Inferring Relevant Gazetteer Instances to a Placename

Yunhui Wu and Stephan Winter

Department of Geomatics, The University of Melbourne, VIC 3010, Australia
Telephone: +61 3 8344 6881
Fax: +61 3 9347 2916
Email: y.wu21@pgrad.unimelb.edu.au, winter@unimelb.edu.au

1. Introduction

In navigation systems, gazetteer services are identifying and geocoding user named places. Gazetteers are dictionaries of geographic names describing location instances. There are three essential components of gazetteer instances: a geographic name, a footprint representing its spatial location, and a type of feature it labels (Hill 2000). Currently, named places are identified by matching an input string to the geographic names of a gazetteer. For example, when a user requests directions to “Royal Melbourne Hospital”, the gazetteer will search for exact or partial matches for this string. However the state-of-the-art string-matching method is insufficient to infer the relevant gazetteer instances if no exact match exists. Firstly users may name a place variously. For example, “RMH”, the abbreviation of “Royal Melbourne Hospital”, is also used by users. Synonyms, vernacular placenames and names in other languages are a known problem for geographic search. Also, a placename does not always unambiguously identify a feature. For instance, “Melbourne” may indicate the city in Florida, USA, or the one in Victoria, Australia. Furthermore, the semantics of placenames are disregarded in this search.

The aim of this paper is to find the most relevant instances to a placename beyond string-matching. To solve this problem, we suggest an approach considering similarity between gazetteer instances and placenames from three aspects: string, ontological and spatial similarities. The following sections discuss the semantics behind a placename and how the similarity approach is designed. A demonstration illustrates how this approach suggests the most relevant instances in a gazetteer.

2. Semantics of a Placename

This section investigates what information can be obtained from a placename. A placename provides a possible geographic name of a feature. Also, it may indicate a type of the feature. The example “Royal Melbourne Hospital” points out that the type is *hospital*. The feature type is frequently more robust than the placename: an identified type helps to resolve variations in notation. The type defines a range of relevant gazetteer instances: e.g., hospital instances more likely include “Royal Melbourne Hospital” or similar. However, feature types may be also named/coded differently in gazetteers. For instance, the type *hospital* (in the placename) may be categorized into *medical building* in the gazetteer. Hence, ontology of a type rather than its string needs to be considered.

In the above example, a type is given explicitly. In other cases, types are implicit, such as “The Royal Melbourne” (hospital). To obtain type information, it is assumed that only lexical nouns can be types. A lexical database can be used to check whether a phrase, by words in their order in the placename, is defined as a lexical noun and thus could indicate a type. Figure 1 demonstrates the process of detecting a type from a placename.

Phrase	Defined by a lexical database?	Possible Feature Type
“Royal Melbourne Hospital”	No	-
“Melbourne Hospital”	No	-
“Hospital”	Yes	hospital

Figure 1. Detecting a feature type in a placename “Royal Melbourne Hospital”

3. Similarity

This section introduces the similarity measurements used to infer most relevant gazetteer instances to placenames. For spatial-scene similarity queries, Nedas and Egenhofer (2008) suggest relaxation of spatial query constraints on spatial objects and spatial relations. In this paper, queries are relaxed by measuring the similarities of feature type ontologies and spatial locations between wayfinders and targeted features. The following sections discuss how similarity can be measured from three aspects: strings, feature types and spatial locations. The larger the similarity, the more relevant the result according to the given aspect. Results are then ranked by a weighting procedure.

3.1 String Similarity

String similarity is a criterion on pure string matching between placenames and geographic names, conventionally applied in gazetteers, navigation systems and the like. The Levenshtein distance $dist_L(s,t)$ (Levenshtein 1966) is a measurement of the similarity between two strings by the number of deletions, insertions or substitutions required to transform the source string *s* (here: a placename) into the target string *t* (here: a geographic name). The greater the Levenshtein distance in proportion to the length of *t*, the less similar the string *s*:

