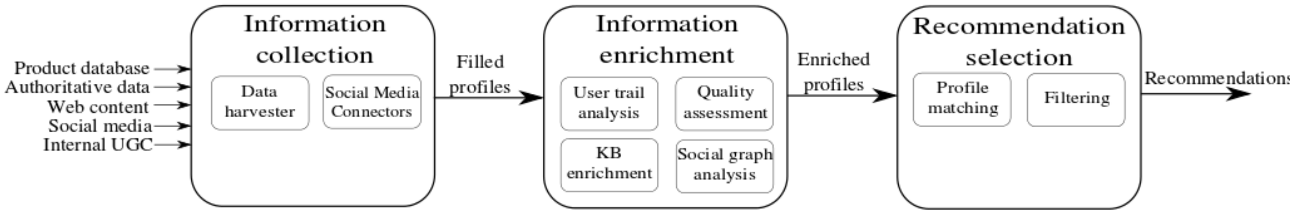

Figure 3: Architecture of a geosocial recommender system

User trail analysis is used to convert the raw GPS trajectories of the internal UGC into semantically meaningful content. This analysis can be done either with or without comparing the coordinates to those of known POIs. Not using POI coordinates is done for example in [5], in which multiple near points on a trail are recognized as a *stay point* (and thus POI), and [6], in which trajectories are simplified based on traveled distance and direction changes.

4.2 Quality assessment

The goal of quality assessment is to filter out irrelevant, imprecise, untrusted or outdated content. UGC is known to be imprecise, and many online reviews lack helpfulness, or even trustworthiness. Coping with the possibility of imprecise content on elsewhere unmentioned items can be done by rating the accuracy of content sources. Such a rating can be based on a comparison of the source's content with other sources, and the type of source (e.g. the website of a restaurant chain is more likely to provide more accurate information on their restaurants than the yellow pages).

4.3 Social graph analysis

Social networks can be represented as a labeled graph, called the *social graph*. Hidden relations can be derived from the social graph using regular graph theory. Using the semantics of the social graph, the uncertainty in this derived graph can be decreased. This is illustrated in Figure 4, where the hometown of the user is derived through the hometown of his brother, rather than from the majority of his online friends.

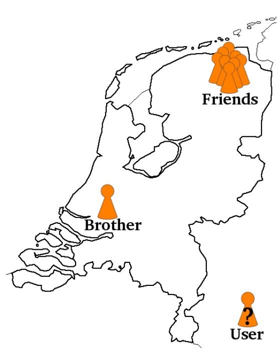

Figure 4: Hometown derivation

4.4 Knowledge base enrichment

The enrichment of the available information using knowledge bases can be used to retrieve more detailed or related content. Typically, information scraped from the web is not rich enough to provide the functionality of a GRS (e.g. geocoded addresses). To solve this problem in a domain-independent way, a possibility needs to be created to detect suitable knowledge bases from the model.

5. RECOMMENDATION SELECTION

The recommendation selection is based on two concepts: *filtering* and *profile matching*. The result of the recommendation selection phase is the intersection of those two individual result sets, as depicted in the Venn diagram in Figure 5. The resulting set of recommendations R for a user u under filtering conditions f is formally given by:

$$R(u, f) = R_f(f) \cap R_{pm}(u),$$

where R_f is the set of recommendations based on filtering, and R_{pm} is the set of recommendations based on profile matching.

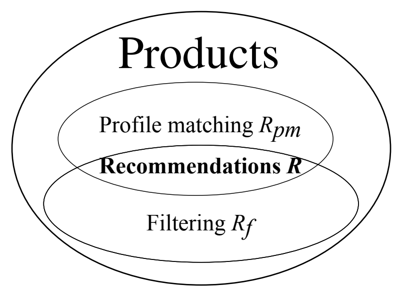

Figure 5: Profile matching and filtering

5.1 Filtering

Filtering is applied to make a selection of the products of the user's interest, based on conditions supplied by the user inputs, or known behavior from the past. In the case of the holiday home broker, a user-filled filter contains for example the start date and end date of the vacation. The complete set of recommendations through filtering R_f is formally given by:

$$R_f(f) = \{\ p \,|\, p \in P \wedge \mathit{filter}(f, p)\ \},$$

where P is the entire set of products, and f are the filtering conditions.

