Contextualized Search by Nearness

Marc Novel

Swiss Federal Institute for Forest, Snow and Landscape Research WSL
marc.novel@wsl.ch

Abstract The spatial expression "near" describes proximity and is frequently used for web search such as "gas stations near the old market". What is "near" depends on the context and I investigate how a context dependent model for "near" can be formulated. For doing so, I investigate the following questions: (i) what is the relevant contextual information for "near"? (ii) how does the identified information influence the interpretation of near? To answer these questions, my research consists of identifying the context factors and then learn from data how these context factors have an effect on a qualitative or quantitative distance measure of "near", which enables me to formulate a contextualized model for "near"

1 Background

As social beings we do not only rely on finding objects in the world, but we also are able to describe and communicate via natural language (NL), where these objects are located. For this reason, NL is full of spatial descriptions and one of its essential building blocks are prepositions such as "left of", "right of", "above"[15]. The prepositions expressing proximity for spatial relations are as follows: "near", "nearby" or "close". These prepositions, which describe proximity between two objects, are sometimes also called qualitative spatial distance relations.

The preposition "near" is used pervasively in many tasks of our daily life. For instance in **route descriptions**:*"Head to the gravel FR 328 at the south end of Moqui Lodge, near the gas station"*[18, p.68], **place descriptions**: *"There is a Starbucks in the Marriott, and a small Mexican restaurant near the gas station opposite the Crowne Plaza."*[1], or **web queries**: *"gas station near my current location"*.

Especially the case of web queries is interesting. This since studies have shown that a substantial percentage $(14\% - 18\%)$ of search queries with traditional search engines contain at least one geographic related term [25,5] and among these search terms the most frequent is "near". Hence, the understanding of "near" is an important and challenging problem in human-machine interaction, query answering systems, semantic search or location based-services (LBS).

Spatial distance relations are usually characterized as a 2-place relation, such as $near(x, y)$, whereas the first argument of the relation is called the to-be-located-object (LO) and the second argument is called reference object (RO) [15].

[1] http://www.tripadvisor.co.uk/ShowUserReviews-g34438-d87115-r75717164-Crowne_Plaza_Miami_Airport-Miami_Florida.html

For instance, at the example "the gas station near the old market", this can be represented as the relation: $near(gasstation, oldmarket)$, whereby "gas station" is the LO and "old market" the RO.

Current models for "near" take a simplistic approach. One approach is by assuming an a priori determined radius around the object of interest. Then, for instance, everything within a radius of $15km$ is deemed to be near. Another approach is to use an implementation of the (k-)nearest neighbor algorithm. A descending list of objects with respect to Euclidean distance is generated and a cut-off point or threshold on this list is assumed ($k \leq n \cong near$). For instance, the value for the cut-off point can be 15 and, consequently, the 15 nearest objects are considered to be near.

For practical reasons, these models, however, ignore the fact, that "near" is context dependent and interpreting "near" in different contexts results in a different distance threshold [6]. For instance, "a gas station nearby" means something different in an urban densely populated area, where gas stations are more common than in rural and sparsely populated areas. In the former case "a gas station nearby" means at most 10 min driving time, in the latter case this can mean up to one hour or more.

It is expected that a contextualized model of "near" not only is more adequate but also more accurate. In order to be able to formulate a contextualized model of "near" the following needs to be identified. (i) what is the relevant contextual information for "near"? (ii) how does the identified information influence the interpretation of near? To answer these questions, a systematic investigation of the relation between context and "near" is needed. However, no such investigation exists so far.

My research answers these questions by identifying the context factors which have an influence on "near" and then learn from data to assess the actual distance of "near" in a given context.

2 Related Work

A way to deal with context for "near" is by assuming that the context can be made explicit. This might be a domain specific model produced by a domain expert [3] or a general model containing user feedback collected with a question answering (QA) system [22]. While the two approaches take systematically context into account, they run into problems when the contextual information needs to be classified automatically. This is problematic in cases where the input by a domain expert or the user is not technically not feasible.

