ISSN 1346-5597

National Institute of Informatics

NII Technical Report

A Generic Query-Based Model for Scalable Clustering

Michael E. Houle

NII-2006-008E
May 2006

URL: http://www.nii.ac.jp/ Tokyo 101-8430, JAPAN +81-3-4212-2000 FAX: +81-3-3556-1916

A Generic Query-Based Model for Scalable Clustering

Michael E. Houle
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan
meh@nii.ac.jp

Abstract

This paper presents a generic model for clustering that requires no direct knowledge of the nature or representation of the data. In lieu of such knowledge, the *relevant-set clustering* (RSC) model relies solely on the existence of an oracle that accepts a query in the form of a data item, and returns a ranked set of items relevant to the query. In principle, the role of the oracle could be played by any similarity search structure, or even a commercial search engine whose ranking function and relevancy scores are kept secret. The quality of cluster candidates, the degree of association between pairs of cluster candidates, and the degree of association between clusters and data items are all assessed according to the statistical significance of a form of correlation among pairs of relevant sets and/or candidate cluster sets. A scalable clustering heuristic based on the RSC model is also presented, and demonstrated for very large, high-dimensional datasets using a fast approximate similarity search structure as the oracle.

1 Introduction

The performance and applicability of virtually all of the well-known, traditional data clustering solutions depend heavily on specific properties or representations of the dataset. Some, such as k-means and its variants [14, 16], require the use of specific measures of data similarity; others, such as the agglomerative method DBSCAN [5] and many hierarchical hybrid methods such as STING [17], BIRCH [18] and CURE [8] pay a prohibitive computational cost when the representational dimension is high, due to a reliance on data representations and search structures that do not scale well to higher dimensions. Still others place assumptions on the distribution of the data that may or may not hold in practice.

Only relatively recently have methods been proposed that do not make heavy assumptions on the nature of the data. The generic Patch Model (PM) and associated PatClust clustering heuristic assumes only the existence of a pairwise similarity measure in order to produce data clusterings [11]. The key to the genericity and performance of PatClust is the SASH approximate search structure [12], which uses sampling techniques together with precomputed links to near neighbors to efficiently produce approximate neighborhoods for query items even when the underlying representational dimension of the data is very high. The SASH index relies on a pairwise distance measure, but otherwise makes no assumptions regarding the representation of the data. As PatClust relies only on the ranked neighbor lists for generating and evaluating cluster candidates, and avoids direct reliance on the data representation, the PatClust-SASH combination can scale to handle datasets of very large size and dimensionality.

This paper extends the generic model of [11] for clustering in the absence of explicit knowledge of the nature or representation of the data. In lieu of such knowledge, the model relies solely on the existence of an oracle that accepts a query in the form of a data item, and returns a ranked set of items relevant to the query. In principle, the role of the oracle could be played by any similarity search

structure capable of generating neighborhood sets with respect to some similarity measure, or even a commercial search engine whose ranking function and relevancy scores are kept entirely secret. Under this *relevant-set clustering* (RSC) model, the quality of cluster candidates, the degree of association between pairs of cluster candidates, and the degree of association between clusters and data items are all assessed according to the statistical significance of a form of correlation among pairs of relevant sets and/or candidate cluster sets.

In the next section, the RSC model itself is presented and contrasted with PM, and a heuristic clustering method based on the model is presented in Section 3. In Section 4, the heuristic is tested on a large protein sequence data set with large (but hidden) representational dimension, using the SASH search structure as the oracle. The paper concludes with a discussion of the potential applications of RSC to such problems as classification, integration of heterogeneous clustering results, cluster-based querying, and navigation of large datasets.

2 The Relevant-Set Correlation Clustering Model

2.1 Assumptions and notation

Let S be a dataset drawn from some domain D. For every item $q \in S$, we further assume the existence of a unique ordering $(q_1, q_2, \ldots, q_{|S|})$ of the items of S, where $i < j$ implies that q_i is deemed more relevant or similar to q than q_j. In practical settings, the item most relevant to q is generally q itself. Nevertheless, unless otherwise stated, we will not require that $q_1 = q$.

