ABSTRACT

Databases of places have become increasingly popular to identify places of a given type that are close to a user-specified location. As it is important for these systems to use an up-to-date database with a broad coverage, there is a need for techniques that are capable of expanding place databases in an automated way. In this paper the authors discuss how geographically annotated information obtained from social media can be used to discover new places. In particular, the authors first determine potential places of interest by clustering the locations where Flickr photos have been taken. The tags from the Flickr photos and the terms of the Twitter messages posted in the vicinity of the obtained candidate places of interest are then used to rank them based on the likelihood that they belong to a given type. For several place types, their methodology finds places that are not yet contained in the databases used by Foursquare, Google, LinkedGeoData and Geonames. Furthermore, the authors' experimental results show that the proposed method can successfully identify errors in existing place databases such as Foursquare.

Keywords: Databases, Discovering, Geographic Information Retrieval, Places of Interest, Social Media

INTRODUCTION

Berners-Lee’s vision of the Semantic Web (Berners-Lee, 2001) has become increasingly popular in the last few years. The World Wide Web would evolve to a highly interconnected network of data that could be easily accessed and understood by machines. Applications could for instance use the Semantic Web to construct customized answers to a particular question. In such applications the user is no longer required to search for information or pore through results. A question can e.g. be ‘What are locations of the restaurants in London?’ To answer this question, a structured dataset has to be available containing places located in London (entities), associated with their location and semantic type (properties).

DOI: 10.4018/ijswis.2013070105
However, a lot of information on the Web is still unstructured or only semi-structured. Therefore, there is a need for automated methods to extend structured datasets using existing Web data. Several methods of this form have been proposed, e.g. YAGO2 (Hoffart, 2013) and BabelNet (Navigli, 2010) are knowledge bases that are constructed using Wikipedia and Wordnet. Other research focuses on establishing structured datasets containing information of a specific type. For instance, LinkedGeoData (Auer, 2009) is a dataset of places constructed using OpenStreetMap, an application in which users can submit geographical data such as place semantics.

In this paper, we will focus on improving existing databases of places. More precisely, we will add new places and discover likely errors using data from the Web. Social media data is particularly promising in this respect, due to the large amounts of geographically annotated data produced by these media. For example, about 1.5% of all Twitter posts (i.e. tweets) are annotated with geographical coordinates (Murdock, 2011). In addition, there are currently more than 190 million geotagged Flickr photos (Flickr, 2013). This data has been used to e.g. automatically detect events (Rattenbury, 2007; Sakaki, 2010; Lee, 2011), to find popular places (Crandall, 2009; Van Canneyt, 2011) and tourist routes (Choudhury, 2010; Jain, 2010).

The main focus of this paper is on how geographically annotated information obtained from social media can be used to discover new places of a given type such as ‘hotel’ or ‘school’ to extend semantic databases of places. Our hypothesis is that the type of a place can be derived from the tags of the Flickr photos and the terms of the Twitter posts associated with locations in the vicinity of the place. For example, if photos around a particular location contain tags such as ‘food’, ‘dinner’ and ‘eating’, this strongly suggests that there is a restaurant at that location. In our previous work (Van Canneyt, 2012a), we have provided evidence for the validity of this hypothesis: given the location of various places of interest (POIs), we addressed the task of identifying those POIs that are most likely to be of a particular type. Our main conclusion was that Flickr tags are a rich source of information for deciding on the type of a place. Using Twitter terms further improved the results although this improvement was more limited. We also considered the correlation between the type of the POIs and the types of the places in the vicinity to categories the POIs. However, this additional information led to a minimal improvement of the performance of our methodology, and in this paper we are mainly interested in the use of social media by itself to improve databases of places. Therefore, we do not consider such correlations here. In (Van Canneyt, 2012b) we considered the more challenging problem of finding locations where places of particular types can be found, without providing a list of candidate locations. Instead, we used a simple grid overlay to find candidate locations and compared the results against existing databases of places. This qualitative analysis demonstrated the potential of the proposed method to find POIs in London that are not yet contained in FourSquare, Google Places, Geonames and LinkedGeoData. Encouraged by these initial results, we improve the proposed methodology in this paper and present a more detailed experimental evaluation. First, the Support Vector Machine classifier used in (Van Canneyt, 2012a; Van Canneyt, 2012b) is replaced by a language modeling approach, which improves the results significantly. Second, we analyze the behavior of different feature selection techniques. We conclude that for the Flickr data correlation coefficient feature selection (Ng, 1997) performs significantly better than $\chi^2$ feature selection. The performance of the proposed methodology can be further improved when names of cities and countries are removed from the considered features. Finally, we perform a large-scale evaluation on 88 different cities, where we examine the results for London in more detail. Based on this evaluation, we can conclude that our approach is able to extend and validate data sets of places. In particular, our method is able to detect new places of a particular type, even when the locations of