In this case another approach of contextual modeling within the setting of supervised learning is more appropriate. In such a setting the contextual features are detected from text [29] and mapped to a set of static attributes (context factors or CFs). This method follows the idea of Adomavicius et al. [1], who claim that contextual modeling can be achieved by representing the context as an enumerated set of CFs. To enable contextual modeling the following elements

need to be known: (i) How can "near" be measured? (ii) How do the CFs modify the distance measure for "near"?

For question (i), defining an adequate distance measure for "near" has proven to be difficult. This as the nearness relation has the peculiarity of being **positive**, **not necessarily symmetric**, and ignoring the **triangle-inequality** [33]. For this reason, various different distance measures have been proposed. Among the **quantitative** distance measure these are: *Eucledian distance* [7,6], *relative distance* [33], or *travel time* [20,14]. Among the **qualitative** distance measure, these are: *network distance* (i.e. street network) [9,3], and *topological connectivity* (i.e. adjacency of regions) [2,10].

For question (ii) it is then hypothesized that the CFs can either induce a threshold [6] on the distance measure or modify the scale of the distance measure. In previous studies, so far, it has been detected that a CF can either shorten or lengthen the "near" distance.

Factors that have a decreasing effect are **barriers** [16,23,4,3], such as rivers, railway tracks, highways or, borders.

Factors that have an increasing effect are as follows. **Channels** [19,11,17]: channels are objects where a person can move along, such as rivers, railway tracks, or highways[2]. **Familiarity** [35,13]: If the subject is familiar with the LO an increased distance is observed. **Reachability** [9,35,3]: The reachability via a network (i.e. route network, public transport) influences "near"and the higher the reachability of the LO, the higher the distance for "near". **Type of the object** [9,31]: In case the RO is a point of interest (POI), such as a museum or a cinema, increased distances for "near" are observed. The distance further increases in case the object is a landmark (Big Ben, Eiffel Tower).

Factors that either increase or decrease the distance are as follows. **Hierarchical information** [27,21,30]: Geographical information of country borders or administrative units influence the distance of "near". This can also be the common upper-level administrative unit of the LO and the RO [10]. **Dimensions of the objects** [28,9,3,22]: a small object has a smaller nearness distance than a bigger one. **Affinity/Attraction** [8,35,4]: If a subject has a special attraction towards the RO or LO, such as political or social affiliation, the distance for "near" increases. In case of a negative attitude towards the LO or RO, the distance for "near" decreases.

While a wide variety of different CFs have been detected so far, these CFs, however, have been studied mostly in isolation. Thus, a systematic study of the CFs is missing, as the CFs have rarely been studied in combination with each other, nor are there any studies investigating the conditions for an appropriate CF.

[2] Objects such as rivers, railway track lines, highways can serve both as barriers and as channels.

3 Research Questions & Hypotheses

The aim of my research is to formulate a contextualized nearness model. Such a model can classify objects as "near/not near" depending on CFs in combination with its appropriate distance measure. For doing so, a systematic study is needed from which we can infer how a CF determines what is considered to be "near/not near" and how well such CFs improve the classification of "near/not near". These assumptions inform my research questions (RQ1 — RQ3): **RQ1** is concerned with context factors in general: **RQ1.1:** Do context factors determine nearness? **RQ1.2:** Which context factors determine nearness? **RQ2** is concerned with the comparability of CFs: Do context factors interact with one of the nearness measures (Euclidean distance, traveling time, network connectivity) or with each other? **RQ3** is concerned with the proximity measures: Which distance measure makes a better predictor for nearness modeling?

Validating the following hypotheses (H1 — H3) will answer my RQs. **H1.1:** Using only a distance measurement (quantitative or qualitative distance measurement) gives us less accurate nearness predictions. **H1.2.** At least the following context factors determine nearness: Mode of Transportation (i.e. driving, walking, public transport), size of the object, reachability, urban/rural and its common upper level administrative division. **H2:** A model consisting of "driving" and "urban" induces a different threshold on the absolute distance than "driving" and "rural" or "walking" and "urban". **H3:** How well a measure predicts "near", depends on the context. For instance, in the context of traveling, traveling time is a better predictor than Euclidean distance or network connectivity.

4 Research Methodology

I will formulate context dependent models of "near" by using an inductive algorithm (supervised prediction). By identifying the relevant CFs, a contextual threshold for "near" or a probability of class membership can be learned.