The relevancy ranking for q induces a collection of sets $\mathrm{Q}(q, k) = \{q_1, \ldots, q_k\}$ for each choice of set size $1 \leq k \leq |S|$. With respect to the ranking, if a dataset query-by-example operation were to be based at item q, $\mathrm{Q}(q, k)$ would represent the top-k relevant set. $\mathrm{Q}(q, k)$ can also represent the result of a k-nearest neighbor (k-NN) query for q with respect to some distance measure *dist* : $D \times D \rightarrow \mathbb{R}^{\geq 0}$. However, the RSC model makes no explicit use of the actual values of any distance function or other scoring function used to determine relevancy rankings. The generic applicability of RSC follows from its use of ranking information, and *only* ranking information, to decide questions of the following nature:

- Given two subsets A and B of S, how strong is the relationship between A and B?
- How strong is the mutual association among the items of A?
- Which of A and B constitutes the more significant aggregation of items?
- How strongly is item v related to the aggregation of items A?

Although most of the concepts for the RSC model are original to this paper, some have been borrowed from other models. To simplify the exposition of RSC, an accounting of the similarities and differences between RSC and other models will be deferred to Section 2.8.

2.2 Measuring inter-set association

For many if not most application areas, individual items can often be naturally assigned to more than one cluster. For example, a newspaper article concerning the ongoing stem cell research controversy in the United States could meaningfully contribute to clusters formed around the larger concepts of medical research, White House policy, the right-to-life movement, as well as the more narrow concept of stem cells per se. Rather than simply regarding such items as an impediment to clustering, they serve as the means by which the relationships between concepts (as represented by clusters of items) can be identified and assessed.

2.2.1 Confidence

Consider now two item sets A and B drawn from S, each associated with some underlying concept relevant to the domain. Even if no additional information is available regarding the nature of A and B, much in the same way as in association rule discovery [1], a *confidence* relationship from A to B can still be assessed according to the relative degree of overlap between the two sets:

$$\textsc{c}(A, B) \triangleq \frac{|A \cap B|}{|A|}.$$

If this confidence value is small, then there is little evidence of impact of the concept underlying B upon that underlying A. On the other hand, if the confidence is large, B can be considered to be strongly related to A.

The confidence measure can be interpreted in terms of the general notions of "precision" and "recall" from information retrieval. Let A be the target item set of a query that returns item set B as a result. Then the recall rate of the query is $\textsc{c}(A, B)$, and the precision of the query is $\textsc{c}(B, A)$.

Taken together, the two directed confidence values can be used to judge the qualitative relationship between two cluster candidates: if $\textsc{c}(A, B)$ and $\textsc{c}(B, A)$ are both high, then the concepts underlying A and B can be regarded as being similar. If only one of the two directed confidences is very high, one of the concepts can be considered to be a sub-concept of the other. However, under the RSC model it will often be more convenient to express the similarity between sets in terms of a single symmetric measure.

2.2.2 The cosine measure

A natural way of combining the two directed confidences between sets A and B into a single symmetric measure is by 'averaging' their orders of magnitudes. This yields the popular *cosine* similarity measure [10]:

$$\begin{aligned} \textsc{cm}(A, B) &\triangleq \sqrt{\textsc{c}(A, B) \cdot \textsc{c}(B, A)} \\ &= \frac{|A \cap B|}{\sqrt{|A|\,|B|}}. \end{aligned}$$

Mut Note that when A and B are equal in size, $\textsc{cm}(A, B) = \textsc{c}(A, B) = \textsc{c}(B, A)$. If A and B are identical, all three confidence values equal 1.

The cosine measure has a useful interpretation in terms of characteristic vectors of sets. Every item of S can be associated with a coordinate of a vector space whose dimension is equal to the size of S. A subset A of S can be represented by a zero-one characteristic vector in this space, where a coordinate value of 1 indicates that the corresponding item is a member of A, and a value of 0 indicates that the item does not belong to A. If $\vec{A}$ and $\vec{B}$ are the respective characteristic vectors of A and B, then

$$\begin{aligned} \textsc{cm}(A, B) &= \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\|\,\|\vec{B}\|} \\ &= \cos\theta_{\vec{A},\vec{B}}, \end{aligned}$$

where $\theta_{\vec{A},\vec{B}}$ is the angle formed by the two vectors. The angle size $\cos^{-1}\textsc{cm}(A, B)$ is known to satisfy all the properties of a distance metric, including the triangle inequality.

