The Relevance of Spatial Relation Terms and Geographical Feature Types

Chunju Zhang, Xueying Zhang, and Chaoli Du

Key Laboratory of Virtual Geography Environment (Nanjing Normal University), MOE, Nanjing, China
zcjtwz@sina.com, zhangsnowy@163.com

Abstract. Spatial relation terms can generally indicate spatial relations described in natural language context. Their semantic representation is closely related to geographical entities and their characteristics e.g. geometry, scale and geographical feature types. This paper proposes a quantitative approach to explore the semantic relevance of spatial relation terms and geographical feature types in text. Firstly, a classification of spatial relation terms is performed. Secondly, the "Overlap" similarity measure is introduced to define the relevance of spatial relation terms and geographical feature types based on a large scale annotation corpus. Thirdly, the relevance is expanded with the semantic distance and hierarchical relationship of the classification system of geographical feature types. Finally, a knowledge base based on protégé is developed to formally represent and visualize geographical feature types, spatial relation classifications, and the relevance of spatial relation terms and geographical feature types. This study indicates that spatial relation terms are strongly relevant to geographical feature types. The semantic representation of topological relation terms is diverse and their relevance with geographical feature types is much stronger than directional relation and distance relation terms, but the annotation quality and the classification granularity of geographical entities in the corpus have a great effect on the performance.

Keywords: spatial relation, geographical feature type, spatial relation term, relevance.

1 Introduction

Natural language describes the nature of people's internal representation of space and is the primary means for representation and exchange of geographical information, such as geographical entities, spatial relations, etc. Spatial relations are the associations or connections between different real world features, and play an important role in spatial data modeling, spatial query, spatial analysis, spatial reasoning, and map comprehension [1]. The semantic research of spatial relations is the premise and basis for the description and expression of spatial relations. Spatial relations have been in a high priority in many research fields, such as linguistics, cognitive science, GIS and spatial reasoning. The linguistics field focus on the words, lexical, syntactic and semantic structure of spatial relation expressions, and the relationship with

T. Washio and J. Luo (Eds.): PAKDD 2012 Workshops, LNAI 7769, pp. 47–56, 2013.

human's spatial cognition [2][3]. In recent years, spatial relations in natural language have become a hot topic of geographical information science. Mark [4] and Lautenschütz [5] investigated the influence of geometry and scale characteristics, spatial relation types and geographical feature types on human's chosen of spatial relation terms by questionnaire method, and then Mark [6] made a further research on the mapping between spatial relation terms and GIS computational model. Shariff [7] and Xu [8] summarized the knowledge rules of different geographical feature types and spatial relation vocabularies to construct a semantic mapping model of spatial relation terms. Du and Wang [9] explored the formal expression of GIS querying sentences described in a restricted syntactic pattern of spatial relation descriptions in natural language.

Spatial relation terms can indicate spatial relations described in natural language context. Different from the early models of spatial relations which focused on the geometry, it is now widely recognized that the semantic meaning of spatial relation terms is also dependent on functional and pragmatic features in situated context [5]. Their semantic descriptions in natural language are closely related to geographical entities and their characteristics of geometry, scale and geographical feature types. Especially, some spatial relation terms can be used for several different geographical feature types, while some are just for a certain geographical feature type. For example, the spatial relation term of watershed is used to indicate the junction between mountains and waters, and cannot describe geographical entities of other geographical feature types. This paper proposes a quantitative approach to explore the relevance of spatial relation terms and geographical feature types from text corpuses and the classification system of geographical feature types. Properly understanding the semantic meaning of spatial relation terms in text will improve geographical information retrieval, GIS natural language query, extraction of spatial relations from text, and qualitative spatial reasoning.

The reminder of this paper is structured as follows: Section 2 investigates the basic categories of spatial relations in natural language, and the classification of spatial relation terms. Section 3 discusses the calculation of relevance of spatial relation terms and geographical feature types based on Corpus and geographical feature types. The semantic knowledge expression of spatial relation terms based on Ontology is in section 4. The conclusion and future work are given in Section 5.