$$Sim_s(s,t) = 1 - \frac{dist_L(s,t)}{len(t)} \tag{1}$$

3.2 Ontological Similarity

From an ontological perspective, Janowicz et al. (2009) state that similarity should not be applied to compare sub- and super-types, but should combine subsumption reasoning to fit user’s requirements. However, those at sub/super level of a type could be counted as less similar than those at the level of this type. To define the ontology of feature types, WordNet (Fellbaum 1998) can be used, with its taxonomy including hypernym-hyponym and part-whole relations. Among measurements based on WordNet, the Jiang-Conrath distance (Jiang and Conrath 1997) is selected here for its outstanding performance (Jurafsky and Martin 2008). The ontological similarity of the feature type is then defined as Equation 2, with t_s the type of placename *s*, t_t the type of the geographic name *t*, and $dist_{JC}(t_s,t_t)$ the Jiang-Conrath distance of these two types.

$$Sim_T(t_s,t_t) = \frac{1}{dist_{JC}(t_s,t_t)+1} \tag{2}$$

3.3 Spatial Similarity

Spatial similarity can be rated as spatial relevance. Tomko and Winter (2009) argue that the more prominent a feature, the less cognitive effort is required for people to recall it. In other words, wayfinders would mention more prominent features together with a placename as they expect that informers would more likely recognize them. They also state that the distance between people's location and targeted features increases the cognitive effort, because the greater search radius the more features of a type have to be considered. For example, in Melbourne CBD when a wayfinder looks for "the market", he/she probably indicates the Queen Victoria Market, because this market is not only a famous local icon meaning prominence, but also the closest market to Melbourne CBD. Assuming the current location of a wayfinder is known, the spatial similarity of a placename with a gazetteer instance can be defined by Equation 3, with l_s the current location of the wayfinder, l_t the location of the targeted feature, $dist(l_s, l_t)$ the distance in a part-of hierarchy between these two locations, and $prom(l_t)$ the prominence of the feature. Prominence measurements were suggested for example by Raubal and Winter (2002), Claramunt and Winter (2007) and Caduff and Timpf (2008).

$$Sim_G(l_s, l_t) = \frac{prom(l_t)}{dist(l_s, l_t) + 1} \qquad (3)$$

3.4 Relevance Procedure

Query results are ranked by the following procedure: Firstly gazetteer instances are ranked by string similarity. If multiple exact matches exist, spatial similarity is measured to rank the results; if no exact match exists ontological similarity is measured to generate the potential result set before spatial similarity is used for further ranking. If a single exact string match exists, no further action is required. The procedure is illustrated in Figure 2.

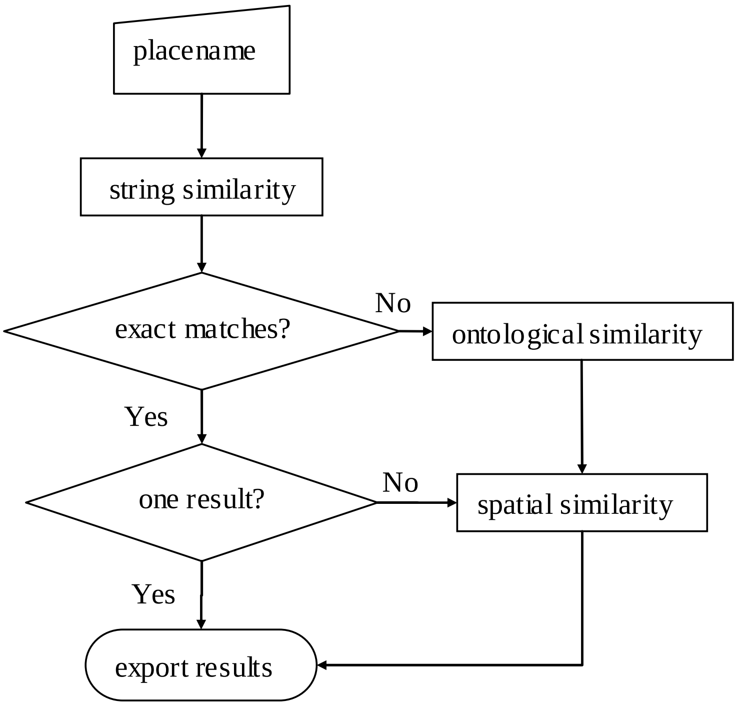


Figure 2. Gazetteer instances ranking procedure.

