Semantic Similarity Measurement and Geospatial Applications

Krzysztof Janowicz
janowicz@uni-muenster.de
Institute for Geoinformatics
University of Münster, Germany

Martin Raubal
raubal@geog.ucsb.edu
Department of Geography
University of California, Santa Barbara, USA

Angela Schwering
aschweri@uos.de
Institute of Cognitive Science
University of Osnabrück, Germany

Werner Kuhn
kuhn@uni-muenster.de
Institute for Geoinformatics
University of Münster, Germany

With the increasing amount of geographic information available on the Internet, searching, browsing, and organizing such information has become a major challenge within the field of Geographic Information Science (GIScience). As all information is ultimately for and from human beings, the methodologies applied to retrieve and organize this information should correlate with human similarity judgments. Semantic similarity measurement, which originated in psychology, is a methodology fulfilling this requirement and supporting geographic information retrieval.

The following special issue presents work on semantic similarity measurement from different perspectives, including cognitive science, information retrieval, and ontology engineering, with a focus on applications in GIScience. It originated in the *Workshop on Semantic Similarity Measurement and Geospatial Applications* held in conjunction with COSIT 2007, the International Conference on Spatial Information Theory (http://www.cosit.info/). A substantial part of the workshop contributions addressed the need for similarity measurement in geographic information retrieval, including applications in web service discovery, knowledge management, and emergency scenarios. The call for papers to this issue was based on these workshop contributions, discussions, and results (http://musil.uni-muenster.de), but open to any submissions on the role of semantic similarity in GIScience. Eleven papers were submitted and then

reviewed by at least three internationally renowned scientists. Out of the selected four papers, two were written by workshop participants and two were contributed by others. The following sections motivate the need for semantic similarity in GIScience and relate the four papers to it.

Motivation

Most information on the Internet is either (annotated) text, multimedia content, or expressed using some kind of logic format, such as description logics in case of the (geospatial) semantic Web (Egenhofer, 2002). Besides classical (Boolean) retrieval methodologies, such as keyword matching or subsumption reasoning, similarity plays an increasing role as a measure of overlap. It spans from syntactic string comparison such as edit-distance (Levenshtein, 1966), to the computation of instance and class similarity used for retrieval and alignment on the semantic Web. In the latter case, similarity measures the degree of overlap between semantic descriptions.

Many geospatial applications offer potential for the integration of similarity-based information retrieval techniques. Web interfaces, such as the similarity-enabled gazetteer interface for the Alexandria Digital Gazetteer (Janowicz et al., 2007) can propose similar geographic features or feature types for a user's query. Location-based services could derive points of interest in the user's current neighborhood from similar, previously visited places. Similar web services could be proposed if a certain service is temporarily offline. Similarity could also be applied to improve ontology engineering, data integration, or ontology alignment and mapping. In general, the more information resources become available (e.g., via the semantic Web), the higher is the need for methods supporting the interaction with these resources. Similarity is one of these methods, as it supports users in retrieving and browsing through information, and hence in knowledge acquisition. In fact, every application that deals with fuzzy or ambiguous input – either from human beings or software agents – is a potential candidate for similarity measurement.

Foundations of Semantic Similarity

Measures of semantic similarity have a long tradition in the cognitive sciences and especially in psychology (Gentner and Markman, 1994; Goldstone and Son, 2005). Similarity estimations are among the fundamental processes underlying human categorization and inference (Medin et al., 1993). Cognitive science research has investigated various kinds of similarities over the last fifty years, including similarity between individuals, classes, complex (pictorial) scenes, and processes. Various approaches to model similarity and reason about it have been developed. While feature-based models (Tversky, 1977) are most prominent, network-based (Rada et al., 1989), geometric (Torgerson, 1965; Shepard, 1987), alignment (Goldstone, 1994), transformation-based (Hahn et al., 2003), and information theoretic (Resnik, 1995) approaches have also evolved. Understanding human cognition forms the main motivation underlying these approaches, but recent research in information science has applied computational similarity theories as reasoning support for information retrieval and organiza-

tion (Rissland, 2006; Möller et al., 1998; Janowicz et al., 2007; Maedche and Staab, 2002; d'Amato et al., 2005; Schwering and Kuhn, forthcoming). In the following, some core characteristics of semantic similarity are described:

