Harvesting large corpora for generating place graphs

Junchul Kim, Maria Vasardani, and Stephan Winter

Department of Infrastructure Engineering, University of Melbourne, Parkville, Victoria 3010, Australia
junchulk@student.unimelb.edu.au,
{maria.vasardani,winter}@unimelb.edu.au

Abstract. This paper proposes a novel approach of harvesting large corpora from the Web for generating place graphs. The approach consists of three main phases: an efficient strategy for harvesting relevant web pages that include place descriptions related to a particular environment, extracting triplets from the descriptions, and generating place graphs from these triplets. The paper also discusses the characteristics of the generated place graphs, and identifies further challenges given the well-known flexibility of natural language.

Keywords: place descriptions, place graphs, web harvesting, mining

1 Introduction

Place descriptions are individual natural language (NL) descriptions of places, which are used as a common way to convey spatial information between people, based on their perception or memory of spatial features in an environment and their relations. The descriptions provide a rich source of human spatial knowledge that is complementary to the knowledge found in current space-based GIS.

Instead of tedious crowd-sourcing of place descriptions [14], this paper investigates whether they also can be harvested from Web platforms such as Wikipedia, business websites, blogs, and social networks services. Throughout the analysis of text-based information, human spatial knowledge about places can be extracted by NL processing in the form of triplets of a locatum, a relatum, and a spatial relationship between them. These triplets can be transformed and stored in a graph structure which is useful for managing place graphs, where nodes represent locata or relata, and edges their relations. These place graphs would provide massive support towards place-based GIS, i.e., for diverse applications such as helping navigation systems, supporting human wayfinding process, and automatic landmark identification.

For this purpose, this paper proposes a novel approach of harvesting relevant web pages that include place descriptions related to a particular environment, extracting triplets from the descriptions, and generating place graphs from these triplets.

CESIP 2015: Cognitive engineering for spatial information processes: From user interfaces to model-driven design. Workshop at COSIT 2015, October 12, 2015. Santa Fe, NM, USA

2 Related work

This section outlines the relevant papers according to major components of the approach, namely Web harvesting of place descriptions, spatial information extracting from place descriptions, place graphs construction.

Web harvesting in order to collect spatial information has been studied before. For example, a method for extracting landmarks from web documents was proposed [15], and a system architecture for extracting geographic information from large collections of web documents was introduced [12]. Also examples of data collections and mining from social networking sites, WikiMapia (www.wikimapia.com), Wikipedia (www.wikipedia.com) or GeoNames (www.geonames.org) through their Application Programming Interfaces were already presented, largely called geographic information retrieval [13]. In addition, existing crowd-sourcing methods from the Web for disambiguating place names were related to Web harvesting [17].

Linguistic researchers have been working on parsers that can process unrestricted language input for spatial content. For example, a conceptual framework had been presented for the tagging and parsing of text to extract movement information [5]. Other work has studied the semantics of locative expressions of various languages [10, 18]. Based on the latter, a method was proposed to assign reference objects, locata and spatial relations roles to linguistic terms [9], and a parser was presented extracting spatial features and qualitative spatial relations between them from natural language descriptions [6, 11]. This parser extracts the triplets that we use in this research.

Place graphs are a special type of a property graph [4], as introduced as a Spatial Property Graph (SPG) [16, 7]. In this research, the concept of a place graph is used for representing and integrating spatial information extracted from place descriptions. The set of locata and the set of relata are the vertices of the place graph, while the spatial relations are the labeled edges, directed from locatum to relatum. In our research, place graphs are stored in a graph database since graph databases are commonly used in modelling general or spatial knowledge [2, 3].

3 Approach

The approach consists of three steps: Web harvesting of documents containing place descriptions of a target environment, extracting the spatial information, and constructing the place graphs.

Web harvesting of place descriptions of a target environment. A target environment represents a specific region defined by a boundary, such as the boundary of a city. Relevant Web pages are harvested by using place search APIs and Web crawling tools. In general, these services provide the spatial search functions as well as detailed information of found places, especially further urls.

