Conceptual Spaces for Data Descriptions

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

Abstract

Cognitively adequate information representations have a great potential to improve interactions between systems and users. They are able to account for a specific user's personal understanding, knowledge and preferences. To utilize this capacity for the personalization of services, we need to represent information on a level that corresponds to human cognition. In this paper, we show how to generate such a description in the form of a conceptual space from an existing landmarks database. We then analyze this process for future automation and generalization.

Introduction

People's interaction with their environment has been an important research topic for a long time. It has been worked on from different perspectives such as activities and behaviour in space, or navigation on different scales and in different modes (e.g. [5], [12], [2]). Cognitively adequate information enables improvements for systems which assist users in performing spatio-temporal tasks in their environment. The representation of information on a level which corresponds to human cognition bears great potential for the personalization of services, and for the integration of semantic descriptions.

The aim to assure a common understanding of the semantics during information exchange has often resulted in fixed interpretations of terms, regardless of the fact that the parties involved might have a different understanding of them. As a result, services such as navigation assistants or location based services provide all users with the same output, irrespective of whether these are useful for a specific user or not.

The approach of cognitive semantics accounts for the fact that the understanding of a term is dependent on an agent's knowledge and context. In contrast to the *realist* approach, where "the meaning of a word or expression is something out there in the world" and "similarity is something that exists objectively [...], independent of any perceptual and other cognitive processes" (pages 151 and 110 in [3]), the *cognitive* view is capable of explaining phenomena such as learning or the change of concepts over time. From a system engineer's point of view, cognitive semantics can be useful in the design of geospatial services, which can adapt to a specific user. To utilize this potential, cognitive descriptions are required both for the users and the system. *Conceptual spaces* provide spatial representations of concepts from the cognitive perspective. They arrange concepts as points in a vector space. This space is spanned by *quality dimensions* which correspond to the properties of the concepts.

The prospect of this paper is to demonstrate how to derive a conceptual space from a given data source. As a case study, a landmark selection scenario in the city of Vienna is used. A database with 58 buildings in Vienna's first district is analyzed and converted into a cognitively adequate representation, namely a conceptual space as formally defined in [1]. Although this research focuses on the first steps towards user-adapting services based on cognitive adequate data

representations, it also provides insight on the differences between human perception of space and the according data representations. Both the use case and the data presented have been developed and collected in [8] to investigate the selection of landmarks for navigation.

In the following, we will first present conceptual spaces. We will then show how to define a conceptual space, and go through the process of creating a conceptual space for a database containing landmarks. This will be followed by an analysis of the extraction process, focussing on automation. We will conclude with an outlook on future work.

Conceptual Spaces

Conceptual spaces have been introduced as a framework for representing information on the conceptual level [3]. Gärdenfors identifies three different levels for representing information in cognitive science: According to the *symbolic* approach cognition is symbol manipulation. Methodologies currently used for the semantic web are based on this approach [4]. The *associationist* approach puts the stress on the associations between symbols, as in artificial neural networks, for example. The third one is the *conceptual* approach, which will be elucidated in the following.

A conceptual space represents the properties forming a concept as quality dimensions with a geometrical or topological structure. Hence, they span a vector space with concept instances being points (i.e., vectors) in that space which take a value for every quality dimension. The calculation of similarity values is based on the inverse distances between the vector representations of the concepts. Depending on the kind of dimension, different metrics apply for the distance calculations. Gärdenfors distinguishes *integral* and *separable* dimensions. A group of integral dimensions is characterized by the fact that one needs to assign a value to every one of them to completely describe the concept (e.g. hue, saturation and value of a color). For integral dimensions, Euclidian metric is usually applied. Separable dimensions, in contrast, can stand alone, such as the height or the age of a building, and apply city block metric[1]. The distinction into these two groups of dimensions stems from the way humans perceive their environment: Integral dimensions are processed holistically, whereas separable dimensions are processed analytically [7]. A set of integral dimensions that is separable from all other dimensions is called a *domain*. To reflect the saliency of particular dimensions, individual weights can be assigned to all dimensions [9] — e.g. in the concept "landmark", the height and color of a building are more important than the number of people living in it. These weights can be task-dependent.

