It Depends

Carsten Keßler
Institute for Geoinformatics, University of Münster, Germany
carsten.kessler@uni-muenster.de

Abstract

Context plays a crucial role for semantic similarity measurements. Depending on aspects such as location, time, task and user preferences, the respects that play a role for the comparison of concepts may vary. This variance should be reflected in a change of the similarity measurements' results. Although this dependency has been observed in various human subject tests and is accounted for in recent similarity theories, there is no generic model of the characteristics of similarity measurement so far. The objective of this research is the formalization of such a model, which clarifies the exact systematics of the dependency of similarity on context. From a practical perspective, this model must allow for the determination of the most influential contextual aspects for similarity measurements in a concrete application.

Introduction & Motivation. *"Are canals more similar to rivers than to harbours?"—"Well, it depends."*

The assessment of the similarity of concepts (and also instances) is heavily influenced by context. Accordingly, similarity measurements must take context into account to produce cognitively plausible results, instead of calculating generic similarity values that do not consider the respects of a comparison [6]. One might even go further and doubt the utility of context-free similarity measurements [2], since people's similarity ratings in everyday situations are always influenced by their current context. The objective of this research is hence the development of a generic context model for semantic similarity measurement that supports the identification of the respects that influence similarity for a given task.

Even though the influence of context is beyond dispute, the modelling of context for similarity measurement has been either disregarded, or handled as a mere add-on to the actual similarity theory so far. A more comprehensive utilization of contextual information in similarity measurement can help to produce results that are better tailored to the current situation of an individual. Moreover, context enables the disambiguation of terms such as "long" or "heavy" on a personalized basis: taking a user's profile into account, which may, for example, contain information on previously taken hiking tours, such terms can be quantified for computation in a similarity measurement. Without contextual information, the required quantification is basically a wild guess.

As mentioned above, the aim of this research is not to generate yet another application specific context model, but the development of a set of formal, generic characteristics of context for similarity measurement, independent of a specific knowledge representation. This generic model should be expressive enough to determine which potential context parameters influence a similarity measurement, and how influential they are. It is obvious that a similarity measurement produces results that are better adapted to the current situation if more (contextual) parameters are taken into account. Thus, it is important to be able to specify which of those parameters are the most influential ones, and which have little impact on the overall result. This is also relevant from an economic perspective, as collection, storage and computation of context parameters is expensive, so that it must be ensured that only relevant parameters are considered. The envisioned formal model must thus support the selection of the relevant parameters for concrete applications.

Similarity and Context in the Geospatial Domain. Both similarity measurement and context have become major research issues within the geographic information community over the last years. Similarity measurement has been investigated with the integration of heterogeneous spatial data sources and the enhancement of geographic information retrieval in mind. Concerning context, location is one of the most important contextual aspects for many applications[1] and plays an essential role in location based services and mobile decision support systems, for example. Previous research within this field of research embarks on different strategies for the integration of similarity measurement and context. The matching-distance similarity measure [8] introduces a context-dependent similarity measure for geospatial entity classes based on activities. The SIM-DL approach [4] focuses on concepts defined in description logics, formalizing the context as a single concept stated together with the similarity query. A geometric approach based on conceptual vector spaces is introduced in [7], where dissimilarity is represented by semantic distance, allowing for different contexts through weighting of the quality dimensions of the space.

As indicated by the examples above, there are already approaches to similarity measurement at hand that provide a reasonable consideration of context. However, the presented notions of context are depending on the corresponding similarity theory, i.e. there is no generic context model that is independent of a specific similarity theory. Moreover, the existing approaches all treat context as a (weighted) subset of the knowledge at hand. This is remarkable because the notion of context used across other fields of research mostly implies that the inclusion of context should *add* information to what is already known.

Similarity Measurement in Practice. Research on similarity measurement, especially from the psychological perspective, has produced substantial

[1] Context is even reduced to location in some cases [9].

findings during the last decades [1]. Nonetheless, applications making use of the developed theories are a long time coming, which is mostly due to the special knowledge representations they require. Theories such as Conceptual Spaces [3] focus on reflecting human cognition as good as possible, but are hard to use in practice because they are not applicable to widespread knowledge representations. Application specific context models must thus obey the technical prerequisites imposed by existing knowledge representations used within the application to ensure practicability. One such prerequisite, for example, is the widespread use of the Web Ontology Language (OWL) for concept representations.

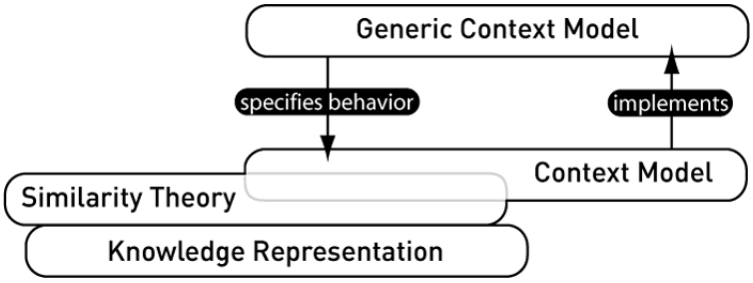

Figure 1: The generic context model specifies the behavior of concrete context models. A concrete context model may also be (partly) covered by the similarity theory, which is tailored to a specific kind of knowlegde representation such as OWL ontologies or conceptual spaces, for example.

