Complex Spatial Relationships

Robert Munro and Sanjay Chawla and Pei Sun
School of Information Technologies
University of Sydney
{rmunro, chawla, psun2712}@it.usyd.edu.au

Abstract

This paper describes the need for mining complex relationships in spatial data. Complex relationships are defined as those involving two or more of: multi-feature colocation, self-colocation, one-to-many relationships, self-exclusion and multi-feature exclusion. We demonstrate that even in the mining of simple relationships, knowledge of complex relationships is necessary to accurately calculate the significance of results. We implement a representation of spatial data such that it contains known 'weak-monotonic' properties, which are exploited for the efficient mining of complex relationships, and discuss the strengths and limitations of this representation.

1. Introduction

A relationship in spatial data is a relationship between features in a Euclidean space, defined in terms of the colocational trends of two or more features over that space. An example is determining the confidence of the *'where there's smoke there's fire'* with respect to a set of coordinates, each representing the feature smoke or fire. The task here would be to determine whether fire occurs in the neighborhood of smoke more than is randomly likely.

Neighborhoods are defined in terms of cliques (also known as neighbor-sets). A clique is defined as any set of items such that *all* items in that set colocate, for example, in Figure 1 and Table 1, the colocational pattern {A,D} occurs three times in four cliques, iv, v/vi & ix. Two features are typically said to colocate if they are positioned within a distance d of one another. As has been assumed in Figure 1, d is usually constant, but it may also be defined as varying locally within the space or with respect to a given feature.

Typically, the mining of information in a spatial domain involves representing the cliques as transactions, and undertaking association rule mining upon these transactions. While association rule mining is a well-developed field [5], the mining of confident cliques as transactions fails to capture many spatial phenomena of interest, due to most association mining techniques being optimized for 'market-basket' data. Spatial data is fundamentally different from market-basket data, both in its basic nature and distributional tendencies.

One factor unique to spatial data is that the number of transactions a single item may participate in is potentially unbounded, while in a market-basket this is limited to one (obviously, two people may purchase the same toothpaste product, but not the same exact tube). Self-colocation is also more likely in spatial data. The upper limit in a market-basket is multiple purchases of only one product, which is less likely than an equivalent spatial situation of an area of monoculture forest. Similarly, there may be direct relationships between features that don't colocate, as between animals displaying territorial behavior, making such relationships intrinsically more interesting in spatial data. A complex relationship is simply any combination of these different relationships. It is important to note that while the relationships are defined as complex, the phenomena they represent are often very simple [8].

Perhaps the most fundamental difference between spatial and other data in a transactional representation is the notion of significance. A colocation is considered significant if it occurs more than is randomly likely. In transactions representing market-baskets, the transactions, by definition, represent the complete space of the data (there are no empty baskets). In such cases, the significance of the data may be represented by frequency of the features in relation to the number of transactions, such as the interest measure proposed in [4]. In spatial data, however, the random likelihood of a colocation depends on the volume of the space from which it was taken. This is discussed in more detail in section 6.1.

1.1 Our contribution

We describe the need for mining complex relationships in spatial data. To the authors' best knowledge, this problem has not previously been addressed:

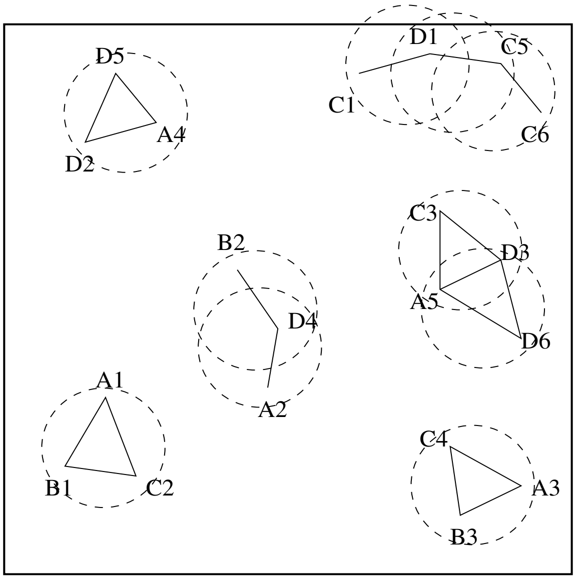

Figure 1. An example of spatial colocational patterns of the features A,B,C and D

1. We demonstrate that complex relationships are more numerous than simple ones and discuss why they are desirable to mine in the spatial domain.