4. Demonstration

Beyond string-matching of placenames, this section demonstrates how the similarity approach is still able to infer relevant results even if the string of a placename is not exactly matched in a gazetteer. A synthetic test gazetteer is shown in Figure 3. The locations of features and their part-of relations are represented in Figure 4, which are used to calculate their hierarchical distances. The current location of the wayfinder is close to the *Seven Eleven* in *North Melbourne*. "Royal Melbourne Hospital" is the requested placename. The similarity approach works by the following steps:

1. A query is run to find gazetteer instances having exact "Royal Melbourne Hospital" in attribute *Geographic Name*. Results are stored in set *A*. The set *A* is here {}.
2. WordNet is used to detect "Hospital" as the feature type.
3. A query is run to find gazetteer instances under type “Hospital”. Results are stored in set *B*. Set *B* is here {“Royal Melbourne, hospital, 0.8”, “St Vincents Private, hospital, 0.7”}.
4. A query is run on attribute *Feature Type*. Instances are added to set *B* by constraining their feature type ontological similarity to “hospital” larger than 0.8 (an arbitrary threshold). *Set B* is here {“Royal Melbourne, hospital, 0.8”, “St Vincents Private, hospital, 0.7”, “Borotto, building, 0.2”}.
5. If set *A* is NOT empty, a query is run to find gazetteer instances under various types from set *A*. Results are stored in set *B*.
6. Based on the current location, outside the *Seven Eleven*, spatial similarity is calculated on the result set made of set *B*. The results (Figure 5) are then ranked by the relevance procedure.

ID	Geographic Name	Feature Type	Prominence
1	Baretto Café	coffee shop	0.3
2	George AA	coffee shop	0.2
3	China Bar BB	restaurant	0.3
4	Flinders Station	station	1
5	Borotto	building	0.2
6	Pizza Hut	restaurant	0.4
7	MCG	station	0
8	Botanical Garden	garden	0.8
9	Royal Melbourne	hospital	0.8
10	St Vincents Private	hospital	0.7
11	Melbourne Medical	medical center	0.4
12	Errol Street	road	0.6
13	North Melbourne	suburb	0.8
14	Flemington Road	road	0.6
15	Seven Eleven	shop	0.2
16	Gidon Kremer concert	event	0.2
17	Carlton	suburb	0.8
18	Melbourne	suburb	1

Figure 3. Gazetteer instances

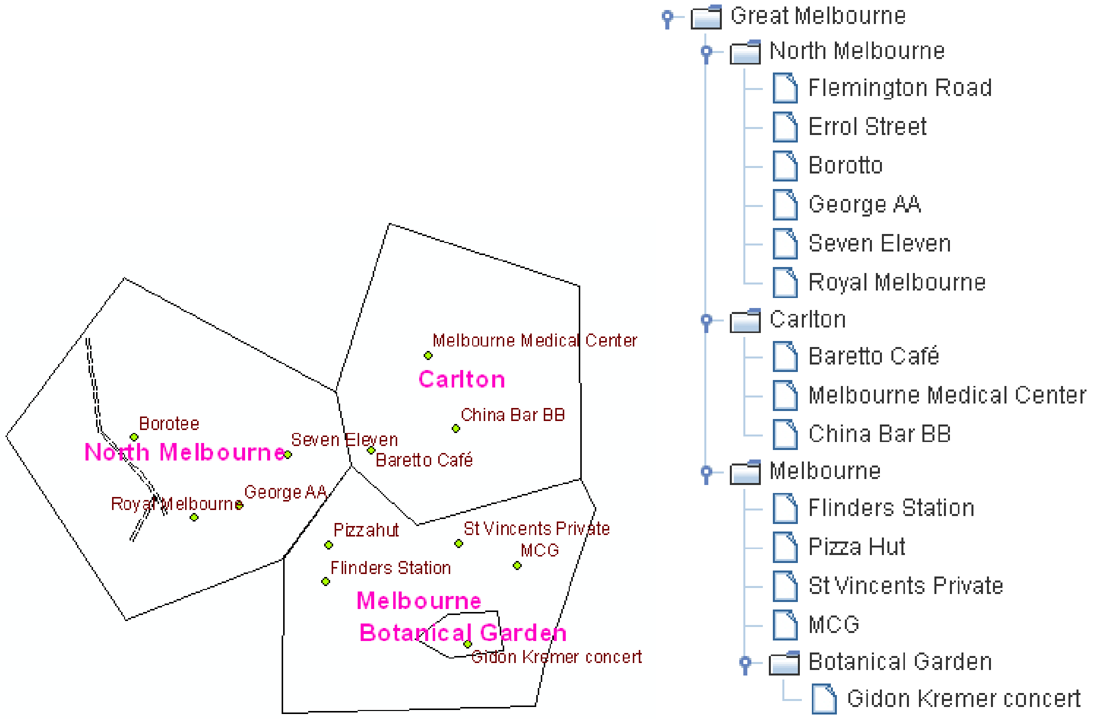
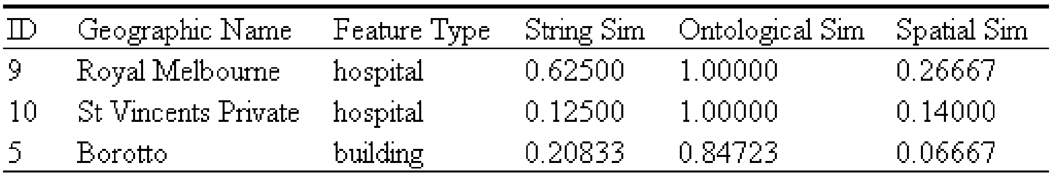