Aisbett and Gibbon (2001) present a general formulation of conceptual spaces [1]. The authors formalize conceptual spaces as a meso level representation embedded between the higher-level symbolic representations (the realist approach) and the lower-level network representations (the associationist approach). This formalization is especially useful for the given task because it explicitly links these three levels of representation to each other. Hence, it provides useful hints on how to infer the conceptual space for the given landmark database, and the results can be checked for coherence with the model.

Criticism on conceptual spaces, as on other geometric models for concept representation such as multidimensional scaling, is based on violations of the basic metric axioms — i.e. minimality, triangle inequality, and especially symmetry [13]. Since this paper develops a conceptual space for a given data set, minimality cannot be violated. Two landmarks can only be identified as identical (similarity = 1, distance = 0 respectively) by the system if a landmark is compared to itself[2]. Based

[1] It has been shown that other kinds of Minkowski metrics may provide improved descriptions of the perceived similarity [6].

[2] However, it must be borne in mind that the user might confuse two landmarks with a high similarity value. For the given scenario, this can be handled by a spatial buffer to make sure that for a given landmark, there is no confusable other landmark within the area of sight.

on the same assumption, i.e. that a computer cannot confuse similar things, the triangle inequality necessarily holds. Even if there are two identical buildings, they are still distinguishable from each other by their location, given by the street name and number in the scenario. The violation of symmetry can easily be shown in subject tests with directed tasks, for example Mexico is usually rated to be more similar to the USA than vice versa. Different similarity values for the two directions of a comparing task stem from the fact that one concept is more prominent than the other one. Hence, extensions have been developed for conceptual spaces that introduce bias values to reflect the prominence of a concept; for an overview, see [6], chapter five. For simplicity, symmetry will be taken for granted below, following the assumption that the user working with the system does not know any of the landmarks in his environment and is therefore not biased concerning prominence.

The System Space

Though conceptual spaces were initially developed to represent the human understanding of terms, they can also be utilized to describe the data and services a machine offers [9]. Representing a data source as a concept improves techniques for service and data source discovery. Current query techniques only find information which exactly matches given criteria, usually provided as keywords. This is due to the fact that current descriptions are mainly based on metadata and ontologies, which reduce relations among concepts to *is-a* relationships (such as "a cathedral is a church is a building"). With these descriptions, similarity measurement is only indirectly possible, through calculations on the ontological tree structure — if at all. Conceptual spaces allow for the discovery of *sufficiently similar* sources of information by calculating similarity values based on the properties of an object. For example, when a tourist searches for a museum for medieval art at his destination, but there is no such museum, the search engine could provide him with historic buildings from the same era. Those alternative results were then based on a high similarity between the query concept and the results, e.g. on the dimensions for age and attractiveness for tourists. Although they do not match the query exactly, they provide valuable information. Beyond that, conceptual spaces, once defined for a service and a user, could improve the personalization of the service. A service which is aware of a specific user's conceptual space knows which information is valuable for that user, and which is useless. Imagine a traveling architect, for example, with his personal device that is aware of his interests and the special knowledge he has due to his profession: His device can select landmarks for navigation based on architectural features that do not stand out for lay-persons, and notify him of architecturally interesting sights on his way. This personalization can even be fine-tuned to special eras or styles, reflecting the detailed preferences by weights on the according dimensions. The high level of individual adaptation is possible because the device acts on a level of information representation that corresponds to the user's understanding.

Conceptual spaces are supposed to be especially useful in a spatial context because we perceive space directly. We see the world when we are walking outside or driving a car, and we look at a representation of the real world when we use a map. It is this perception that distinguishes conceptual spaces from other semantic descriptions such as ontologies: while ontologies show how symbols are related to each other, without being anchored in the real world, conceptual spaces are based on people's perceptions as the fundamental quality dimensions. Other, more complex quality dimensions may build upon them, but human perception is what ties conceptual spaces to the real world.