2. We demonstrate that a representation and mining strategy for spatial data is possible such that it facilitates the efficient mining of complex relationships.

3. Most significantly, we demonstrate that it is necessary to mine complex relationships to accurately calculate the significance of results, even when the goal is only the mining of simple relationships.

1.2 Outline

In sections 2 and 3 we give the problem definition and a discussion of related work. In section 4 we describe and discuss the use of a participation threshold, rather than a support threshold, in the mining of spatial colocations. It should be noted that we are not redefining / refining any association rule algorithms; rather than throwing out the baby with the bathwater, we explore new applications and interpretations of existing ones. In section 5 we define and give examples of the various types of relationships in spatial data, including complex relationships. In section 6 we demonstrate that the knowledge of complex relationships in spatial data is necessary, even when the goal is the mining of only simple relationships. In section 7 we implement a transactional representation of spatial data such that it contains weak-monotonic properties, which are exploited by the maxPI algorithm [6] for the efficient mining of complex relationships, and discuss the strengths and limitations of this representation. In section 8 we conclude and discuss possible future directions.

No	Clique
i.	C_1, D_1
ii.	C_5, D_1
iii.	C_5, C_6
iv.	A_4, D_2, D_5
v.	A_5, C_3, D_3
vi.	A_5, D_3, D_6
vii.	A_1, B_1, C_2
viii.	B_2, D_4
ix.	A_2, D_4
x.	A_3, B_3, C_4

Table 1. Cliques in Figure 1

2 Problem Definition

Given: a set of items, $S = s_1, ...s_n$, each representing some entity with one or more features at a given coordinate and a rule confidence threshold, c

Find: all complex spatial relationships with confidence greater than c.

Constraints: The discovery of all rules of a given confidence is an intractable problem, so any method that can improve the efficiency of mining these rules is paramount. The data must be represented in a way that facilitates the mining of complex rules. (a transactional representation, as in Table 1, is the most commonly used in mining spatial data, as it allows the inclusion and the discovery of the interrelation of non-spatial features)

For mining colocations, this is a 3-step process: generate a set of the cliques in a representation that facilitates the mining of colocations; apply a mining algorithm to the cliques, returning a set of colocations and their confidences, the constituency of which is determined by given pruning and confidence thresholds; and calculate the significance of the mined colocations.

The first two steps are typically combined, so as to not to generate cliques already known to be below the given thresholds. In this paper, we assume that the first step has already taken place.