Properties of Semantic Similarity One of the reasons why developing computational semantic similarity theories is a challenging task is that in many cases neither *symmetry*, *transitivity*, *triangle inequality*, nor *minimality* hold (Tversky, 1977; Goldstone and Son, 2005). For instance, it cannot be assumed that A is similar to C because of sim(A, B) and sim(B, C). Most relevant for information retrieval is the fact that similarity estimations are not necessarily symmetric. Consequently, the majority of theories developed for geospatial applications define similarity as an asymmetric relationship (Rodríguez and Egenhofer, 2004; Janowicz et al., 2007).

Semantic Similarity depends on Context As Goodman (1972) puts it, the similarity between A and B is meaningless without stating with respect to what both are similar or at least without defining a reference C for comparison. But similarity does not only depend on the compared characteristics or the set of compared individuals and classes. It has been demonstrated that age, cultural background, as well as user motivation and application also play key roles (Medin et al., 1993; Janowicz, 2008). Consequently, there is no universal similarity measure and each measure is only applicable to particular use cases.

Semantic Similarity depends on Representation Whether and to which degree A and B are similar depends on what is said (in a computational representation) about both. Therefore, similarity theories are bound to a particular representation language such as description logics or semantic networks. In addition, classes and individuals change over time, as do their computational representations (Raubal, forthcoming 2008). For instance, the assumption that rivers are linear waterbodies while lakes are not, may not hold in the case of a flooding event. Hence, from a similarity point of view, flooded rivers are similar to lakes (Keßler et al., 2007).

Semantic Similarity, Usability, and Cognitive Plausibility Comparing computational similarity ratings to human similarity assessments has become a well-established strategy to evaluate semantic similarity measures: The similarity rankings obtained by comparing individuals, classes, or scenes using a computational theory have to correlate with human similarity rankings (Rodríguez and Egenhofer, 2004). Nevertheless, a computational similarity theory is not per se cognitively plausible. Human participants tests are therefore required to investigate whether a certain measure is cognitively plausible or not. As similarity depends on representation, one of the most critical challenges for such testing is to ensure that both the computational representation and the descriptions handed out to the participants are comparable. If the rankings correlate, this does not mean that the concrete similarity values also do. In fact, humans tend toward individual gradings of what is considered similar and what not.

Additionally, if a particular measure is cognitively plausible, this does not imply that the results are represented in a way accessible to human reasoning. For instance, similarity rankings could be represented using decreasing font sizes

or different colors instead of purely numerical representations (Janowicz et al., 2007).

The Role of Semantic Similarity for GIScience

While it is out of scope for this introduction to give a comprehensive overview of ongoing research on semantic similarity measurement, we point to some of this work to demonstrate the various application areas of similarity within GIScience. A detailed overview was recently given by Schwering (2008).

By extending Tversky's classical feature-based model (Tversky, 1977), the Matching Distance Similarity Measure (MDSM), proposed by Rodríguez and Egenhofer (2004), was one of the first similarity measures, which has been developed specifically for the geospatial domain. It also includes an initial context theory that laid the background for further work on the context dependency of similarity (Keßler et al., 2007; Janowicz, 2008).

Raubal (2004) proposed geometric similarity measures based on conceptual spaces (Gärdenfors, 2000) for landmark-based navigation. Facades of buildings, which are locally most dissimilar to the neighboring facades, were selected as prominent landmarks for route instructions in a pedestrian navigation service in Vienna. Schwering and Raubal (2005) extended this approach to implement inter-class similarity measures by integrating spatial relations.