Most training data will consist of geo-processed nearness information from various NL corpora (within the domain of geographic description and tourism): Geograph[3] and TripAdvisor [4]. I will follow related work on extracting the nearness relation [12,34,31,26]. This task consists of identifying the word "near", its synonyms and the arguments of the nearness relation (RO & LO). In case of the Geograph.co.uk corpus, a sentence such as "This large substation is near the towns of Wrexham and Rhosllanerchrugog in Wales"[5] can be found. The LO is identified as the GPS-coordinates of the photo and the RO is identified by using a NL-syntax parser by looking for the named entity within the syntactic sub-tree of the prepositional phrase. The nearness relation is extracted together with the event or activity such as "near for a daily commute" or "near for grocery

[3] http://geograph.co.uk

[4] http://archive.ics.uci.edu/ml/datasets/OpinRank+Review+Dataset

[5] http://www.geograph.org.uk/photo/39134

shopping" so that the context can be identified. Additionally, the objects (LO,RO) are georeferenced, which enables to calculate the necessary metrics for our CFs:

- distance metrics (Euclidean distance between LO and RO via Haversine Formula)
- travelmetrics (walking, driving, biking time between LO and RO via Openstreetmap)
- the size of RO
- properties of the LO and RO (urban, rural, county/district, point of interest) [6]
- determine the connectedness of the RO and LO

Additionally I will also use specialized datasets for "near", such as the collection of recreation areas near Swiss urban centers [17].

My contextualized models are schematized as follows: $near/\neg near = \mu + \beta_1 + \beta_2 + \ldots + \beta_n$. That is, a model consists of a distance measure (μ): quantitative (Euclidean distance, travel time) or qualitative (network connectivity) and a set of context factors ($\beta_1 + \beta_2 + \ldots + \beta_n$). Hence, a model in the domain of traveling might look as follows: near = traveltime + dimension object + urban + type. For each model I will use supervised prediction to learn a threshold or class membership probability for near, based on context factors and a distance measure. The methods for doing so will be by logarithmic regression and Bayesian classifiers. For model comparison, I will use Receiver Operating Characteristic (ROC) curves [24], which compare the different models with respect to their expected outcome (sensitivity & specificity). This, because I assume a skew in my dataset. I will also validate my models via cross-validation to avoid overfitting of my contextualized models.

My hypotheses are tested as follows: For **H1.1.** I formulate two models. The first model contains only a distance measure without any CF and the second contains a distance measure and all CFs. H1.1. is confirmed when the prediction accuracy of the model with the CFs is higher than the model without the CFs.

H1.2: I formulate a model with all the context factors. The second model contains all the context factors minus one CF. If the accuracy decreases then the context factor has a contributive effect and H1.2. is confirmed.

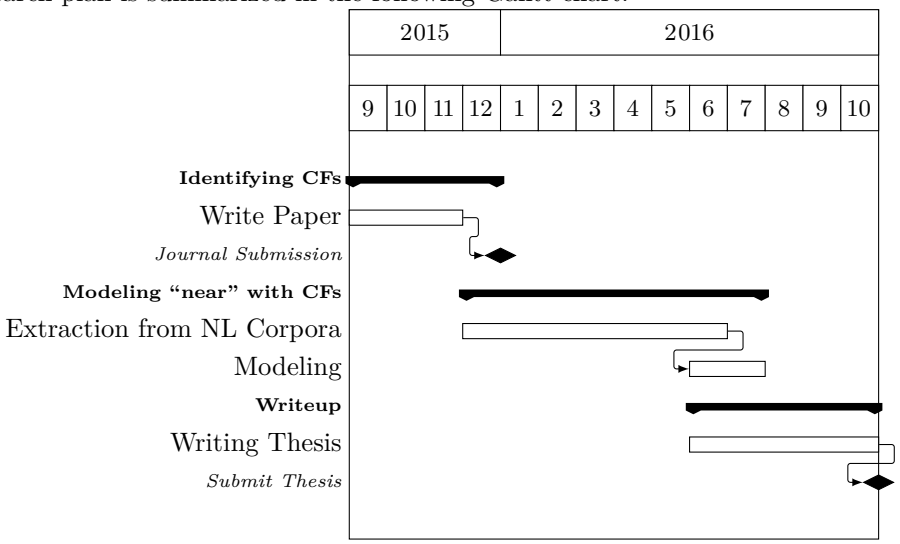
H2: To check if there is an interaction between either the distance measure and its context factors or between the context factors, I check if the distance measure is conditionally dependent on one of the context factors. If the context factors are conditionally dependent on each other, interaction effects are present. To test this I will use the Naïve Bayes classification algorithm, which assumes that all the independent variables are conditionally independent of each other. The AODE algorithm [32] weakens the independence assumption If the accuracy of Naïve Bayes is as good or better than AODE (compared via ROC), no interaction among the independent variables occur. If, however, the accuracy of AODE is better, interaction among the variables occur.