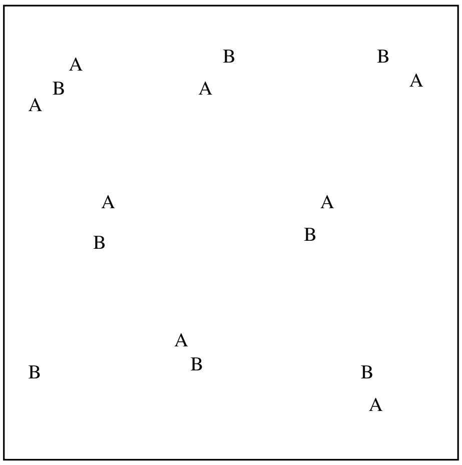

Figure 2. A positive relationship $A \to B$

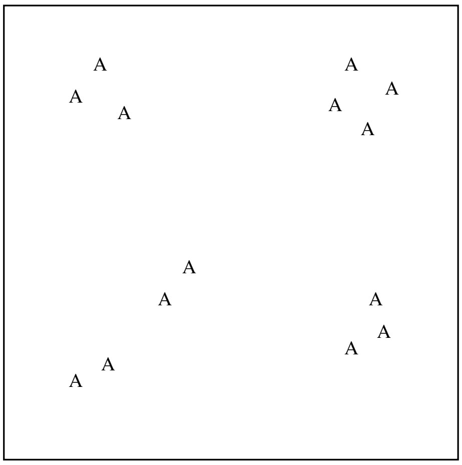

Figure 3. A self-colocating relationship $A \to A$

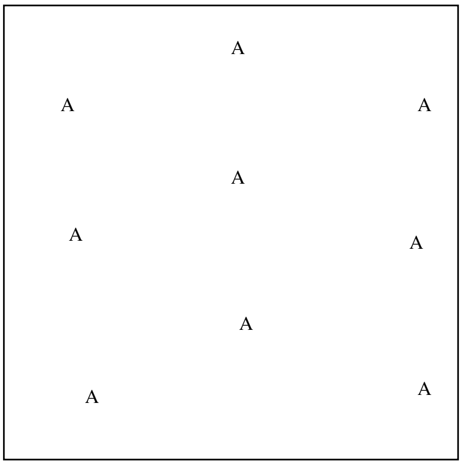

Figure 4. Self-exclusion, low confidence for $A \to A$

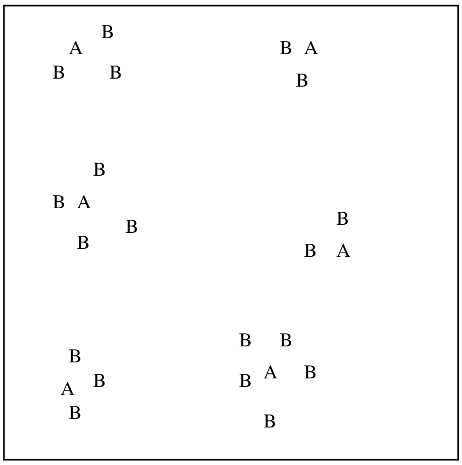

Figure 5. A one-to-many relationship $A \to B+$

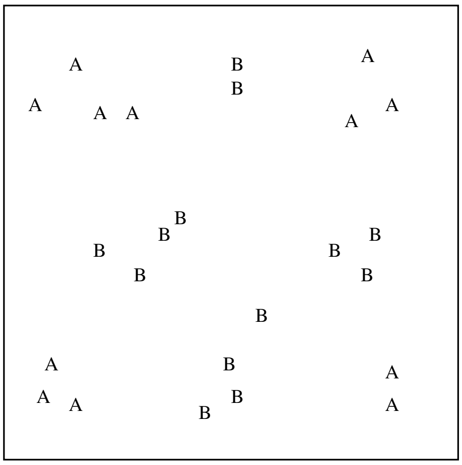

Figure 6. A multi-feature exclusive relationship $A \to -B$

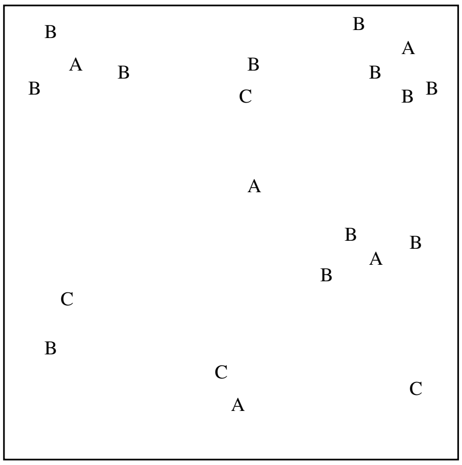

Figure 7. A complex relationship $A, -C \to B+$

3 Related Work

The first extension of the Apriori paradigm to spatial data was in [7]. However in their method they materialized all the possible spatial relationships that they intended to mine. This is equivalent to determining the universe of candidate interesting relationships. Thus, in some ways, their technique was 'hypothesis driven' rather than 'hypothesis generating.'