Ahlqvist (2004) combined the cognitive theory of conceptual spaces with a formal representation of semantic uncertainty based on rough fuzzy sets. This approach allows for modeling the uncertainty often found in geographic classes. He also investigated the role of semantic similarity to detect category and land cover change (Ahlqvist, 2005).

Starting with the work of Bruns and Egenhofer (1996), various researchers addressed the question of how to compare spatial scenes for similarity and how to construct similarity-based queries (Nedas and Egenhofer, 2003) for information retrieval. In their Topology-Direction-Distance (TDD) model, Li and Fonseca (2006) investigated the fundamental components of spatial similarity assessments combining feature-based, alignment-based, and transformational approaches to similarity.

Motivated by the gap between description-logics-based ontologies on the geospatial semantic Web and existing similarity measures, which were not able to deal with the expressivity of these languages, Janowicz et al. (2007) implemented the SIM-DL theory. It allows for measuring similarity between classes specified using various expressive description logics.

Gahegan et al. (2007), developed ConceptVista, an ontology management and learning environment that uses similarity for browsing through classes, but also for negotiation, i.e., to establish a common agreement among domain experts.

Further Challenges for Similarity Measurement

This section points to some of the forthcoming challenges and research directions for semantic similarity.

Explanation of Similarity Values For a given pair of individuals, classes, or scenes, similarity theories provide numerical values (usually between 0 and 1) as a measure for their degree of overlap. To support the user in interpreting these results, a similarity reasoner should also explain the results. Besides providing the overall similarity through a numeric value and ranking, a reasoner could display which characteristics were selected for comparison and how they perform. This would not only be of great benefit for information retrieval, but also for ontology engineering and knowledge organization.

Approximation of Similarity Values Semantic similarity measures are complex and, in most cases, expensive in terms of computation time. This is especially the case for ontology alignment and information retrieval in large knowledge bases consisting of thousands of compared classes or individuals. The approximation of similarity values is a promising approach to significantly reduce computation time. This can be achieved in two ways, either by improving the selection process of the compared entities, or by approximating similarity values at first and only investigate those in detail which are above a certain threshold.

Integration of Extended Context Theories While an increasing amount of similarity measures is context-aware, the notion of context is often reduced to the selection of compared-to classes, individuals, or scenes, and the weighting of particular characteristics. More sophisticated theories should take contextual information into account to alter the similarity functions as such as well as the computational representations of the compared entities (Janowicz, 2008). One of the difficulties of such an approach is the immense variety of contextual information. For instance, in case of a mobile recommendation service, contextual information may span from the user's age, personal preferences, and available public transportation to the current weather. Hence, one important pre-condition to context-aware similarity measures is the extraction of the most relevant contextual information. Context Impact Measures (CIM) are such an approach (Keßler et al., 2007).

Semantic Similarity between Perdurants Up to now, most similarity theories have focused only on endurants. Endurants, such as a human being or vehicle, persist over time and all of their characteristics can be perceived (and hence compared) at each temporal snapshot. In contrast, perdurants are not entirely present at all times. This includes all kinds of activities or events. As not all their characteristics are perceivable simultaneously, their comparison is more complex and remains an open issue.

Semantic Similarity and Analogy Two entities are commonly judged as similar if they share many commonalities and have only few differences. However, it is not only the ratio of common and distinct features as Tversky (1977) suggested, but also the kinds of matches that matter. Gentner (1989) distinguishes between two types of matches: common attributes and common relations. She suggests to decompose similarity into finer subclasses of which we discuss four prominent types here: literal similarity, analogy, mere appearance, and metaphor. While literal similarity is characterized by common attributes

and common relations, analogy is characterized by a common relational system. This means that in literal similarity, wide parts of the description of one entity are also applicable to the other entity. Between analogous entities it is mainly the structural descriptions that match. Mere appearance similarity denotes similarity primarily based on common attributes, but there are no deeper structural relationships. Metaphors are a special case of similarity, where two entities share only very limited common attributes or relations. Sometimes, these entities do not have any similarity in advance, but it is only created through the metaphorical relationship Indurkhya (1992). One must be aware that similarity can be perceived based on various criteria. Therefore, depending on the task, different similarity measures must be chosen.