In our experiment, we chose the Wikimapia place APIs for selecting start urls. The main reason is that Wikimapia, as a user-generated place database,

would contain place descriptions shared by many people. Another reason is that it provides directly links to Wikipedia documents linked to the searched places, which again are plain language and include abundant place descriptions shared by many users. After selecting start urls, the web harvesting starts crawling web sites linked to the start urls and scrapying their contents. The following data sources were discovered as the harvested websites:

- *WikiMapia sites*: User-generated place information including place descriptions
- *Wikipedia sites*: User-generated contents including place descriptions
- *Business sites or official sites*: Place descriptions related to locations of companies, shops, restaurants, and so on
- *Tourist sites*: Popular places and attractions in travel guides
- *Blogs*: Place descriptions focused on individual interests

Extracting the spatial information. After HTML parsing to extract texts from the collected sites, the above mentioned natural language parser, trained on these forms of texts, is applied in order to extract the spatial information in form of triplets (locatum, relation, relatum). Each site produces one set of triplets.

Due to the flexibility of natural language and the limitations of current natural language processing, the extracted triplets include also non-locative expressions. Therefore, the triplets have further been filtered by a set of rules based on the more frequent exceptions found in this experiment, such as:

- Temporal expressions such as 'in 1999', and 'in May' in a relatum.
- Event or game names such as 'Melbourne Cup' in a locatum or relatum
- Composite locative expressions such as 'Grattan Street and Swanston Street'
- Personal pronouns such as 'I', 'You', and 'They' in a locatum or relatum

Constructing the place graphs. In the third step, the individual place graphs generated from the triplets are merged into a composite place graph. Corresponding nodes among graphs are identified via similarity measures, using a combination of three types of similarity scores: typographic, linguistic, and spatial similarities. The typographic similarity, based on existing string matching algorithms, is used to measure the similarity of node names. The linguistic similarity is calculated by a WordNet-based similarity of node names in terms of the words' meanings. The spatial similarity between nodes represents the level of spatio-relational similarity of the associated neighbours having similar spatial relations. An overall similarity score with equal weights of these similarities controls the matching of corresponding nodes in the graph amalgamation process (the highest precision 0.82), which is fully documented [8]. The (potentially large) composite place graph reflects the collective human spatial knowledge in the target environment, maintaining also place name synonyms as they were identified.

4 Experiments and observed findings

An experiment was conducted to show the effectiveness of this approach for harvesting large corpora as well as to discuss the characteristics of the generated

place graphs. Figure 1 shows the workflow of the implemented approach to generate place graphs from large corpora. We use the Wikimapia place search APIs for Web crawling and scraping, and the Scrapy tool[1] for its flexibility as a web crawling tool, including the extraction of text from HTML/XML sources. Two target environments were chosen for comparison: Melbourne, Victoria, Australia, and Santa Fe, New Mexico, USA.

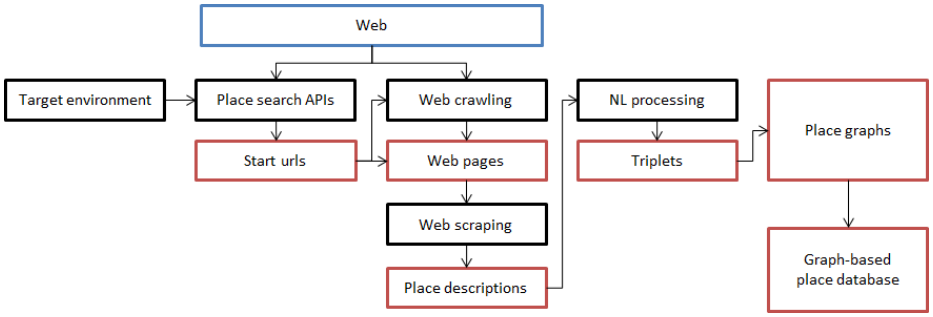

Fig. 1. Workflow of generating place graphs

For Melbourne, first the start urls of places were searched by using the WikiMapia place search APIs. Figure 2a shows the locations of these start urls on a map. From the start urls, the number of web pages extracted by Web crawling is 19,125, and the number of texts extracted by Web scraping is 16,527. From these texts, 2,911 sets of triplets have been identified by NL processing, i.e., 17.6% (=2,911/16,527) of the texts contained a recognized place description. Figure 2b shows the composite place graph for Melbourne generated from the descriptions harvested from the Web. This graph contains 2,736 unique nodes, representing the unique place names identified, and 3,221 edges, representing the identified spatial relations between them.