2.2.3 Set membership correlation

The cosine measure can also be expressed in terms of the Pearson correlation between corresponding entries of set characteristic vectors. For sequences of variables $(x_1, x_2, \ldots, x_n)$ and $(y_1, y_2, \ldots, y_n)$

with means $\bar{x}$ and $\bar{y}$, respectively, the Pearson correlation is given by the following formula:

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}. \end{aligned}$$

Applying the formula to the characteristic vectors of sets A and B, and noting that $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i = n\bar{x}$ whenever $x_i \in \{0,1\}$, we obtain the following *set correlation* formula:

$$\begin{aligned} \textsc{r}(A,B) &= \frac{|A\cap B| - \frac{|A|\,|B|}{|S|}}{\sqrt{|A|\,|B|(1-\frac{|A|}{|S|})(1-\frac{|B|}{|S|})}} \\ &= \frac{|S|}{\sqrt{(|S|-|A|)(|S|-|B|)}}\left(\textsc{cm}(A,B) - \frac{\sqrt{|A|\,|B|}}{|S|}\right). \end{aligned} \quad (1)$$

In situations where the set sizes of A and B are both small compared to the size of the domain (as is often the case in clustering applications), their cosine measure and set correlation values are nearly equal. The cosine measure can thus be legitimately viewed as a close variant of the Pearson correlation for set membership.

2.3 Measuring intra-set association

When only relevancy ranking information is available for the items of a dataset, we can no longer make use of existing measures of intra-cluster association based on density or other distance-based estimates of the data distribution. Instead, RSC assesses the internal association of set A in terms of correlations involving relevant sets based at the members of A.

Intuitively speaking, if an item $v \in A$ is strongly associated with the remaining items of A, it is likely that the items of S that are highly relevant to v also belong to set A. Alternatively, if A as a whole were to have a high degree of internal cohesion, one would expect many if not most of its items to have relevant sets that overlap significantly with one another. These intuitions form the motivation for two intra-set association measures under the RSC model.

2.3.1 First-order self-confidence

The *first-order self-confidence* measure assesses intra-set association as the expectation of the cosine measure value between $|A|$ and the relevant set of size A based at a randomly-selected item of A:

$$\begin{aligned} \textsc{sc}_1(A) &\triangleq \frac{1}{|A|}\sum_{v\in A} \textsc{cm}(A, \textsc{q}(v,|A|)) \\ &= \frac{1}{|A|^2}\sum_{v\in A} |A\cap \textsc{q}(v,|A|)|. \end{aligned} \quad (2)$$

A self-confidence value of 1 indicates perfect association among the members of A, whereas a value approaching 0 indicates little or no internal association within A. Using Equation (1), the self-confidence can be expressed in terms of the Pearson correlation as:

$$\textsc{sc}_1(A) = \frac{|S|-|A|}{|S|}\textsc{sr}_1(A) + \frac{|A|}{|S|}, \quad (3)$$

where

$$\mathrm{SR}_1(A) \triangleq \frac{1}{|A|}\sum_{v\in A}\mathrm{R}(A,\mathrm{Q}(v,|A|)),$$

the *first-order self-correlation*, is the expected correlation between A and the relevant set of size $|A|$ based at a randomly-selected item of A.