Semantic Similarity within Semantic Reference Systems The notion of Semantic Reference Systems, as proposed by Kuhn (Kuhn, 2003; Kuhn and Raubal, 2003), suggests that all conceptual representations can be interpreted as semantic spaces, and that these spaces have not only topological structure (e.g., a class includes another class), but also order and metric relations. Semantic similarity measures are the obvious candidates to provide the latter. It remains to be seen how they can contribute to Semantic Reference Systems, helping to establish conceptual relations across ontologies and to ground semantics in physical or cognitive foundations.

The Contributions in this Issue

This special issue is rather a snapshot of ongoing work on semantic similarity than a comprehensive compilation of achievements. It is primarily meant as an enticement to initiate more work at this fruitful intersection of engineering, computing, mathematics, ontology, and cognition. The four papers following this introduction cover a broad spectrum of topics, ranging from spatial-scene similarity queries, over a novel approach to web service discovery and similarity-based ontology alignment, to knowledge management and negotiation for improving the interaction with domain experts during ontology engineering.

Konstantinos Nedas and Max Egenhofer investigate the foundations for plausible reasoning about *Spatial-Scene Similarity Queries*. One of the major difficulties in comparing spatial scenes, i.e., sets of spatial objects together with their spatial arrangement, is the appropriate matching of corresponding elements in both scenes. The authors propose a methodology for similarity queries that incorporates cognitively motivated approaches about scene comparisons combined with explicit domain knowledge on spatial objects and their relationships. Their formalization is based on three observations from psychology: subjects tend to match only those spatial objects that are sufficiently similar in preserving the correspondences among their relations (to other objects); they ignore very dissimilar scenes, i.e., interpret low similarity as dissimilarity; and in the presence of alternative solutions they choose those scenes requiring the least amount of change.

In *Structural Alignment Methods with Applications to Geospatial Ontologies*, Isabel Cruz and William Sunna present their revised and extended results from taking part in the Ontology Alignment Evaluation Initiative (OAEI). The authors extended their software called AgreementMaker by two novel similarity

algorithms, the Descendant's Similarity Inheritance (DSI) method, which relies on the relation between ancestor classes, as well as the Sibling's Similarity Contribution (SSC) method, which takes the relations between sibling classes into account. From a total of seven alignment methodologies taking part in the OAEI competition, their approach came in third place. The authors also provide insights into performance tuning techniques to significantly reduce the runtime of their alignment process.

In *A Platform for Visualizing and Experimenting with Measures of Semantic Similarity in Ontologies and Concept Maps*, Mark Gahegan, Ritesh Agrawal, Anuj Jaiswal, Junyan Luo, and Kean-Huat Soon describe their open platform (which is a part of ConceptVista) for experimenting with similarity measures. The platform also supports the visualization and communication of similarity values. Most importantly, one of the motivations for developing this platform is to set up an open environment for experiments to analyze which similarity measures are appropriate for particular applications and how they could be combined. The authors introduce their tool by pointing to various examples and test cases, e. g., using external domain level ontologies.

Based on the Natural Semantic Metalanguage, which specifies a set of more than 60 not further reducable and culturally independent semantic primitives, Kristin Stock presents an approach for *Determining Semantic Similarity of Behavior Using Natural Semantic Metalanguage to Match User Objectives to Available Web Services*. As both the user goals and the web services are described by semantic primitives, they can be compared to determine whether a particular service will be helpful for the user. Stock's approach consists of two separate phases. First, the primitives describing the services and the user's goal are compared for similarity. Next, the order of the primitives within the compared semantic explications is compared using their edit distance. The results from two test cases demonstrate that this methodology allows for determining which web services are most similar to the user's objectives.