According to the formal definition provided in [1], a conceptual space includes the following elements (summarized; see the original paper for the exact, extensive definition):

1. A base conceptual space with a distance metric and a betweenness relation
2. A concept space and a symbol space
3. A set of dimensions, composed of subsets of integrate dimensions (domains) and separable dimensions

Beyond that, the definition includes an attention buffer and copies of the base conceptual space, specifying levels. The attention buffer formalizes the process of highlighting a region in a conceptual space because of its importance for a given task. Striving for a conceptual space representing the given landmark database independent of a specific task, this aspect will be ignored in the following. The copies of the base conceptual space provide a representational solution for the *binding problem*: In a complex concept, these levels specify which property refers to which sub-concept. As the case study does not include complex concepts, this part will also be omitted[3]. The process that will be analyzed can be summarized as shown in figure 1.

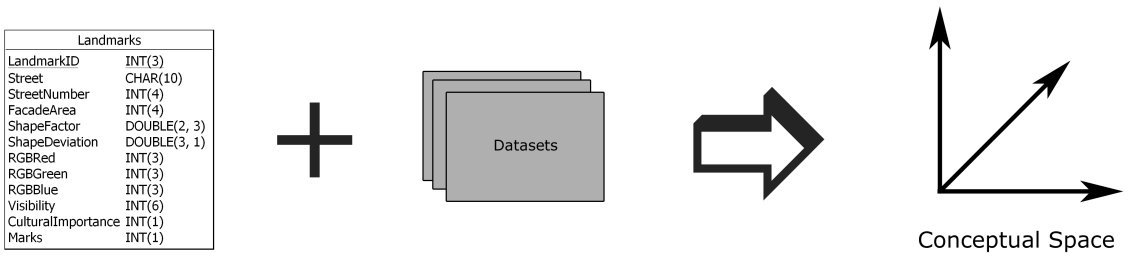

Figure 1: The database schema and the set of 58 landmarks are used to define the conceptual space.

Specifying the Database's Conceptual Space

So far, we have pointed out the potential usages of conceptual spaces, focusing on applications with a spatial context. However, hardly any of the above techniques have been implemented or tested so far. To come to working results, we will next go through the process summarized in figure 1. The results will be described informally; for a general formalization of the derivation process, which is in line with the formal definition of conceptual spaces in [1], further research is required.

The starting point for the definition of the conceptual space for the landmark database is the symbol space given by the database schema. Every field in the database stands for a property (i.e. the property's name), and every dataset assigns specific values to the properties, describing the landmarks on the symbolic level (see table 1 for an extract). In the first step, we generate dimensions according to the fields of the database. We use the data type defined by the database schema and the measurement scale [11] of every field to assign the appropriate metrics and betweenness relation to every dimension. Note that the measurement scales are not defined by the schema; additional knowledge about the semantics of the fields is required to infer them. Table 2 shows an overview of all database fields with data types, descriptions and measurement scales. For instance, the values in the field *Visibility* are on the ratio scale, whereas the entries for *ID* are on the nominal

[3]Moreover, the formation of complex concepts requires a detailed understanding of the underlying basic concepts. Since these are still a research topic with many open questions, they should be addressed first before investigating complex concepts in the future.

scale, although both of them are stored as integers. To come to this conclusion, we must know that *Visibility* refers to the size of the area from which the landmark is visible, and that an area is on the ratio scale because it allows for the determination of the equality of ratios. The values for *ID*, however, are arbitrarily assigned identification numbers, which are not ordered and thus only allow for determination of equality. Accordingly, *Cultural Importance* is on the ordinal scale, because a landmark with value 2 is more important than another one with value 1, but it does not make sense to calculate differences. The measurement scales for the remaining fields are derived correspondingly[4].