2.3.2 Second-order self-confidence

The *second-order self-confidence* measure assesses intra-set association as the expectation of the cosine measure value between the relevant sets of two items randomly selected from A, with replacement, where the relevant sets are of the same size as A. Although a formulation involving only unordered pairs of distinct items is possible, the following definition will be seen to have useful properties in the context of ranking of cluster items:

$$\begin{aligned}\mathrm{SC}_2(A) &\triangleq \frac{1}{|A|^2}\sum_{v\in A}\sum_{w\in A}\mathrm{CM}(\mathrm{Q}(v,|A|),\mathrm{Q}(w,|A|)) \\ &= \frac{1}{|A|^3}\sum_{v\in A}\sum_{w\in A}|\mathrm{Q}(v,|A|)\cap\mathrm{Q}(w,|A|)|.\end{aligned} \tag{4}$$

Again, a value of 1 indicates perfect association among the members of A, whereas a value approaching 0 indicates little or no internal association within A. The second-order self-confidence can be expressed in terms of the Pearson correlation as:

$$\mathrm{SC}_2(A) = \frac{|S|-|A|}{|S|}\mathrm{SR}_2(A)+\frac{|A|}{|S|}, \tag{5}$$

where

$$\mathrm{SR}_2(A) \triangleq \frac{1}{|A|^2}\sum_{v\in A}\sum_{w\in A}\mathrm{R}(\mathrm{Q}(v,|A|),\mathrm{Q}(w,|A|)),$$

the *second-order self-correlation*, is the expected correlation between the relevant sets of two items randomly selected from A, with replacement, where the relevant sets are of the same size as A.

2.4 Significance testing

In general, when making inferences involving Pearson correlation, a high correlation value alone is not considered sufficient to judge the significance of the relationship between two variables. When the number of variable pairs is small, it is much easier to achieve a high value by chance than when the number of pairs is large. For this reason, to help interpret correlation scores, statisticians resort to tests of significance such as the t-test that account for variation in the number of pairs. The correlation score is considered significant if it deviates sufficiently from zero (randomness), as measured by the t statistic.

As can be seen from their formulations in terms of correlations, a significance test would also be useful for both forms of self-confidence. The need for a significance test is illustrated by the two-dimensional example in Figure 1, where the relevancy ranking is induced by the Euclidean distance measure. Set A consists of 20 points with first-order self-confidence and self-correlation scores of 0.8525 and 0.815625, respectively, whereas set B consists of only 5 points but has higher self-confidence and self-correlation scores, both equal to 1. Set C consists of 10 points with self-confidence 0.45 and self-correlation $\frac{7}{18}\approx 0.3889$. Of the three sets, A appears to constitute the most significant aggregation of points, while C appears to be the least significant aggregation.

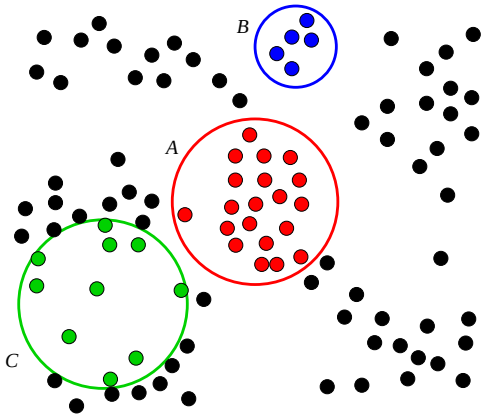

Figure 1: Set A has smaller first-order self-confidence and self-correlation than B, but is a more significant aggregation.

2.4.1 The randomness hypothesis

One might hope to make use of existing tests for the significance of Pearson correlation for judging the significance of self-confidence values. However, in the correlation formulations of (3) and (5), the number of variable pairs corresponds to the size of the domain, $|S|$, and is usually fixed. For self-confidence values, any significance test must take into account differences in the size of the candidate item sets themselves.

During the clustering process, instead of verifying whether or not the self-confidence of a candidate set meets a minimum significance threshold, we will more often need to test whether one candidate has a more significant self-confidence value than another. To do this, we test against the assumption that each relevant set contributing to the self-confidence score is independently generated by means of uniform random selection from among the available items. In practice, of course, the relevant sets are far from being random. However, this admittedly unlikely situation serves as a convenient reference point from which the significance of actual self-confidence values can be assessed.

Under the randomness hypothesis, the mean and standard deviation of the self-confidence score can be calculated (as will be shown below). Standard scores (also known as Z-scores) of the two actual candidates can then be generated and compared. The more significate candidate would be the one whose standard score is highest — that is, the one whose self-confidence score exceeds the expected value by the greatest number of standard deviations.