An efficient algorithm to mine a kind of spatial colocations was presented in [11]. The concepts of neighborhood, participation ratio and participation index were defined. Instead of support, the participation index was used as a pruning measure in the conventional Apriori-like technique.

The drawback of above method is that some confident colocation rules with low support are also pruned. In order to solve this problem, [6] proposed the concept of maximal participation index and it was used as pruning measure to replace the participation index . We will discuss these measures in detail in the next section, as they are central to our approach.

In [12], an algorithm was used to mine both positive and negative association rules. Negative rules are generated from infrequent item sets and interest is used as a further pruning measure. Their algorithm involves no spatial component.

4 Maximal Participation Ratio

In this section we will briefly describe the notion of Maximal Participation Index (maxPI) as described in [6] where more details can be found.

4.1 Participation ratio

Given a colocation pattern L and a feature $f \in L$, the participation ratio of f, $pr(L, f)$, can be defined as the support of L divided by the support of f. For example, in Figure 1, the support of $\{A, B, C\}$ is 2 and the support of C is 6, so $pr(\{A, B, C\}, C) = 2/6$.

4.2 Maximal participation index

Given a co-location pattern L, the maximal participation index of L, $maxPI(L)$ can be define as the maximal participation ratio of all the features in L, that is:

$$maxPI(L) = max_{f \in L}(pr(L, f)). \tag{1}$$

For example, in Figure 1, $maxPI(\{A, B, C\}) = max(pr(\{A, B, C\}, A), pr(\{A, B, C\}, B), pr(\{A, B, C\}, C)) = max(2/5, 2/3, 2/6) = 2/3$. A high maximal participation index indicates that at least one spatial feature strongly implies the pattern. By using maxPI, low frequency confident rules can be found which would be pruned by a support threshold [6].

4.3 The weak monotonic property of maxPI

Maximal participation index is not monotonic with respect to the pattern containment relations. For example, in the Figure 1, $(maxPI(A, C) = 3/5) < (maxPI(A, B, C) = 2/3)$. Interestingly, as pointed out in [6] the maximal participation index does have the following weak monotonic property: If P is a k-colocation pattern, then there exists at most one $(k - 1)$ subpatterns P' of P such that $maxPI(P') < maxPI(P)$. Relying on this weak monotonic property, the Apriori-like algorithm can be modified to mine confident patterns by using a maxPI threshold.

5 Relationship Definitions

5.1 Notation used

Feature: In this paper, a feature is represented as a capital letter, for example A.

Item: An instance of a feature (one item) is represented as the feature followed by an id number unique for that feature, for example A_2.

Absence: The absence of an item is represented by negation, for example $-A$. (Note: this is not the equivalent of the set-theory, $\neg A$, meaning the presence of any item other than A).

Self-colocation: Multiple instances of a feature (multiple items) are represented by a '+' following the feature, for example $A+$.

5.2 Types of Spatial Relationships

Here we define the different types of spatial relationships that are desirable to mine:

Positive (Simple) Relationships: This is the most common type of relationship mined, describing, for example, the fraction of A's that colocate with a B. eg: $A \rightarrow B$

Definition 1: A positive relationship (multi-feature colocation) in spatial data is a set of features that colocate at a ratio greater than some predefined threshold.

Self-colocation / Self-exclusion: This is the measure of which a feature tends to colocate with itself. Formally, it is the average cardinality of an item in a clique with respect to the expected cardinality of a random distribution. Extreme self-exclusion will be a perfectly uniform distribution with respect to the data space. eg: $A \rightarrow A+$

Definition 2: A feature is defined as self-colocating in spatial data if the items representing that feature colocate with each other at a ratio greater than some predefined threshold.