Figure 4. Gazetteer instances and their topological relations.



ID	Geographic Name	Feature Type	String Sim	Ontological Sim	Spatial Sim
9	Royal Melbourne	hospital	0.62500	1.00000	0.26667
10	St Vincents Private	hospital	0.12500	1.00000	0.14000
5	Borotto	building	0.20833	0.84723	0.06667

Figure 5. Results ranked by the relevance procedure

From Figure 5, the most relevant gazetteer instance is “Royal Melbourne”, which has the most similar geographic name, exact type match and highest spatial similarity.

5. Conclusion

Beyond the state-of-the-art string-matching approach, this paper suggests using semantics of placenames and spatial relevance based on wayfinders’ locations and targeted features to refine the inferred results. This approach will improve the efficiency of the inference process.

Acknowledgements

This work is supported under the Australian Research Council's Discovery Projects funding scheme (no. 0878119).

References

Caduff, D.; Timpf, S., 2008: On the assessment of landmark salience for human navigation. Cognitive Processing, 9 (4): 249-267.

Claramunt, C.; Winter, S., 2007: Structural salience of elements of the city. Environment and Planning B: Planning and Design, 34: 1030-1050.

Fellbaum, C., 1998: WordNet: An Electronic Lexical Database. The MIT Press, Cambridge MA, 445 pp.

Hill, L., 2000: Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In: Borbinha, J.e.; Baker, T. (Eds.), Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science, 1923. Springer, Berlin, pp. 280-290.

Janowicz, K.; Schwarz, M.; Wilkes, M., 2009: Implementation and Evaluation of a Semantics-based User Interface for Web Gazetteers, Visual Interfaces to the Social and the Semantic Web (VISSW 2009) Workshop in conjunction with the International Conference on Intelligent User Interfaces (IUI 2009), Sanibel Island, Florida.

Jiang, J.J.; Conrath, D.W., 1997: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy, Proceedings of ROCLING X, Taiwan, pp. 15.

Jurafsky, D.; Martin, J.H., 2008: Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Pearson Prentice Hall, Upper Saddle River, New Jersey, 988 pp.

Levenshtein, V.I., 1966: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10 (8): 707-710.

Nedas, K.A.; Egenhofer, M.J., 2008: Spatial-Scene Similarity Queries. Transactions in GIS, 12 (6): 661-681.

Raubal, M.; Winter, S., 2002: Enriching Wayfinding Instructions with Local Landmarks. In: Egenhofer, M.; David, M. (Eds.), Geographic Information Science. Lecture Notes in Computer Science, 2478. Springer, Berlin, pp. 243-259.

Tomko, M.; Winter, S., 2009: Pragmatic Construction of Destination Descriptions for Urban Environments. Spatial Cognition and Computation, 9 (1): 1-29.