2.4.2 First-order significance of sets

Assume that we are given a fixed set $U \subseteq S$ and a second set V chosen uniformly at random (without replacement) from the items of S. Then $X = |U \cap V|$ is a hypergeometrically-distributed random variable with expectation

$$\mathbf{E}[X] = \frac{|U|\,|V|}{|S|}$$

and variance

$$\mathbf{Var}[X] = \frac{|U|\,|V|\,(|S| - |U|)(|S| - |V|)}{|S|^2(|S| - 1)}.$$

Consider now the first-order self-confidence value $\mathrm{SC}_1(A)$ of some non-empty subset $A \subseteq S$, as expressed by the formula (2). Let $\underline{\mathrm{SC}}_1(A)$ denote the first-order self-confidence value for A under the

assumption that each relevant set $\mathrm{Q}(v, |A|)$ is independently replaced by a set $\underline{\mathrm{Q}}(v, |A|)$ consisting of $|A|$ items selected uniformly at random from S. Then $\underline{\mathrm{SC}}_1(A)$ is a random variable with expectation

$$\begin{aligned}\mathbf{E}[\underline{\mathrm{SC}}_1(A)] &= \frac{1}{|A|^2}\sum_{v\in A}\mathbf{E}[|A\cap\underline{\mathrm{Q}}(v,|A|)|] \\ &= \frac{1}{|A|^2}\sum_{v\in A}\frac{|A|^2}{|S|} = \frac{|A|}{|S|}\end{aligned}$$

and variance

$$\begin{aligned}\mathbf{Var}[\underline{\mathrm{SC}}_1(A)] &= \frac{1}{|A|^4}\sum_{v\in A}\mathbf{Var}[|A\cap\underline{\mathrm{Q}}(v,|A|)|] \\ &= \frac{1}{|A|^4}\sum_{v\in A}\frac{|A|^2(|S|-|A|)^2}{|S|^2(|S|-1)} \\ &= \frac{(|S|-|A|)^2}{|A|\,|S|^2(|S|-1)}.\end{aligned}$$

With respect to the randomness hypothesis, the significance value for $\mathrm{SC}_1(A)$ is thus the standard score

$$\begin{aligned}Z_1(A) &\triangleq \frac{\mathrm{SC}_1(A)-\mathbf{E}[\underline{\mathrm{SC}}_1(A)]}{\sqrt{\mathbf{Var}[\underline{\mathrm{SC}}_1(A)]}} \\ &= \frac{|S|}{|S|-|A|}\sqrt{|A|\,(|S|-1)}\left(\mathrm{SC}_1(A)-\frac{|A|}{|S|}\right).\end{aligned}$$

Expressed in terms of the Pearson correlation using (3), the significance value simplifies to

$$Z_1(A) = \sqrt{|A|\,(|S|-1)}\,\mathrm{SR}_1(A).$$

Under the same assumptions, it is not difficult to show that the standard score for $\mathrm{SR}_1(A)$ also equals $Z_1(A)$, and thus there is no difference between the confidence formulation and the correlation formulation when it comes to testing the significance of an aggregation of items. Accordingly, we shall simply refer to the value $Z_1(A)$ as the *first-order significance* of A.

Returning to the example in Figure 1, the first-order significances of the three sets are $Z_1(A) = \frac{783}{160}\sqrt{55} \approx 36.29$, $Z_1(B) = 3\sqrt{55} \approx 22.25$, and $Z_1(C) = \frac{7}{6}\sqrt{110} \approx 12.24$. These values conform with our earlier intuition regarding the relative significance of A, B and C.

The randomness hypothesis, as stated above, does not take into account the possibility that the relevant set $\mathrm{Q}(v, |A|)$ may be guaranteed to contain v. If such a guarantee were provided, the randomness hypothesis could be varied so that $\underline{\mathrm{Q}}(v, |A|)$ comprised v together with $|A|-1$ items selected uniformly at random from among the items of $S \setminus \{v\}$. Moreover, if the set A were itself known to be a relevant set of some item $a \in S$, then one may opt to select random relevant sets only for the $|A|-1$ summation terms where $v \neq a$. These choices lead to slightly different (and less elegant) formulations of the significance measure, the details of which are omitted here.