[6] This information can be obtained from Corine Landcover dataset: `http://www.eea.europa.eu/data-and-maps`

H3 I build several models ($M = \mu_1 + \beta_1 + \beta_2 + \ldots + \beta_n$), whereby the distance measure (μ_n) varies for each model. If the accuracy of one model is better in one context, H3.1. is confirmed.

5 Research Plan

My research consists of the following tasks: (i)identifying the CFs, (ii) extract the information from NL-corpora and (iii) use the gained information to build context dependent models for "near". Currently, I am finishing my first task, which aims to identify the CFs in the literature. This work cumulates into a literature review paper which is in preparation and is scheduled to be finished in November. My second task is to build contextualized models for "near". The necessary data for my modeling task will be data extracted from NL-corpora. I assume that the extraction from NL-corpora will be the most time consuming task (November 2015 — June 2016) and upon completion I can formulate and evaluate my context-dependent models for "near". In parallel I will also start my write-up and finish and submit my thesis (July 2016 — October 2016) and my research plan is summarized in the following Gantt chart:



6 Expected Contributions

The goal of my research is to formulate contextualized models for "near". I do this by empirically testing how to determine what is near given the information obtained from the context. By doing so, I expect to get an insight on which contextual information is relevant for "near" and I will make the following contributions. From an engineering point of view, I will provide nearness models with an increased accuracy which also needs to make fewer a priori assumptions. From a cognitive point of view I will provide the conditions for and when a CF

has a contributing effect on determining what is near and henceforth give more insight on the interplay between context and "near".

References

1. Adomavicius, G., Mobasher, B., Ricci, F., Tuzhilin, A.: Context-aware recommender systems. AI Magazine 32(3), 67–80 (2011)
2. Bera, R., Claramunt, C.: Topology-based proximities in spatial systems. Journal of Geographical Systems 5(4), 353–379 (2003)
3. Brennan, J., Martin, E.: Spatial proximity is more than just a distance measure. International Journal of Human-Computer Studies 70(1), 88–106 (2012)
4. Carbon, C.C., Leder, H.: The wall inside the brain: overestimation of distances crossing the former Iron Curtain. Psychonomic bulletin & review 12(4), 746–750 (2005)
5. Delboni, T.M., Borges, K.A.V., Laender, A.H.F., Jr, C.A.D.: Research Article Semantic Expansion of Geographic Web Queries Based on Natural Language Positioning Expressions. Transactions in GIS 11(3), 377–397 (2007)
6. Denofsky, M.: How near is near? Tech. rep., Massachusetts Institute of Technology (1976)
7. Dolbear, C., Hart, G., Goodwin, J.: From theory to query: Using ontologies to make explicit imprecise spatial relationships for database querying. In: Conference on Spatial Information Theory (COSIT) (2007)
8. Ekman, G., Bratfisch, O.: Subjective distance and emotional involvement. A psychological mechanism. (1965)
9. Gahegan, M.: Proximity operators for qualitative spatial reasoning. In: Frank, A.U., Kuhn, W. (eds.) Spatial information theory a theoretical basis for GIS, pp. 31–44. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (1995)
10. Grütter, R., Scharrenbach, T., Waldvogel, B.: Vague Spatio-Thematic Query Processing: A Qualitative Approach to Spatial Closeness. Transactions in GIS 14(2), 97–109 (2010)
11. Gryl, A., Moulin, B., Kettani, D., Gryll, A., Moulin, B., Kettani, D.: A Conceptual Model for Representing Verbal Expressions used in Route Descriptions. Spatial Language pp. 19–42 (2002)
12. Hall, M.M., Jones, C.B.: A field based representation for vague areas defined by spatial prepositions. In: The Workshop Programme Methodologies and Resources for Processing Spatial Language (2008)
13. Hall, M.M., Jones, C.B.: Cultural and Language Influences on the Interpretation of Spatial Prepositions. In: GI_Forum 2012: Geovisualization, Society & Learning. Berlin/Offenbach (2012)
14. Helming, I., Bernstein, A., Grütter, R., Vock, S.: Making close to suitable for web search : A comparison of two approaches. In: Terra Cognita - Foundations, Technologies and Applications of the Geospatial Web. No. October, Bonn, Germany (2011)
15. Herskovits, A.: Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English. Studies in Natural Language Processing, Cambridge University Press (1986)
16. Hirtle, S.C., Jonides, J.: Evidence of hierarchies in cognitive maps (1985)
17. Kienast, F., Degenhardt, B., Weilenmann, B., Wäger, Y., Buchecker, M.: GIS-assisted mapping of landscape suitability for nearby recreation. Landscape and Urban Planning 105(4), 385–399 (2012)