2 Classification of Spatial Relation Terms

Spatial relations are considered to be one of the most distinctive aspects of spatial information. According to Egenhofer and Franzosa's argument, spatial relations can be grouped into three different categories of topological relations, direction relations and distance relations. Natural-language spatial relations are spatial relations described in natural language among people's daily communication, it is much closer to people's habit than GIS spatial relations [10]. For example, the description of "Yangtze River is across Nanjing city in the northwest, and is 10 kilometers from XinJieKou Shop", there are a topological and direction relation between Yangtze

River and Nanjing city, and a distance relation between Yangtze River and XinJieKou Shop. The description and expression forms of spatial relations in natural language and GIS are very different. Spatial relations in natural language are richer, but with a qualitative, fuzzy, uncertainty and unstructured characters, while spatial relations in GIS are quantitative, structured, and accurate. Topological relations have long been considered as the most important spatial relations in GIS while direction and distance relations are with the highest using frequency in people's daily life. Spatial relations in natural language are expressed through a series of spatial relation terms. In different language, those terms are with different diversity and complexity. Taking the spatial relation term of "crossing" in English for example, in Chinese it can be expressed as "穿越(chuanyue, crossing)", "交叉(jiaocha, crossing)", "横贯(hengguan, crossing)", etc. Meanwhile, some spatial relation terms in Chinese indicate more than one spatial relation type. For example, the spatial relation term "北靠(beikao, north and near)" not only expresses the north direction but also implies a topological relation of extended connection. In addition, there are some spatial relation terms in text descriptions whose semantic meanings cannot be expressed with existing calculation models of GIS spatial relations. Taking the spatial relation term of "支流(zhiliu, tributary)" for example, it may describe an including relation of the main vein and tributaries of a river, however, this semantic relation cannot be expressed in GIS spatial relation models.

Region connection calculus (RCC) model takes geographical entities in the real world as a region and describes spatial relations with the region connectedness [11]. Therefore, it is in accordance with human's cognition habit and more suitable for qualitative representation and reasoning of spatial relations. The ternary point configuration calculus (TPCC) describes directions such as front, back, left and right [12]. Distance relations specify that how far the object is away from the reference object. Based on RCC8, TPCC, and the frequency of spatial relation terms in natural language context, basic categories of spatial relations and classifications of spatial relation terms are described in Table 1.

From table 1, we can see that one spatial relation category may include multi-spatial relation terms, and one spatial relation term may correspond to more than one spatial relation categories. Also, there are some commonly used spatial relation terms which cannot be clustered into these categories, such as between, round, etc. Here it should be noted that this paper only discusses a binary instance of spatial relations between two geographical entities, not consider the composite spatial relations. For some compound spatial relation terms, the classification will be determined by the last direction word, such as "中南部 (zhongnanbu, central south)" with a direction of south. Also, there are some connected words which cannot reflect topological or direction relations but provide the connection between the source and target objects. So they play a role in auxiliary judgments of spatial relations, such as "located", "is", "as", "with", "by", etc.