2.4.3 Second-order significance of sets

Consider next the second-order self-confidence value $\mathrm{SC}_2(A)$ of some non-empty subset $A \subseteq S$, as expressed by the formula (4). Let $\underline{\mathrm{SC}}_2(A)$ denote the second-order self-confidence value for A under

the randomness hypothesis. Then $\underline{\mathrm{SC}_2}(A)$ is a random variable with expectation

$$\begin{aligned}\mathbf{E}[\underline{\mathrm{SC}_2}(A)] &= \frac{1}{|A|^3}\sum_{v\in A}\sum_{w\in A}\mathbf{E}[|\underline{\mathrm{Q}}(v,|A|)\cap\underline{\mathrm{Q}}(w,|A|)|]\\ &= \frac{1}{|A|^3}\sum_{v\in A}\sum_{w\in A}\frac{|A|^2}{|S|}=\frac{|A|}{|S|}\end{aligned}$$

and variance

$$\begin{aligned}\mathbf{Var}[\underline{\mathrm{SC}_2}(A)] &= \frac{1}{|A|^6}\sum_{v\in A}\sum_{w\in A}\mathbf{Var}[|\underline{\mathrm{Q}}(v,|A|)\cap\underline{\mathrm{Q}}(w,|A|)|]\\ &= \frac{1}{|A|^6}\sum_{v\in A}\sum_{w\in A}\frac{|A|^2(|S|-|A|)^2}{|S|^2(|S|-1)}\\ &= \frac{(|S|-|A|)^2}{|A|^2|S|^2(|S|-1)}.\end{aligned}$$

The significance value for $\mathrm{SC}_2(A)$ is the standard score

$$\begin{aligned}Z_2(A) &\triangleq \frac{\mathrm{SC}_2(A)-\mathbf{E}[\underline{\mathrm{SC}_2}(A)]}{\sqrt{\mathbf{Var}[\underline{\mathrm{SC}_2}(A)]}}\\ &= \frac{|S|}{|S|-|A|}|A|\sqrt{|S|-1}\left(\mathrm{SC}_2(A)-\frac{|A|}{|S|}\right).\end{aligned}$$

Expressed in terms of the Pearson correlation using (5), the significance value simplifies to

$$Z_2(A) = |A|\sqrt{|S|-1}\,\mathrm{SR}_2(A).$$

The standard score for $\mathrm{SR}_2(A)$ also equals $Z_2(A)$, and thus we can refer to the value $Z_2(A)$ as the *second-order significance* of set A.

As was the case with first-order significance, the randomness hypothesis does not take into account the possibility that the relevant set $\mathrm{Q}(v,|A|)$ may be guaranteed to contain v, or that $\mathrm{R}(\mathrm{Q}(v,|A|),\mathrm{Q}(w,|A|))$ always equals 1 whenever $v=w$. If such a guarantee were provided, the randomness hypothesis could be varied so that the contributions are dropped for the case $v=w$, and so that $\underline{\mathrm{Q}}(v,|A|)$ comprised v together with $|A|-1$ items selected uniformly at random from among the items of $S\setminus\{v\}$. This choice leads to a slightly different formulation of second-order significance, the details of which are omitted.

2.5 Inter-set significance

The significance of the inter-set measures $\mathrm{CM}(A,B)$ and $\mathrm{R}(A,B)$ can also be analyzed with respect to an assumption of randomness. Let $\underline{\mathrm{CM}}(A,B)$ denote the cosine value between A and B under the assumption that $|B|$ is independently replaced by a set consisting of $|B|$ items selected uniformly at random from S. Then $\underline{\mathrm{CM}}(A,B)$ is a random variable with expectation

$$\mathbf{E}[\underline{\mathrm{CM}}(A,B)]=\frac{\sqrt{|A|\,|B|}}{|S|}$$

and variance

$$\mathbf{Var}[\underline{\mathrm{CM}}(A,B)]=\frac{(|S|-|A|)(|S|-|B|)}{|S|^2(|S|-1)}.$$

The expectation and variance do not change if $|A|$ is selected at random instead of B, or even if both sets are selected at random.