18. Lankford, A.: Biking the Grand Canyon Area. Westcliffe Publisher, Inc., Englewood, CO (2003)
19. Lynch, K.: The Image of the City. MIT Press, Cambridge, MA (1960)
20. MacEachren, A.M.: Travel Time As the Basis of Cognitive Distance. The Professional Geographer 32(1), 30–36 (1980)
21. McNamara, T.P.: Mental representations of spatial relations. Cognitive Psychology 18(1), 87–121 (1986)
22. Minock, M., Mollevik, J.: Context-dependent near and far in spatial databases via supervaluation. Data & Knowledge Engineering 86(270019), 295–305 (2013)
23. Montello, D.: The perception and cognition of environmental distance: Direct sources of information. In: Hirtle, S., Frank, A.U. (eds.) Spatial Information Theory A Theoretical Basis for GIS. Proceedings of COSIT. pp. 297–311. Springer, Berlin (1997)
24. Provost, F., Fwacett, T.: Robust Classification for Imprecise Environments. Machine Learning 42(3), 203–231 (2001)
25. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: SIGIR Workshop on Geographic Information Retrieval. vol. 2 (2004)
26. Skoumas, G., Pfoser, D., Kyrillidis, A.: Location Estimation Using Crowdsourced Geospatial Narratives (aug 2014), http://arxiv.org/abs/1408.5894
27. Stevens, A., Coupe, P.: Distortions in judged spatial relations. Cognitive psychology 10(4), 422–437 (1978)
28. Talmy, L.: How language structures space. In: Toward a Cognitive Semantics, vol. 1, chap. 3, pp. 177–254. MIT Press, Cambridge, MA (2000)
29. Turney, P.: The Identification of Context-Sensitive Features: A Formal Definition of Context for Concept Learning. 13th International Conference on Machine Learning, Workshop on Learning in Context-Sensitive Domains, Bari, Italy pp. 53–59 (1996), http://arxiv.org/abs/cs.LG/0212038
30. Tversky, B.: Distortions in cognitive maps. Geoforum 23(2), 131–138 (1992)
31. Wallgrün, J.O., Klippel, A., Baldwin, T.: Building a Corpus of Spatial Relational Expressions Extracted from Web Documents. In: SIGSPATIAL '14. ACM, Dallas/Fort Worth, TX, USA (2014)
32. Webb, G.I., Boughton, J.R., Wang, Z.: Not so naive Bayes: Aggregating one-dependence estimators. Machine Learning 58, 5–24 (2005)
33. Worboys, M.F.: Metrics and topologies for geographic space. In: Proc. 7th Intl. Symp. Spatial Data Handling. Delft, Netherlands (1996)
34. Xu, S., Klippel, A.: Linking context and proximity through web corpus. In: Proceedings of the 7th Workshop on Geographic Information Retrieval - GIR '13. pp. 45–46. ACM Press, New York, New York, USA (2013)
35. Yao, X., Thill, J.C.: How Far Is Too Far? - A Statistical Approach to Context-contingent Proximity Modeling. Transactions in GIS 9(2), 157–178 (2005)