Table 1. Samples of classifications of spatial relation terms

Spatial Relations	**Spatial Relation Terms**
Topological relation	
…..IN(tangential and non-tangential proper parts)	包含(baohan, including), 属于(shuyu, belong to)
…..EC(extended connection)	相接(xiangjie, touch), 流入(liuru, flow into)
…..DC(discrete connection)	相离(xiangli, discrete connection), 相距(xiangju, apart)
…..PO(Partially overlap)	贯穿(guanchuan, run through), 交叠(jiaodie, overlap)
…..EQ(equality)	相等(xiangdeng, equal), 别名(bieming, alias)
Directional relation	
Relative direction	
…..F(front)	前头(qiantou, front), 前部(qianbu, anterior)
…..BE(behind)	后端(houduan, back-end), 后面(houmian, behind)
…..L(left)	左边(zuobian, left side), 左面(zuomian, left)
…..R(right)	右边(youbian, right), 右端(youduan, right)
…..A(above)	上端(shangduan, above), 上面(shangmian, above)
…..BW(below)	下端(xianduan, below), 下(xia, below)
…..INT(inner)	内(nei, in), 内部(neibu, inner), 里面(limian, inside)
…..EXT (exterior)	外(wai, outer), 外部(waibu, exterior), 外头(waitou, outside)
Absolute direction	
…..E(east)	东方(dongfang, east), 东端(dongduan, east), 东(dong, east)
…..W(west)	西端(xiduan, west), 西部(xibu, west), 西(xi, west)
…..S(south)	南部(nanbu, south), 南(nan, south), 南方(nanfang, south)
…..N(north)	北面(beimian, north), 北方(beifang, north), 北(bei, north)
…..C(centre)	中部(zhongbu, middle), 中心(zhongxin, center)
…..NE(northeast)	东北面(dongbeimian, northeast),东北方(dongbeifang, northeast)
…..SE(southeast)	东南边(dongnanbian, southeast), 东南方(dongnanfang, southeast)
…..NW(northwest)	西北(xibei, northwest), 西北部(xibeibu, northwest)
…..SW(southwest)	西南部(xinanbu, southwest), 西南边(xinanbian, southwest)
Distance relation	距离(juli, distance), 相离(xiangli, distance), 相距(xiangju, apart from)

3 Calculation Method of Relevance

3.1 Calculation Based on Corpus

A binary spatial relation could be formalized as < geographical entity A, spatial relation terms, geographical entity B > in natural language context. Obviously, one spatial relation term should associate with a pair of geographical entities. For the concept characteristics of geographical entities could be defined by the type of geographical features, therefore, a single spatial relation can be further abstracted as < feature type of geographical entity A′ , spatial relation term, feature type of geographical entity B′ >. There is an order for target and reference objects in spatial relation descriptions. To simplify the calculation, this order between A and B is not distinguished in this paper. In linguistics, a text corpus is a large and semi-structured set of texts which are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. This paper takes the large scale annotation corpus (Geocorpus) of spatial relations of "Chinese Geography Encyclopedia" from paper [13] as an experimental data, and summarizes 600 commonly used spatial relation terms. Overlap is a classic calculation method for semantic relations, and it is based on the co-occurrence frequency of two events in a data set [14]. Therefore, the relevance of spatial relation terms and the type of geographical entities based on Geocorpus can be defined as in formula 1.

$$R(T, A', B') = \frac{|T \cap A'B'|}{\min(|T|, |A'B'|)} \quad (1)$$

In formula 1, T represents the occurrences of a spatial relation term in the Geocorpus, A ' and B' denote the occurrence of two geographical feature types, R indicates the relevance degree between T and a pair of A 'and B'. Taking the spatial relation term of "流入 (liuru, flow into)" and "北部(beibu, north)" as example, the results of the relevance are just as shown in Table 2.

Table 2. Relevance of spatial relation terms and geographical feature types based on corpus

Spatial Relation Terms T	Geographical Feature Types A′	Geographical Feature Types B′	Relevance
流入 (liuru, flow into)	river	ocean elements	0.8333
	river	lake	1.0000
	river	river	0.1333
北部 (beibu, north)	resident	natural landscape	0.1428
	resident	river	0.0714
	Natural landscape	natural landscape	0.0444
	natural landscape	Lake	0.2222

The annotation and experiment result shows that spatial relation terms are strongly relevant to geographical feature types. The term of “北部(beibu, north)” is related to resident, river, natural landscape, and the other geographical feature types, while the term of “流入(liuru, flow into)” is just in co-occurrence with ocean, lake and river in the corpus. However, some of the relevance has a higher R-value, and some is lower. This is because that there is a natural phenomenon of imbalance of geographical concepts in the real world. Text is a main expression vector of people’s cognition from real world, so the geographical concepts in the Geocorpus are not in an imbalance. Meanwhile, some geographical concepts in the Geocorpus have a coarse granularity. Therefore, the R value is higher when the phenomenon is more common.