ID	Street	No	Façade Area	Shape Factor	Shape Deviation	RGB Red	RGB Green	RGB Blue	Visibility	Cultural Importance	Marks
24	Stephansplatz	1	1266	0,752	33	91	96	109	4738	3	0
25	Stephansplatz	4	679	0,94	0	154	164	184	2190	1	2
26	Stephansplatz	5	1702	0,755	0	173	184	205	4578	1	1
27	Stephansplatz	6	2279	0,735	0	154	167	194	2803	1	1
28	Stephansplatz	1	3309	1,27	57,7	52	55	64	11051	3	0
29	Stephansplatz	4	763	0,798	0	130	139	162	4398	1	1
30	Churhausgasse	1	819	0,712	0	105	107	122	1853	1	1
31	Churhausgasse	2	1035	0,824	0	113	115	133	1058	2	1
32	Stephansplatz	3	2296	0,693	0	123	129	151	3832	2	0

Table 1: Extract from the landmarks database.

To derive the appropriate metrics for a dimension, we also need to know whether it is a separable dimension, or if it is integrate with other dimensions, forming a domain. The three dimensions for the façade color domain in RGB mode combine using Euclidian metrics, because all three dimensions are required to completely describe the color. Thus, the dimensions are integrate. The remaining dimensions are separable from each other and should consequentially be assigned city block metrics; however, this is not always possible. Looking at the measurement scales, we see that the dimensions for the database fields with a nominal scale cannot be used to calculate any distances. Instead, we have to fall back on a Boolean metrics, which only allows us to specify whether two landmarks are identical on the according dimensions, e.g. whether they are in the same street. Similarity values other than 0 and 1 are not possible. Note that it is only possible to use city block metrics on the dimensions for the fields with an ordinal scale because they are already expressed in numbers from 0 to 3. If they were identified by keywords (such as “no marks”, “used commercially”, “commercially used by a well-marked venue” etc. for the field *Marks*), it would be necessary to assign numbers to them, which reflect the order of the original values [10].

The betweenness relation for every dimension required for the complete definition of the conceptual space is strongly related to the metrics. For the dimensions combining with Euclidian or city block metrics, betweenness is implied in the metrics. For the dimensions applying Boolean metrics, which stem from the database fields on the nominal scale, this is not possible. The first condition for the betweenness relation *B(a,b,c)* requires the variables a, b and c to take different values ([1], p. 199), which is not possible with Boolean metrics, providing only two distinct values. Hence, the betweenness relation on these dimensions is empty. However, it is still possible to determine whether a concept representation is between two others, using only those dimensions which allow for the computation of betweenness.

[4]There is no field with interval scale in this database; *year of construction* would be an example.

Field Name	SQL Data Type	Description	Measurement Scale
ID	INT	Unique ID	Nominal
Street	CHAR	Street name	Nominal
No	INT	Street number	Nominal
Façade Area	INT	Façade area in m^2	Ratio
Shape Factor	DOUBLE	Proportion of height to width of façade	Ratio
Shape Deviation	DOUBLE	Deviation from rectangular shape	Ratio
RGB Red	INT	RGB value for red	Ratio
RGB Green	INT	RGB value for green	Ratio
RGB Blue	INT	RGB value for blue	Ratio
Visibility	INT	Size of the area from which the façade is visible in m^2	Ratio
Cultural Importance	INT	Four ordered classes with increasing importance	Ordinal
Marks	INT	Four ordered classes with increasing recognizability	Ordinal

Table 2: Field names from database schema with according data types and measurement scales.

Automating the Process

To utilize data descriptions based on conceptual spaces, automating the process described in the previous section is desirable. Particularly, this would allow for the generation of cognitively adequate representation for existing data without going through a cumbersome manual process. However, as the analysis of the process to describe a given landmarks database has shown, it is not possible to completely automate this process. Manual intervention is required at several stages to integrate additional knowledge, which cannot be retrieved from the database schema or the datasets.