In the same way place descriptions for Santa Fe were extracted. 590 texts were found, and 114 sets of triplets were extracted from these texts. Figure 2c shows the start urls and 2d the composite place graph for Santa Fe. Despite of the small area of Santa Fe (96.9 km^2, compared to Melbourne with its 9,990 km^2), a total of 238 unique place names and 218 spatial relations between them were identified.

In the composite graph of two target areas, 69 unique spatial relations were found. Directional relations have a large portion (55%) in comparison to topological relations (35%) and the other qualitative spatial relations such as order.

The effectiveness of the Web harvesting process was evaluated by computing the number of the texts in which places correctly belong to their target area, and accuracy in percent. For non-georeferenced text geoparsing, we applied an open source library, CLAVIN [1] to identify locations from the texts. The locations were checked whether they relate to their target areas, Melbourne and Santa

[1] http://scrapy.org

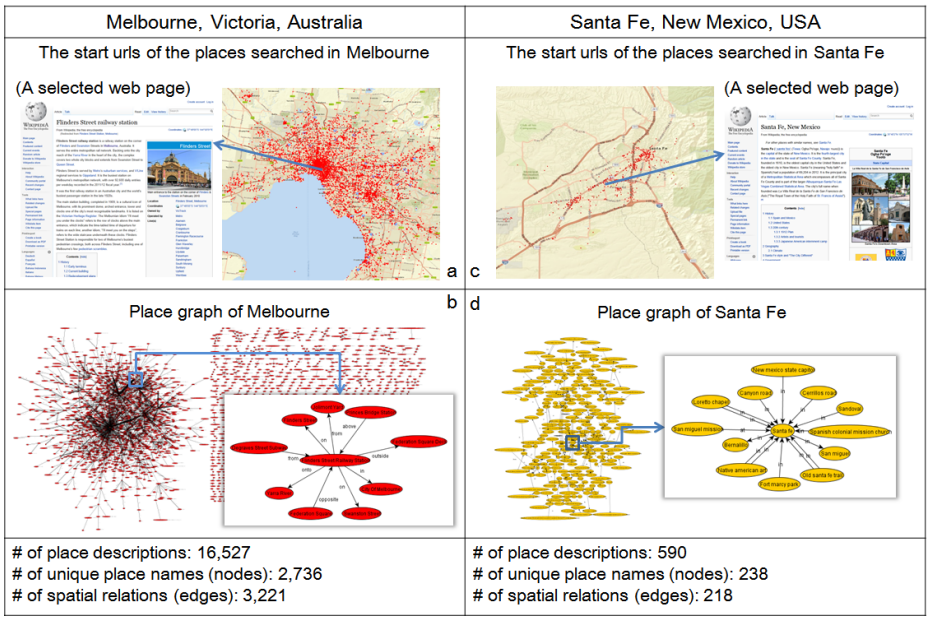

Fig. 2. Place graphs for Melbourne and Santa Fe

Fe. Note that it only considers automatically extracted geospatial entities from CLAVIN (75% accuracy for entity resolution). As a result, the total ratio of relevant texts is 78.5% since 2,373 texts correctly related to their target area from 3,025 texts (2,911 texts for Melbourne and 114 for Santa Fe).

5 Conclusions

This paper introduces a new approach for generating place graphs from large corpora, which provides an efficient way of harvesting place descriptions for a target environment from the Web, and a strategy of generating place graphs from spatial information extracted from the descriptions via NL processing. The system implemented based on the proposed approach was tested for two different target environments. As a result of the experiment, the main achievement is that the approach harvested place descriptions effectively and generated place graphs based on human spatial knowledge from them. The place graphs also represented a feasible and reliable structure to store and manage abundant place names and their spatial relations. Future work includes studying further how to extract salient objects as cognitive landmarks from these composite place graphs.