Definition 3: A feature is defined as self-excluding in spatial data if the items representing that feature colocate with each other at a ratio less than some predefined threshold.

One-to-Many relationships: This explicitly captures the cardinality of a relationship between two features. eg: $B+ \rightarrow A$.

Definition 4: Two features are defined as having a one-to-many relationship in spatial data if one feature occurs multiple times in the presence of the other feature, greater than some pre-defined threshold. Included within this definition are two-way one-to-many (many-to-many) relationships.

Multi-feature exclusive relationships: These are exclusive relationships with respect to two or more features. In terms of a transactional representation, they are negative rules, which are explored in [12]. eg: $A \rightarrow -B$

Definition 5: A multi-feature exclusive relationship in spatial data is defined as where a feature is absent from a given colocation at a ratio greater than a predefined threshold.

Complex relationships: These are any combination of two or more spatial relationship types.

Definition 6: A complex relation in spatial data is any relationship containing two or more of the properties defined in definitions 1-5. The independent application of the above rules may produce complex relationships such as $A+ \rightarrow B$ *and* $A \rightarrow -C$, but will not produce complex relationships such as $\{A+, -C\} \rightarrow B+$.

5.3 Sparse data and the mining of absence

A participation index directly addresses the problem that a certain item may have low support resulting in it being absent from very many cliques, and hence having a high negative support, in that it with therefore have a low participation ratio for each of those cliques. For example, if D is an infrequent feature, then the rule $A \rightarrow -D$ will most likely be confident. However, the participation ratio of $-D$ in $\{A, -D\}$ will be very low, as $-D$ will occur in many cliques. In other words the participation ratio of $-D$ in $\{A, -D\}$ will only be high if D is *atypically* absent from cliques containing A. Therefore, when using a participation ratio for sparse data, there is, in fact, a gain in efficiency. The results in section 7 confirm this.

6 Statistical Applications of Complex Relationships

Complex relationships are not restricted to mining complex rules. Complex relationships can be used to provide stronger definitions and more accurate significance testing for simple relationships.

6.1 Significance as a Complex relationship

In terms of confidence, the significance of a rule is given by the extent to which the observed confidence of a rule differs from the expected confidence given by a random distribution.

Lemma 1: The significance of a confidence rule of the form $A \rightarrow B$, is independent of the self-colocation/exclusion of A, but is dependent on the self-colocation/exclusion of B.

Proof Outline: In general, given that A occurs with frequency $F(A)$, and B with frequency $F(B)$, in a two dimensional space with dimensions x and y, with cliques formed by a distance d the random chance of $A \rightarrow B$ can be given by the product of the fraction of the total volume that each features occupies:

$$\frac{F(A)\pi d^2}{xy}\frac{F(B)\pi d^2}{xy} = \frac{F(A)F(B)\pi^2 d^4}{x^2y^2} \qquad (2)$$

The problem with the above, however, is that B may not be exclusively distributed with respect to itself.

Note that the random chance will not change with respect to the self-colocation/exclusion of A. The two extreme cases are: A self-excludes such that no A's are in the same clique (the equation given above); and A self-colocates such that all A's co-occur in one clique. In the second case all A's occupy an effective total space of πd^2, giving $F(A)$ an effective value of 1. However, if a B exists in that clique, then all A's in that clique colocate with it, so the equation must be multiplied by the number of A's in the clique - in this case $F(A)$. The equation is, therefore the equivalent of where A self-excludes.