3.2 Calculation Based on Geographical Feature Types

There is a level and hierarchical relationship in the classification system of geographical feature types. It could be seen as a semantic network diagram. In this diagram, each node represents geographical feature types, edges indicate their relationship, and the weights of edges represent the semantic distance of geographical feature types. With this distance, the semantic relation and relevance between geographical feature types can be analyzed. Based on the relevance from corpus calculation, a quantitative approach to expand the relevance of spatial relation terms and geographical feature types from the classification system of geographical feature types is proposed.

In theory, a pair of geographical feature types with the father-son relationship has a higher semantic relation than brotherhood or nephew relationship. Terms which describe spatial relations between the father geographical feature types cannot describe their son geographical feature types. However, spatial relation terms for the son geographical feature types can also describe their father geographical feature types. Therefore, the semantic relevance between spatial relation terms and geographical feature types should be with a consideration of the semantic distance and the inheritance direction of geographical concept. Assuming that C1 and C2 represent a pair of geographical feature types respectively and their semantic relation in the classification system is R′, and α is the semantic relation value. The specific calculation rules are as follows:

- If C1 and C2 have a father-son relationship, then $R'(C1, C2) = \alpha$ ($0 < \alpha < 1$, the default value is 0.75); if C1 and C2 have a reverse relationship, then $R'(C1, C2) = 1$;
- If C1 is inherited indirectly from C2 among n concepts, then $R'(C1, C2) = \alpha^n$（$0<\alpha<1$，the default value is 0.75）; If C2 is inherited indirectly from C1 among n concepts, then $R'(C1, C2) = 1$;
- If C1 and C2 have a brother relationship, then $R'(C1, C2) = \alpha$（$0<\alpha<1$，the default value is 0.75）;
- Other relations are defined with the above composition.

With the above calculation rules, if the spatial relation term T is in co-occurrence with geographical feature types A ' and B' in the Geocorpus, then taking A 'and B' as a starting point and R as weight value to expand the semantic relevance between T and the pair of A ' and B'. For the hierarchical relationship of geographical feature types in the classification system, the relevance value is expanded with an iterative calculation. This calculation will stop until no new semantic relevance occurs.

With the 600 commonly used spatial relation terms and classification system of geographical feature types (GB/T 13923-2006), the relevance is a large net structure chart. Taking the term "流入 (liuru, flow into)" as an example, the relevance is as in Figure 1.