Acknowledgments

We sincerely acknowledge the great efforts of the submitting authors and reviewers. The participants of the workshop at COSIT 2007 gave us the confidence that the topic warrants follow-up activities such as this issue. Too many colleagues and friends to name individually have encouraged, accompanied, taken up, and, most importantly, criticized our work in this area. This work is partly funded by the 'SimCat' project granted by the German Research Foundation (DFG Ra1062/2-1), as well as the 'Modelling of predictive analogies by heuristic driven theory projection' project (DFG KU1949/2-1).

List of reviewers

The following colleagues wrote very helpful reviews for one or more submissions to this issue: Riccardo Albertoni, Ola Ahlqvist, Boyan Brodaric, Christophe Claramunt, Isabel Cruz, Claudia d'Amato, Diarmuid O'Donoghue, Max Egenhofer, Fred Fonseca, Anna Formica, Andrew Frank, Mark Gahegan, Helmar Gust, Stephen Hirtle, Eva Klien, Alexander Klippel, Kai-Uwe Kühnberger,

Michael Lutz, David Mark, Dan Montello, Catharina Riedemann, Andrea Rodríguez, and Stephan Winter.

References

O. Ahlqvist. A parameterized representation of uncertain conceptual spaces. *Transactions in GIS*, 8(4):493–514, 2004.

O. Ahlqvist. Using uncertain conceptual spaces to translate between land cover categories. *International Journal of Geographical Information Science*, 19(7): 831–857, 2005.

H.T. Bruns and M.J. Egenhofer. Similarity of spatial scenes. In J. M. Kraak and M. Moleenar, editors, *Proceedings of the 7th International Symposium on Spatial Data Handling (SDH'96)*, pages 173–184, Delft, The Netherlands, August 1996. Taylor and Francis.

C. d'Amato, N. Fanizzi, and F. Esposito. A semantic similarity measure for expressive description logics. In *CILC 2005, Convegno Italiano di Logica Computazionale*, Rome, Italy, 2005.

M. Egenhofer. Toward the semantic geospatial web. In *GIS '02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pages 1–4, New York, NY, USA, 2002. ACM.

M. Gahegan, R. Agrawal, T. Banchuen, and D. Dibiase. Building rich, semantic descriptions of learning activities to facilitate reuse in digital libraries. *International Journal on Digital Libraries*, 7(1):81–97, 2007.

P. Gärdenfors. *Conceptual Spaces - The Geometry of Thought.* MIT Press, Cambridge, MA, 2000.

D. Gentner. *The mechanisms of analogical learning. Similarity and analogical reasoning*, pages 197–241. Cambridge University Press, New York, NY, USA, 1989.

D. Gentner and A. B. Markman. Structural alignment in comparison: No difference without similarity. *Psychological Science*, 5(3):152–158, 1994.

R. L. Goldstone. Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20:3–28, 1994.

R. L. Goldstone and J. Son. Similarity. In K. Holyoak and R. Morrison, editors, *Cambridge Handbook of Thinking and Reasoning.* Cambridge University Press, 2005.

N. Goodman. Seven strictures on similarity. In *Problems and projects*, pages 437–447. Bobbs-Merrill, New York, 1972.

U. Hahn, N. Chater, and L. B. Richardson. Similarity as transformation. *Cognition*, 87:1–32, 2003.

B. Indurkhya. *Metaphor and Cognition.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992.

K. Janowicz. Kinds of contexts and their impact on semantic similarity measurement. In *5th IEEE Workshop on Context Modeling and Reasoning (CoMoRea) at the 6th IEEE International Conference on Pervasive Computing and Communication (PerCom'08)*, Hong Kong, March 2008. IEEE Computer Society.