The random chance will, however, change with respect to the self-colocation/exclusion of B. Assume the two extreme cases: B self-excludes such that no B's are in the same clique (again, this is the equation given above); and B self-colocates such that all B's co-occur in one clique. In the second case, all B's occupy an effective total space of πd^2, giving $F(B)$ an effective value of 1. If one A exists in that clique, the number of B's in that clique has no effect on the confidence as, by definition 1, only one unique A colocates with a B. The expected value of $A \rightarrow B$ is therefore given by:

$$\frac{F(A)1\pi^2 d^4}{x^2 y^2} \quad (3)$$

As the two extreme cases demonstrate, the expected value of $A \rightarrow B$ exists in a range with boundaries differing by a factor of $F(B)$. The consequence of this is that an accurate measure of the significance of a rule $A \rightarrow B$ must also include the measure of B's self-colocation/exclusion. The importance of this becomes obvious when, in spatial data, $F(B)$ may literally be the number of stars in the sky. Therefore, by definition 6, in order to measure the significance of a simple relationship $A \rightarrow B$, it is necessary to know a complex relationship.

An approximation of self-exclusion may be given by the ratio of the number of cliques containing B to the total number of B's. Assuming B occurs in $cf(B)$ cliques. This can underestimate the random chance, as it doesn't take into account the intersection of cliques in the data space where two or more B's are greater than distance d but less than distance $2d$ apart, or over-estimate, as it doesn't take into account the distance between items within a clique.

Alternatively, a calculation of deviance from expected behaviour can be found by observing the original coordinate distributions with a metric such as Ripley's K-function [2]. This will not necessarily give a more accurate measure, as it relates to global distributions not cliques, but it's relationship with colocation mining and clique representations is, in itself, an interesting area deserving further investigation.

Perhaps the most intuitive reason for above is because in the spatial domain, the rule $A \rightarrow B$ cannot be divorced from it's spatial properties. Even when A and B represent coordinates, *A must be thought of as coordinates and B must be thought of as (potentially overlapping) volumes.* This is a general truism for spatial data and will hold whether a constant or variable d is used, and when a simpler clique definition is given, such as the division of the feature space into 'grids', or a more complicated definition, such as the result of a clustering algorithm.

Lemma 2: The potential range of confidence rules of the form $A \rightarrow B$, will depend on the self-colocation/exclusion of both A and B.

Proof Outline: While the self-colocation/exclusion of A does not affect the significance of a confidence rule, it can limit the possible range of the observed confidence. Assuming that all A's and B's self-exclude, then, as in market-basket data, the maximum possible confidence for $A \rightarrow B$ is simply given by:

$$min\left[\frac{F(B)}{F(A)}, 1\right] \quad (4)$$

Where $F(A) > F(B)$, this will obviously be less than 1. However, if A self-colocates such that $cf(A) \leq F(B)$ than the maximum possible confidence will be 1.

Similarly, if B self-colocates, then the maximum possible confidence may be lower. The exact measure for the maximum possible observed confidence is:

$$min\left[\frac{cf(B)}{cf(A)}, 1\right] \quad (5)$$

A further factor that is not discussed here is where the potential size of some cliques extend beyond the boundaries of the measured space. Again, the random likelihood of this will relate to the ratio between d and the dimensions of the space. Here, it is simply assumed that it is very low.

6.2 Exclusion and maxPI

As a support threshold can prune confident rules with low frequency, a maxPI threshold can prune confident rules with low participation. While maxPI will return the complete set of items that satisfy both thresholds of maxPI and the minimum confidence, there may be the case such that a high confidence rule will not have a high corresponding maximal participation index.

An improved measure of participation includes the atypical exclusion of an item. We posit that by including the absence of items (negative items), we may discover a more robust measure for a participation index measure.

7 A Representation of Spatial Data for Mining Complex Relationships

In this section we propose and test one simple representation of spatial data that facilitates the efficient mining of complex relationships.