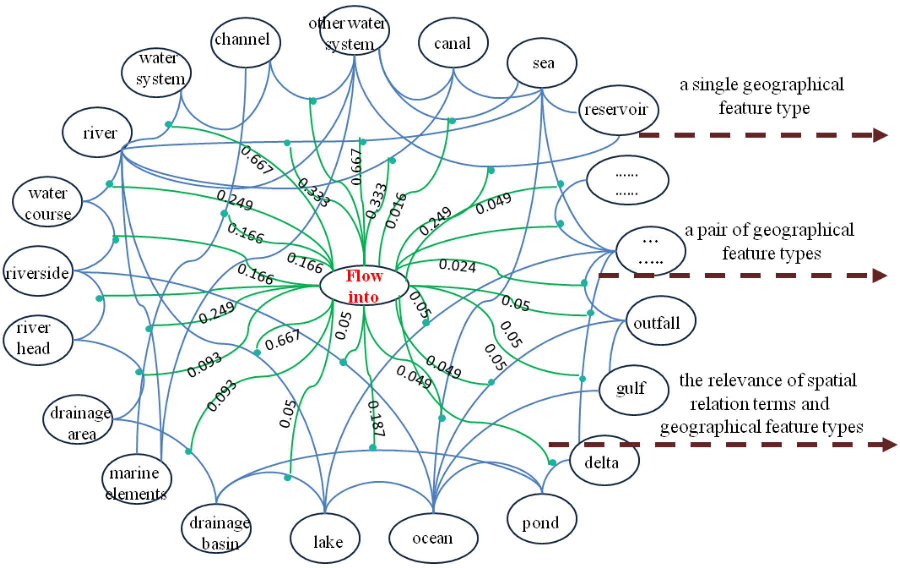

Fig. 1. The relevance of "流入 (liuru, flow into)" and geographical feature types

In the experiment, the R-value stands for the relevance degree between spatial relation terms and geographical feature types. With the classification system of geographical feature types, the relevance value is expanded. However, some of the relevance has a higher R-value, and some is lower than 0.05. In order to keep the balance of the relevance we can set and adjust a threshold to filter the uncommon relevance in a text corpus, such as 0.05 for "流入 (liuru, flow into)". Then the term "流入 (liuru, flow into)" only describes spatial relations of geographical entities of water system, such as river, ocean and lake. As we all know, there are a lot of spatial relation terms to describe spatial relations of river, ocean, lake, etc, however, people are used to choosing "流入 (liuru, flow into)" to describe them in daily life. Therefore, this result comparatively conforms to people's language and cognitive habit. In addition, the relevance of topological relation terms and geographical feature types are significantly stronger than directional relation and distance relation terms.

4 Semantic Knowledge Expression Based on Ontology

Ontology formally represents rich knowledge as a set of concepts and the relationships between those concepts within a domain. It can improve the consistency, accuracy, reusability and sharing features of knowledge to understand and use. In this paper, a knowledge base is developed based on protégé to formally represent and visualize geographical feature types, spatial relation classifications, spatial relation terms and their relevance with geographical feature types (see Figure 2). Geographical feature types and spatial relation classifications are expressed with a class in OWL language, and the hierarchy relationship is established by the subClassOf. The ObjectProperty expresses the semantic relations between spatial relation terms and geographical feature types, and the quantitative constraint values are organized in DatatypePoperty. Then the relevance is defined a property with "rdfs: domain" and "rdfs: range" respectively, which can restrict the application fields and scope. Finally, instances of spatial relation terms can be made according to the semantic relevance of ObjectProperty and DatatypePoperty. This knowledge base can improve GIS natural language query, extraction of spatial relations from text, geographical information retrieval, qualitative spatial reasoning, etc.

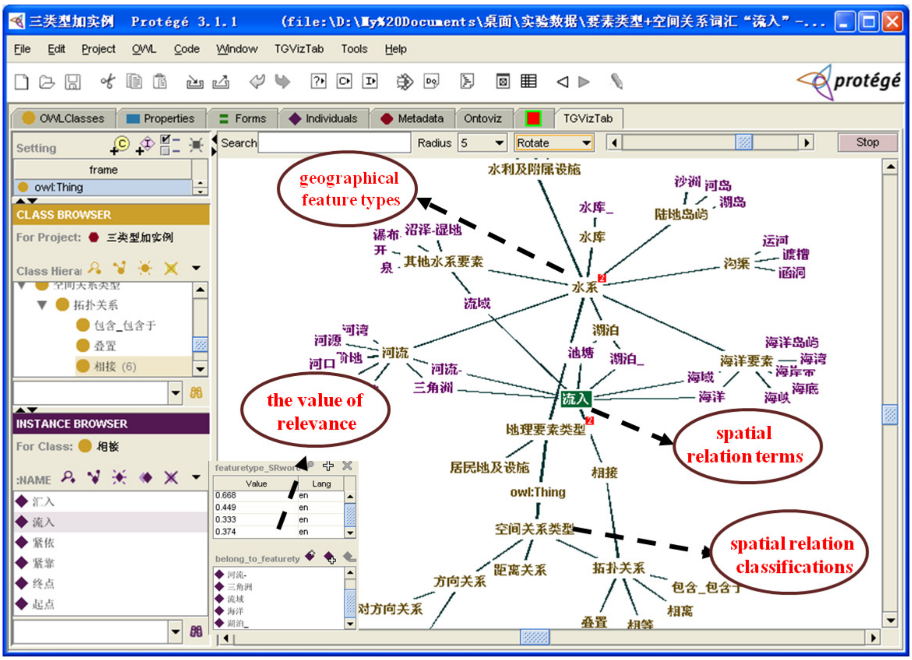

Fig. 2. Knowledge base of the relevance of spatial relation terms and geographical feature types