5.2 Profile matching

Profile matching is used to make a selection of the relevant products based on *reviews* by the user or similar users on that product or similar products. The set of relevant recommendations by profile matching R_{pm} for a user u is formally given by:

$$R_{pm}(u) = \sigma_\tau\big(\{\ (p, s) \,|\, p \in P \wedge s = M(u, p)\ \}\big),$$

where σ_τ is a selection function based on the threshold τ on the score s, P again is the entire set of products, and M is the matching function between the user and the product, which is formally given by:

$$M(u,p) = \underset{\substack{u' \in U \\ p' \in P}}{\Lambda}, M_u(u,u') * M_p(p,p') * M_r(u',p'),$$

where U is the set of users, P is again the entire set of products, M_u is the *user matching function*, M_p is the *product matching function*, and M_r is the *match between a user and a product based on a review*. This function is illustrated in Figure 6, where each matching function is depicted as an arrow. Note the difference between the matching functions based on reviews M_r for users u' and u'', and the calculated match M for user u.

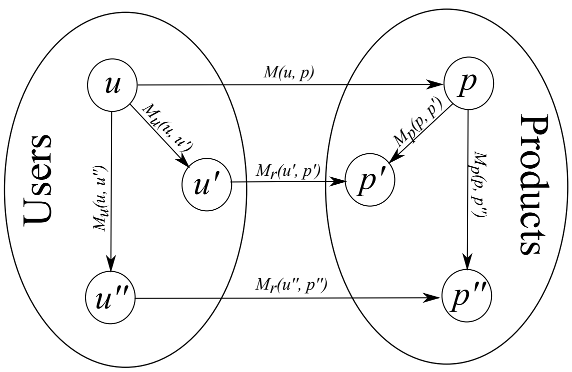

Figure 6: Profile matching function *M(u, p)*

6. RELATED WORK

Geosocial recommender systems are at the intersection of three research areas: geographic information science, online communities, and recommender systems. Much research has been done in all three areas, but the research at this intersection is scarce.

An architecture for a social network with geotagged recommendations is proposed in [7]. This architecture is based on the creation of a new social network, rather than interpreting the activities in other social media. That same approach is taken in [1], where data sparsity became a problem upon evaluation.

Zheng et al. present the architecture for GeoLife 2.0 in [8]. Their architecture has components for user similarity detection, trajectory analysis, and recommendation selection, which can be seen as examples of information enrichment and recommendation selection. User similarity detection is described by this research group in more detail in [9], trajectory analysis in [10] and recommendation selection in [11].

7. CONCLUSION

This paper presents an architecture for a GRS, consisting of three main components: information collection, information enrichment, and recommendation selection. For each of these components, several example approaches have been proposed.

Other research on the creation of a GRS has focused on setting up a new internal social network, whereas this architecture does neither require, nor forbid the user to provide additional reviews or other UGC inside the GRS.

During the research project for which this paper presents the outline, the example approaches are investigated further, using the running example as the validation scenario. The methods and algorithms that follow from this research will form the basis of a toolset for GRS development.

8. REFERENCES

[1] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "Geo-social recommendations," in *ACM Recommender Systems 2011 (RecSys) Workshop on Personalization in Mobile Applications*, 2011.

[2] D. Trieschnigg, K. Tjin-Kam-Jet, and D. Hiemstra, "Ranking XPaths for extracting search result records." Submitted to WSDM'12.

[3] M. B. Habib and M. van Keulen, "Named entity extraction and disambiguation: The reinforcement effect.," in *Proceedings of the 5th International Workshop on Management of Uncertain Data, MUD 2011, Seatle, USA*, vol. WP11-02 of *CTIT Workshop Proceedings Series*, (Enschede), pp. 9–16, August 2011. Centre for Telematics and Information Technology, University of Twente.

[4] C. Strapparava and R. Mihalcea, "Learning to identify emotions in text," in *Proceedings of the 2008 ACM symposium on Applied computing*, pp. 1556–1560, ACM, 2008.

[5] Y. Zheng and X. Xie, "Learning location correlation from GPS trajectories," in *Mobile Data Management (MDM), 2010 Eleventh International Conference on*, pp. 27–32, IEEE, 2010.

[6] Y. Chen, K. Jiang, Y. Zheng, C. Li, and N. Yu, "Trajectory simplification method for location-based social networking services," in *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pp. 33–40, Citeseer, 2009.

[7] C. Chow, J. Bao, and M. Mokbel, "Towards location-based social networking services," in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp. 31–38, ACM, 2010.

[8] Y. Zheng, Y. Chen, X. Xie, and W. Ma, "GeoLife 2.0: A location-based social networking service," in *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*, pp. 357–358, IEEE, 2009.

[9] X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Finding similar users using category-based location history," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 442–445, ACM, 2010.

[10] Y. Zheng and X. Xie, "Learning travel recommendations from user-generated GPS traces," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 1, p. 2, 2011.

[11] Y. Zheng, L. Zhang, Z. Ma, X. Xie, and W. Ma, "Recommending friends and locations based on individual location history," *ACM Transactions on the Web (TWEB)*, vol. 5, no. 1, p. 5, 2011.