The significance value for $\textsc{cm}(A,B)$ is the standard score

$$\begin{aligned} Z(A,B) &\triangleq \frac{\textsc{cm}(A,B) - \mathbf{E}[\underline{\textsc{cm}}(A,B)]}{\sqrt{\mathbf{Var}[\underline{\textsc{cm}}(A,B)]}} \\ &= |S|\sqrt{\frac{|S|-1}{(|S|-|A|)(|S|-|B|)}}\left(\textsc{cm}(A,B) - \frac{\sqrt{|A|\,|B|}}{|S|}\right) \\ &= \sqrt{|S|-1}\,\textsc{r}(A,B). \end{aligned}$$

Under the same assumptions, it is not difficult to show that the standard score for $\textsc{r}(A,B)$ also equals $Z(A,B)$, and thus there is no difference between the confidence formulation and the correlation formulation when it comes to testing the significance of the association between two sets. Interestingly, since the coefficient $\sqrt{|S|-1}$ does not depend on A or B, the analysis shows that the set correlation measure $\textsc{r}(A,B)$ fully captures the significance of the relationship between two subsets of $|S|$.

2.6 Partial significance

Within any highly-significant set A, the contributions of some relevant sets to the self-confidence (or self-correlation) scores may be substantially greater than others. Those items whose relevant sets contribute highly can be viewed as better associated with the concept underlying aggregation A than those whose contributions are small. However, to compare the contributions of a single item with respect to several different sets, or the contributions of several different item-set pairs, a test of significance is needed.

2.6.1 First-order partial significance

The contribution to $\textsc{sc}_1(A)$ attributable to item $v \in A$ is given by

$$t_1(A,v) \triangleq \frac{1}{|A|}\textsc{cm}(A, \textsc{q}(v,|A|)).$$

The *first-order significance* of the relationship between v and A is defined as the standard score for $t_1(A,v)$ under the randomness hypothesis:

$$Z_1(A,v) = \sqrt{|S|-1}\,\textsc{r}(A, \textsc{q}(v,|A|)).$$

The details of the derivation are omitted, as the analysis is essentially the same as that of $\textsc{cm}(A,B)$, with $B = \textsc{q}(v,|A|)$. Note that the same standard score would be obtained when considering the contributions of v to $\textsc{sr}_1(A)$ instead of $\textsc{sc}_1(A)$.

The first-order significance of A can be concisely expressed in terms of the sum of its partial significances:

$$Z_1(A) = \frac{1}{\sqrt{|A|}}\sum_{v\in A} Z_1(A,v). \tag{6}$$

2.6.2 Second-order partial significance

The contribution to $\mathrm{SC}_2(A)$ made by item $v \in A$ is given by

$$\begin{aligned} t_2(A,v) &\triangleq \frac{1}{|A|^2}\sum_{w\in A}\mathrm{CM}(\mathrm{Q}(v,|A|),\mathrm{Q}(w,|A|)) \\ &= \frac{1}{|A|^3}\sum_{w\in A}|\mathrm{Q}(v,|A|)\cap\mathrm{Q}(w,|A|)|. \end{aligned}$$

Let $\underline{t}_2(A,v)$ denote the associated random variable under the randomness hypothesis. Its expectation is

$$\begin{aligned} \mathbf{E}[\underline{t}_2(A,v)] &= \frac{1}{|A|^3}\sum_{w\in A}\mathbf{E}[|\mathrm{Q}(v,|A|)\cap\mathrm{Q}(w,|A|)|] \\ &= \frac{1}{|A|^3}\sum_{w\in A}\frac{|A|^2}{|S|} = \frac{1}{|S|} \end{aligned}$$

and its variance is

$$\begin{aligned} \mathbf{Var}[\underline{t}_2(A,v)] &= \frac{1}{|A|^6}\sum_{w\in A}\mathbf{Var}[|\mathrm{Q}(v,|A|)\cap\mathrm{Q}(w,|A|)|] \\ &= \frac{1}{|A|^6}\sum_{w\in A}\frac{|A|^2(|S|-|A|)^2}{|S|^2(|S|-1)} \\ &= \frac{(|S|-|A|)^2}{|A|^3|S|^2(|S|-1)}. \end{aligned}$$