Dependence of Similarity on Context. The purpose of a generic context model is to specify the behavior of concrete context models, as shown in figure 1. The question how much can actually be said about context at such a generic level is still subject to future research; more detailed specifications of the behavior of context may require definitions at a level that is bound to a specific knowledge representation. In the following, a general process in the computation of context-aware similarity measures and a first draft for a set of generic characteristics of context based on sets is described [5].

The following process has been observed for similarity measurements that take contextual information into account:

1. Context capturing, either automatic or by explicit input
2. Context injection, where contextual information that is not already part of the knowledge base is added to it
3. Selection of the domain of application
4. Weighting of the domain of application
5. Similarity measurement

The set-based formalization of context characteristics is based on the assumption that contextual information C can be distinguished into *internal* and *external* context: whereas internal context is already present in the knowledge base K, the external context needs to be added to the knowlegde base first (referred to as *context injection* in the process described above). To enable this injection which creates the extended knowledge base K_E (eq. 1), context and knowlegde base must refer to a shared vocabulary, i.e. the knowledge base in this case (eq. 2):

$$K_E = C \cup K \tag{1}$$

$$C \cap K \neq \emptyset \tag{2}$$

The requirement that the context should only contain information that affects the similarity measurements in a significant way is formalized by introduction of a threshold value δ: the context is reduced to parameters that have a minimum impact on similarity measurements (eq. 3). The impact is defined as the mean difference between a similarity measurement in a context *with* the parameter compared to one *without* the parameter (eq. 4):

$$C = \{c | imp(c) > \delta\} \tag{3}$$

$$imp(c_n) = \frac{\sum |sim_{(c_n \in C)} - sim_{(c_n \notin C)}|}{|C|} \tag{4}$$

As outlined above, the domain of application D is then defined as the subset of all concepts in the extended knowledge base K_E that are used to define one of the compared concepts a, b (eq. 5). The function w assigns weights to the concepts in D according to their importance for the current context (eq. 6):

$$D = \{c \in K_E | c \sqsupseteq a \sqcup c \sqsupseteq b\} \tag{5}$$

$$w : D \times D \longrightarrow [0, 1], \sum w = 1 \tag{6}$$

While the equations above make general statements about context and similarity, they do not contain information on the systematics of the influence of context on similarity. Context does not only have *some* impact on a similarity measurement, but this impact follows a certain systematics: intuitively, the more similar two contexts are, the less a similarity measurement should change under those two contexts. In other words, the difference between the results of a similarity measurement in two different contexts converges to 0 with a growing similarity of the two contexts (eq. 7):

$$\lim_{sim(C_1, C_2) \to 1} sim_{C_1}(a, b) - sim_{C_2}(a, b) = 0 \tag{7}$$

Application Scenarios and Outlook. One prospect of this research is to demonstrate the fitness for use of current similarity measurement approaches, that make use of a context model for more user-oriented results. For this purpose, a Web portal for the exchange of cycling routes is planned, including a similarity-based search module. In this case, context will mostly consist of the users' profiles containing information about their preferences, equipment and previously taken routes. Moreover, current work within the SimCat project[2] (which also provides the scope for this research) aims at the development of a similarity server, based on the SIM-DL approach [4]. The current use case for the similarity server is a gazetteer interface that makes use of similarity and a basic context theory, accessing an OWL ontology defining geospatial feature types. The similarity server will also be used for human subjects tests in order to verify the aspired characteristics of the context model, and to assure that it is cognitively plausible. However, before such tests can be made, the current set of characteristics of the generic context model [5] must be completed and formalized. One major research issue that needs to be addressed to complete the model is the question how to compare different contexts.

Going back to the example at the beginning of this short paper, the envisioned context model should be able to answer the question *"on what does it depend?"*

Acknowlegdements. This research is partly funded by the German Research Foundation (DFG) under the project title "Semantic Similarity Measurement for Role-Governed Geospatial Categories". Special thanks go to Martin Raubal and Krzysztof Janowicz for helpful comments and fruitful discussions.

References

[1] R. L. Goldstone and J. Son. Similarity. In K. Holyoak and R. Morrison, editors, *Cambridge Handbook of Thinking and Reasoning*, pages 13–36. Cambridge University Press, 2005.

[2] N. Goodman. Seven strictures on similarity. In N. Goodman, editor, *Problems and projects*. Bobbs-Merrill, 1972.

[3] Peter Gärdenfors. *Conceptual Spaces - The Geometry of Thought*. MIT Press, 2000.

[4] K. Janowicz. Sim-dl: Towards a semantic similarity measurement theory for the description logic alcnr in geographic information retrieval. In Meersman et al., editor, *SeBGIS 2006, OTM Workshops 2006*, volume 4278 of *LNCS*. Springer, Berlin, 2006.

[2] http://sim-dl.sourceforge.net/

[5] C. Keßler. Similarity measurement in context. In *Sixth International and Interdisciplinary Conference on Modeling and Using Context*. Lecture Notes in Artifical Intelligence, Springer, Roskilde, Denmark, to appear 2007.

[6] R. Goldstone Medin, D. and D. Gentner. Respects for similarity. *Psychological Review*, 1993.

[7] M. Raubal. Formalizing Conceptual Spaces. Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004). 2004.

[8] A. Rodríguez and M. Egenhofer. Comparing Geospatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, 18(3), 2004.

[9] A. Schmidt, M. Beigl, and H. W. Gellersen. There is more to context than location. *Computers & Graphics*, 23(6):893–901, 1999.