In the first step, we created the dimensions of the conceptual space according to the fields of the database. We used information on the measurement scales and on integrate and separable quality dimensions to define the appropriate metrics for these dimensions; neither of them can be computed from the database schema or the datasets without additional knowledge. Without knowing about the semantics of the fields, it is not possible to determine the measurement scale. The same applies for the specification of separable and integrate quality dimensions. Looking at the three fields defining the RGB color value for the landmarks' façades, there is no hint in the database schema on their semantic relationship. Since the three values are numerically independent, statistical correlation analysis cannot reveal their semantic dependence either. Consequently, this part of the process is heavily relying on external input.

Once the measurement scales and the integrate and separable quality dimensions have been specified, the appropriate metrics can be assigned automatically. Separable dimensions apply city block metrics, whereas integrate dimensions apply Euclidian metrics. If this is not possible due to the fact that the according database field's values are on the nominal scale, Boolean metrics must be applied instead. The betweenness relation is then implied in the metrics, or empty in the case of Boolean metrics.

Conclusions and Future Work

We have outlined some initial ideas on the use of conceptual spaces for data and service descriptions. Beyond small examples of how to utilize them in the geospatial domain, we have used the case of a small database with landmarks in the city of Vienna to show what is necessary to derive a conceptual space representation from such a model. It was demonstrated that some informa-

tion can be extracted, while other parts are merely derivable without interpretation and addition of external information. This leads to the basic conclusion that generating conceptual spaces from descriptions that are settled on the symbol level, comparable to a database, cannot be completely automated. It must also be noted that the fields in the database example are very closely related to properties humans can perceive directly. It can be assumed that the process gets more complicated with increased abstractness and complexity of the data model.

Future work should focus on developing best practices in how to generate conceptual spaces for existing data sources and services. The results of this paper should be used to specify a process for the derivation of a conceptual space from an existing database. This process should be formally in line with the definition of a conceptual space in [1]. Beyond that, it should especially focus on automating the derivation as far as possible and assisting the user when adding external information. In this context, it needs to be investigated whether external data sources are useful for the further automating. We should then strive for the generalization of the process to be applicable on other kinds of symbolic information representations, and finally evaluate the results in a case study with human subjects tests.

Acknowledgements

Special thanks go to Martin Raubal and the two anonymous reviewers for their valuable comments.

References

[1] Janet Aisbett and Greg Gibbon. A general formulation of conceptual spaces as a meso level representation. *Artificial Intelligence*, 133:189–232, 2001.

[2] Reginald G. Golledge. Place recognition and wayfinding: Making sense of space. *Geoforum*, 23:199–214, 1992.

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

[4] Peter Gärdenfors. How to make the semantic web more semantic. In A.C. Vieu and L. Varzi, editors, *Formal Ontology in Information Systems*, pages 19–36. IOS Press, 2004.

[5] Torsten Hägerstrand. What about people in regional science? *Papers of the Regional Science Association*, 24:7–21, 1970.

[6] Mikael Johannesson. *Geometric Models of Similarity*. Lund University Coginitve Studies 90. Lund University Congnitive Science, Lund, Sweden, 2002.

[7] Mikael Johannesson. The problem of combining integral and separable dimensions. *Lund University Cognitive Studies, Lund, Sweden*, 87, 2002.

[8] Clemens Nothegger, Stephan Winter, and Martin Raubal. Selection of salient features for route directions. *Spatial Cognition and Computation*, 4(2):113–136, 2004.

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

[10] Angela Schwering and Martin 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*, Lecture Notes in Computer Science, pages 259–269, Klagenfurt, Austria, 2005. Springer, Berlin.

[11] S.S. Stevens. On the theory of scales of measurement. *Science*, 103(2684):677–680, 1946.

[12] Sabine Timpf, Frank Ostermann, Andreas Lusti, and Lorenz Boeckli. Claiming personal space in public parks. In Martin Raubal, Harvey J. Miller, Andrew U. Frank, and Michael F. Goodchild, editors, *GIScience 2006 Abstracts*, Münster, 2006. ifgi Prints.

[13] Amon Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.