7.1 Mining complex relationships using the maximal participation index

In terms of the steps in the problem definition, the steps taken are: Generate all positive cliques in a transactional representation adding features representing the absence of

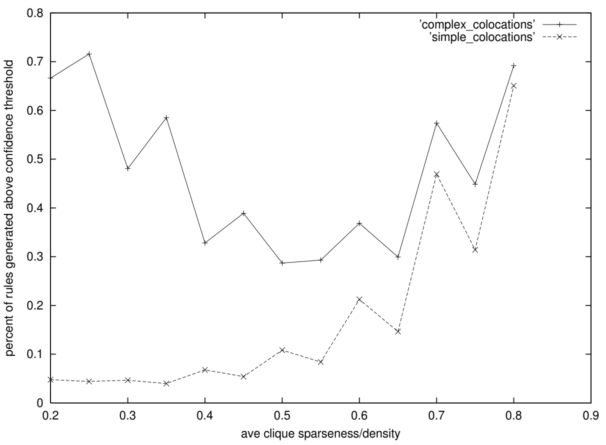

Figure 8. comparison of efficiency from sparse to dense data

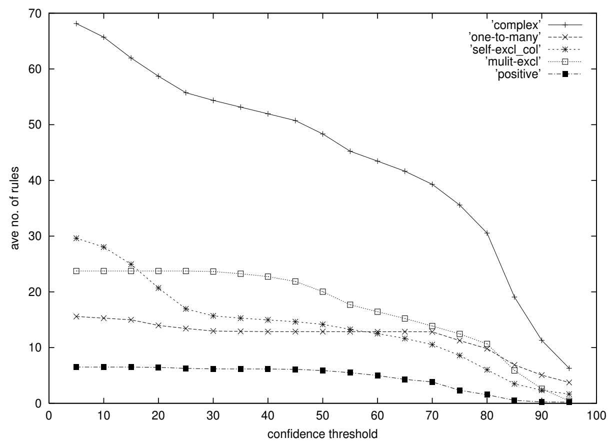

Figure 9. comparison of relationship type frequencies

features and the presence of multiple features; apply the maxPI algorithm to the transactions, as described in section 4, automatically pruning trivial/nonsensical collocations such as $A \rightarrow -A$. (For an analysis of the efficiency and application across different spatial data sets, see [6]) and return a set of colocations and their confidences; and calculate the significance of the confidences of the mined relationships, with respect to their significance, as described in section 6.1.

7.2 Test Sets

Synthetic data sets were created similar to those described in [1], but with the specific properties of spatial data, such the occurrence of a single item in many cliques and the occurrence of many items representing a single feature in one clique. Set constituency was varied according to sparseness, the number of features, and the number of items. The mining of relationships was varied according to the participation and confidence thresholds. A comprehensive set of tests corresponding was completed across approximately 100,000 different set/parameter combinations. A summary of results is given below.

Testing was undertaken to compare the efficiency of mining complex relationships to the mining of simple relationships with maxPI, and to investigate the relative frequencies of the different relationship types.

7.3 Results: Efficiency

As Figure 8 shows, the ratio of rules generated to confident rules found is typically more efficient for the mining of complex rules, especially when the data is sparse. Although it was never the case here, we do not rule out the possibility of the existence of a set such that the mining of simple relationships is more efficient than the mining of complex relationships. The results in Figure 8 are the average ratios for approximately 10,000 randomly generated data sets, which were varied according to sparseness/density: the average probability of a feature occurring in a given clique. The maxPI and confidence were held constant at 0.6 and 0.8, respectively. Varying the maximal participation index had little impact on the respective ratios. Varying the confidence threshold varied the scale of the ratio, but did not affect the scale of the two distributions with respect to each other.

A constant maintained across the generation of all sets were the inclusion of skews in the data such as: 'the probability of C appearing in a clique increases by 0.15 if A and B are present'. These were originally generated randomly, then maintained as averages about which all random sets were created. It is the interaction of such skews with the various thresholds that cause the unevenness in distributions in Figure 8.

7.4 Results: Frequency of relationship types

The results in Figure 9 are averages for approximately 1000 separate data sets, each with 10 features. The number of features is the most sensitive variable in the relative frequencies, due to the fact that there is the possibility of exponentially more exclusive and therefore complex sets with respect to the number of features in a clique, as discussed in section 5.3.