5 Conclusion

Based on a large scale text corpus and the geographical feature classification scheme this paper proposed a method to explore the relevance of spatial relation terms and geographical feature types. The experiment indicates that our proposed approach can effectively obtain meaningful results. However, the annotation quality of the corpus and the classification granularity of geographical entities have a great effect on the performance, especially for a general dataset. In our future work, we will start the classification on different kinds of texts describing the same kind of data (e.g. documents addressing only water, sea) in order to better extract relations specified for a particular domain. Moreover, to simplify the calculation, the order between geographical entities of A and B is notdistinguished in this paper. In addition to geographical feature types, geometric features and spatial scales of geographical entities also have an effect on spatial relations. Obviously, the semantic relevance of spatial relations and geographical feature types can be further improved with a comprehensive consideration of the description order, scale and geometric features in a further research.

Acknowledgement. This work was supported in part by the National Nature Science Foundation under grant number 40971231, and the Innovative Postgraduate Projects funded by Jiangsu province under grant number CXLX11_0874.

References

1. Egenhofer, M.J., Franzosa, R.: Point-Set Topological Spatial Relations. International Journal of Geographical Information Systems 5, 161–174 (1991)
2. Herskovits, A.: Language and Spatial Cognition: an Interdisciplinary Study of the Prepositions in English. Cambridge University Press, Cambridge (1986)
3. Levinson, S.C.: Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge (2003)
4. Mark, M.D., Comas, D., Egenhofer, M.J., et al.: Evaluating and Refining Computational Models of Spatial Relations through Cross-linguistic Human-subjects Testing. In: Kuhn, W., Frank, A.U. (eds.) COSIT 1995. LNCS, vol. 988, pp. 553–568. Springer, Heidelberg (1995)
5. Lautenschütz, A.-K., Davies, C., Raubal, M., Schwering, A., Pederson, E.: The Influence of Scale, Context and Spatial Preposition in Linguistic Topology. In: Barkowsky, T., Knauff, M., Ligozat, G., Montello, D.R. (eds.) Spatial Cognition V. LNCS (LNAI), vol. 4387, pp. 439–452. Springer, Heidelberg (2007)
6. Mark, M.D.: Calibrating the Meaning of Spatial Predicates from Natural Language: Line-Region Relations. In: Proceedings of the Sixth International Symposiun on Spatial Data Handing, pp. 538–553 (1994)
7. Shariff, A.R., Egenhofer, M.J.: Natural-language Spatial Relations between Linear and Areal Objects: The Topology and Metric of English-language Terms. International Journal of Geographical Information Science 12(3), 215–246 (1998)
8. Xu, J.: Formalizing Natural-language Spatial Relations between Linear Objects with Topological and Metric Properties. International Journal of Geographical Information Science 21(4), 377–395 (2007)

9. Du, S.H., Wang, Q., Luo, G.: A Model for Describing and Composing Direction Relations between Overlapping and Contained Regions. Information Sciences: An International Journal Archive 178(14), 2928–2949 (2008)
10. Egenhofer, M.J., Rashid, A., Shariff, B.M.: Metric Details for Natural-Language Spatial Relations. ACM Transactions on Information Systems 16(4), 295–321 (1998)
11. Li, S.J., Wang, H.Q.: RCC8 Binary Constraint Network Can be Consistently Extended. Artificial Intelligence 170, 1–18 (2006)
12. Dylla, F., Moratz, R.: Empirical Complexity Issues of Practical Qualitative Spatial Reasoning about Relative Position. In: Proceedings of Workshop on Spatial and Temporal Reasoning at ECAI, pp. 169–175 (2004)
13. Shen, Q.J., Zhang, X.Y., Jiang, W.M.: Annotation of Spatial Relations in Natural Language. In: Proceedings of 2009 International Conference on Environmental Science and Information Application Technology, pp. 418–421 (2009)
14. Smith, E.P.: Statistical Comparison of Weighted Overlap Measures. Transactions of the American Fisheries Society 114(2), 250–257 (1985)