K. Janowicz, C. Keßler, M. Schwarz, M. Wilkes, I. Panov, M. Espeter, and B. Baeumer. Algorithm, Implementation and Application of the SIM-DL Similarity Server. In *Second International Conference on GeoSpatial Semantics (GeoS 2007)*, number 4853 in Lecture Notes in Computer Science, pages 128–145, Mexico City, Mexico, 2007. Springer.

C. Keßler, M. Raubal, and K. Janowicz. The effect of context on semantic similarity measurement. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshop SWWS 2007*, number 4806 in Lecture Notes in Computer Science, pages 1274–1284, Vilamoura, Portuga, 2007. Springer.

W. Kuhn. Semantic reference systems. *International Journal of Geographic Information Science (Guest Editorial)*, 17(5):405–409, 2003.

W. Kuhn and M. Raubal. Implementing semantic reference systems. In Stphane Coulondre Michael F. Gould, Robert Laurini, editor, *6th AGILE Conference on Geographic Information Science*, pages 63–72, April 24-26, 2003; Lyon, France, 2003.

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

B. Li and F. Fonseca. Tdd - a comprehensive model for qualitative spatial similarity assessment. *Spatial Cognition and Computation*, 6(1):31–62, 2006.

A. Maedche and S. Staab. Measuring similarity between ontologies. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, number 2473 in Lecture Notes in Computer Science, pages 251 – 263. Springer, 2002.

D. Medin, R. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993.

R. Möller, V. Haarslev, and B. Neumann. Semantics-based information retrieval. In *IT&KNOWS-98: International Conference on Information Technology and Knowledge Systems*, pages 49–56, Budapest, Hungary, September 1998.

K. Nedas and M. Egenhofer. Spatial similarity queries with logical operators. In T. Hadzilacos, Y. Manolopoulos, J. Roddick, and Y. Theodoridis, editors, *SSTD '03 - Eighth International Symposium on Spatial and Temporal Databases*, volume 2750 of *Lecture Notes in Computer Science*, pages 430–448. Santorini, Greece, 2003.

R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19:17–30, 1989.

M. Raubal. Formalizing conceptual spaces. In A. Varzi and L. Vieu, editors, *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, volume 114 of *Frontiers in Artificial Intelligence and Applications*, pages 153–164. IOS Press, Amsterdam, NL, 2004.

M. Raubal. Representing concepts in time. In N. Newcombe C. Freksa and P. Gärdenfors, editors, *Spatial Cognition VI - Proceedings of the International Conference Spatial Cognition 2008*, Lecture Notes in Artificial Intelligence, Freiburg, Germany, forthcoming 2008. Springer, Berlin.

P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453, 1995.

E. L. Rissland. AI and similarity. *IEEE Intelligent Systems*, 21(3):39–49, 2006.

A. Rodríguez and M. Egenhofer. Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, 18(3):229–256, 2004.

A. Schwering. Approaches to semantic similarity measurement for geo-spatial data - a survey. *Transactions in GIS*, 12(1):5–29, 2008.

A. Schwering and W. Kuhn. A hybrid similarity measure for spatial information retrieval. *Journal of Spatial Cognition and Computation. Lawrence Erlbaum Associates*, forthcoming.

A. Schwering and M. Raubal. Spatial relations for semantic similarity measurement. In J. Akoka, S. Liddle, I-Y. Song, M. Bertolotto, I. Comyn-Wattiau, W-J. vanden Heuvel, M. Kolp, J. Trujillo, C. Kop, and H. Mayr, editors, *Perspectives in Conceptual Modeling: ER 2005 Workshops CAOIS, BP-UML, CoMoGIS, eCOMO, and QoIS.*, volume 3770 of *Lecture Notes in Computer Science*, pages 259–269. Springer, Klagenfurt, Austria, October 2005.

R. N. Shepard. Toward a universal law of generalization for psychological science. *Science*, pages 1317–1323, 1987.

W. S. Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30: 379–393, 1965.

A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.