Typically, the number of complex relationships found was greater than but correlated with the number of other relationship types found. As Figure 9 shows, the number of complex relationships at a given confidence threshold was sensitive to the variance in the number of the other relationship types. Self-exclusion and self-colocation were modeled together in Figure 9 emphasize the complementary relationship between the two, as described in section 5.2. This is revealed in the corresponding steepness of gradient for self-exclusion/colocation at confidence < 0.3 and confidence > 0.7.

7.5 Limitations/Strengths of the representation

While there are representational issues with any type of data, appropiate representation is particularly important in the spatial domain [9].

Limitations. In one-to-many relationships, this model doesn't capture interesting ranges or distributions in the 'many', which is a task better suited for mixture modelling, or the techniques described in [3]. As pointed out in [10], the cost of fully transcribing spatial data into a transactional representation can, in some cases, be more expensive than the mining of the colocations, but as a full representation is necessary to accurately add the features representing absent and multiple items, a solution to this in the current representation may be problematic.

Strengths. The most obvious strength of this representation is that, currently, it is the only model that allows the mining of complex relationships in spatial data. A major strength of a transactional representation of spatial data not explored here is that it may be combined with non-spatial data, and so the addition of non-spatial data to the representation described here would be uncomplicated.

8 Conclusions / Future Work

We have defined the concept of complex relationships in spatial data.

We have described how, even in transactional representations, spatial data is fundamentally different from other forms of data, making the need to mine complex relationships of inherent interest.

We have demonstrated that even when simple relationships are the goal of mining spatial data, the mining of complex relationships is necessary for determining the significance of those relationships.

We have implemented and demonstrated a transactional representation of spatial data that allows the efficient mining of complex relationships, and discussed its limitations and strengths.

8.1 Future Work:

Apart from investigating improvements to the representation to address the limitations mentioned in 7.5, there are several future directions evident, such as the application to other types of data with a spatial component, such as spatio-temporal data and to a lesser extent natural language and biological systems.

One important step would be the combination of spatial coordinate features with spatial volume features (this is especially important in Geographic Information Systems, where a volume may represent the area of a lake, valley etc.). As we have demonstrated that with a purely coordinate system B in $A \rightarrow B$ must be treated as a volume, the inclusion of features that explicitly represent volumes should prove interesting.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[2] T. C. Bailey and A. T. Gatrell. *Interactive spatial data analysis*. Longman Scientific & Technical, 1995.

[3] S. Brin, R. Rastogi, and K. Shim. Mining optimized gain rules for numeric attributes. *IEEE transactions on knowledge and data engineering*, 15, 2003.

[4] G. Piatetsky-Shapiro. *Discovery, analysis and presentation of strong rules*. AAAI/MIT Press, 1991.

[5] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In W. Chen, J. Naughton, and P. A. Bernstein, editors, *2000 ACM SIGMOD Intl. Conference on Management of Data*, pages 1–12. ACM Press, 05 2000.

[6] Y. Huang, H. Xiong, and S. Shekhar. Mining confident colocation rules without a support threshold. In *Proc. 18th ACM Symposium on Applied Computing (ACM SAC)*, 2003.

[7] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In M. J. Egenhofer and J. R. Herring, editors, *Proc. 4th Int. Symp. Advances in Spatial Databases, SSD*, volume 951, pages 47–66. Springer-Verlag, 6–9 1995.

[8] R. Munro, S. Chawla, and P. Sun. Complex spatial relationships. *University of Sydney, School of Information Technologies Technical Report 539*, 2003.

[9] D. J. Peuquet. *Representations of space and time*. Guilford Press, 2002.

[10] S. Shekhar and S. Chawla. *Spatial Databases, A Tour*. 2002.

[11] S. Shekhar and Y. Huang. Discovering spatial co-location patterns: A summary of results. *Lecture Notes in Computer Science*, 2121, 2001.

[12] X. Wu, C. Zhang, and S. Zhang. Mining both positive and negative association rules. In *19th International Conference on Machine Learning (ICML-2002)*, 2002.