References

1. Berico Technologies. CLAVIN: Cartographic Location And Vicinity INdexer. `http://clavin.bericotechnologies.com/`, 2012–2013

2. Angles, R., Gutierrez, C.: Survey of graph database models. ACM Comput. Surv. 40(1), 1–39 (2008)
3. Basiri, A., Amirian, P., Winstanley, A.: Use of graph databases in tourist navigation application. In: Murgante, B., Misra, S., Rocha, A., Torre, C., Rocha, J., Falcao, M., Taniar, D., Apduhan, B., Gervasi, O. (eds.) Computational Science and Its Applications ICCSA 2014, Lecture Notes in Computer Science, vol. 8583, pp. 663–677. Springer International Publishing (2014)
4. Hidders, J.: Typing graph-manipulation operations. In: Calvanese, D., Lenzerini, M., Motwani, R. (eds.) Database Theory – ICDT 2003, Lecture Notes in Computer Science, vol. 2572, pp. 394–409. Springer, Berlin (2002)
5. Hornsby, K.S., Naicong, L.: Conceptual framework for modeling dynamic paths from natural language expressions. Transactions in GIS 13(s1), 27–45 (2009)
6. Khan, A., Vasardani, M., Winter, S.: Extracting spatial information from place descriptions. In: Proceedings of the First ACM SIGSPATIAL International Workshop on Computational Models of Place. pp. 62–69. COMP '13, New York, NY (2013)
7. Kim, J., Vasardani, M., Winter, S.: From descriptions to depictions: A dynamic sketch map drawing strategy. Spatial Cognition & Computation, accepted (2015)
8. Kim, J., Vasardani, M., Winter, S.: Similarity matching for integrating spatial information extracted from place descriptions. submitted (2015)
9. Kordjamshidi, P., Van Otterlo, M., Moens, M.F.: From language towards formal spatial calculi. In: Workshop on Computational Models of Spatial Language Interpretation (CoSLI 2010, at Spatial Cognition 2010) (2010)
10. Kracht, M.: On the semantics of locatives. Linguistics and Philosophy 25(2), 157–232 (2002)
11. Liu, F., Vasardani, M., Baldwin, T.: Automatic identification of locative expressions from social media text: A comparative analysis. In: Proceedings of the 4th International Workshop on Location and the Web. pp. 9–16. ACM (2014)
12. Mário, J.S., Bruno, M., Marcirio, C., Ana Paula, A., Nuno, C.: Adding geographic scopes to web resources. Computers, Environment and Urban Systems 30(4), 378–399 (2006)
13. Purves, R.: Methods, examples and pitfalls in the exploitation of the geospatial web. In: Hesse-Biber, S. (ed.) The Handbook of Emergent Technologies in Social Research, pp. 592–622. Oxford University Press, New York (2011)
14. Richter, K.F., Winter, S.: Harvesting user-generated content for semantic spatial information: The case of landmarks in OpenStreetMap. In: Hock, B. (ed.) Proceedings of the Surveying and Spatial Sciences Biennial Conference 2011. pp. 75–86. Surveying and Spatial Sciences Institute
15. Tezuka, T., Tanaka, K.: Landmark extraction: A web mining approach. In: Cohn, A., Mark, D. (eds.) Spatial Information Theory. Lecture Notes in Computer Science, vol. 3693, pp. 379–396. Springer, Berlin (2005)
16. Vasardani, M., Timpf, S., Winter, S., Tomko, M.: From descriptions to depictions: A conceptual framework. In: Tenbrink, T., Stell, J., Galton, A., Wood, Z. (eds.) Spatial Information Theory. Lecture Notes in Computer Science, vol. 8116, pp. 299–319. Springer, Berlin (2013)
17. Vasardani, M., Winter, S., Richter, K.: Locating place names from place descriptions. International Journal of Geographical Information Science 27(12), 2509–2532 (2013)
18. Zlatev, J.: Spatial semantics. In: Cuyckens, H., Geeraerts, D. (eds.) The Oxford Handbook of Cognitive Linguistics, pp. 318–350. Handbook of Cognitive Linguistics, Oxford University Press, Oxford (2007)