The *second-order significance* of the relationship between v and A is the standard score for $t_2(A,v)$,

$$\begin{aligned} Z_2(A,v) &\triangleq \frac{t_2(A,v)-\mathbf{E}[\underline{t}_2(A,v)]}{\sqrt{\mathbf{Var}[\underline{t}_2(A,v)]}} \\ &= \frac{|S|}{|S|-|A|}\sqrt{|A|\,(|S|-1)}\left(\mathrm{CM}_2(A,v)-\frac{|A|}{|S|}\right) \\ &= \sqrt{|A|\,(|S|-1)}\,\mathrm{R}_2(A,v), \end{aligned}$$

where

$$\mathrm{CM}_2(A,v) \triangleq \frac{1}{|A|}\sum_{w\in A}\mathrm{CM}(\mathrm{Q}(v,|A|),\mathrm{Q}(w,|A|))$$

is the average of the cosine measure values between the relevant set $\mathrm{Q}(v,|A|)$ and all other relevant sets of the same size based at items of A, and

$$\mathrm{R}_2(A,v) \triangleq \frac{1}{|A|}\sum_{w\in A}\mathrm{R}(\mathrm{Q}(v,|A|),\mathrm{Q}(w,|A|))$$

is the average of the correlation between these same pairs of relevant sets. Once again, the same standard score $Z_2(A,v)$ would be obtained when considering the contributions of v to $\mathrm{SR}_2(A)$ instead of $\mathrm{SC}_2(A)$.

In terms of its partial significances, the second-order significance of A reduces to

$$Z_2(A) = \frac{1}{\sqrt{|A|}}\sum_{v\in A}Z_2(A,v). \tag{7}$$

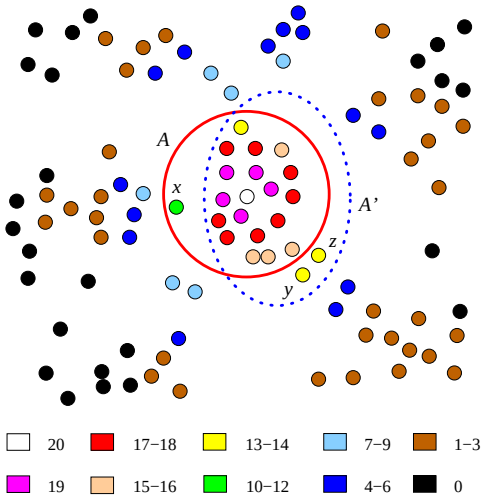

Figure 2: Rankings of points according to first-order partial significance with respect to A. The value ranges shown are of $|A \cap \mathrm{Q}(v, |A|)|$, which determines the same ranking as $Z_1(A, v)$.

2.7 Cluster queries and cluster reshaping

Partial significances, whether first-order or second-order, can be directly used to rank the items of A according to their level of association with A, much like the items of a relevant set are ranked with respect to an individual query item. Moreover, the ranking can be extended to all items of S, as the definitions of partial significance are meaningful regardless of whether v is actually a member of A. In this case, A can be regarded as a form of *cluster query* that returns a set of items ranked according to $Z_1(A, v)$ or $Z_2(A, v)$. Although in principle A could be any set of items, equation (6) indicates that the relevancy scores are high only when A is itself a significant aggregation of items — that is, when A is itself a 'reasonably good' cluster candidate. From the definition of first-order partial significance, ranking according to $Z_1(A, v)$ is easily seen to be equivalent to ranking according to $\mathrm{CM}(A, \mathrm{Q}(v, |A|))$ or $\mathrm{R}(A, \mathrm{Q}(v, |A|))$.

Figure 2 illustrates the first-order cluster query ranking for the point set A from Figure 1. In this example, the partial significance ranking manages a rough approximation of the original Euclidean distance ranking as measured from a central location within the cluster, despite the lack of knowledge of the individual Euclidean distance values themselves.

It is worth noting that two items lying outside A (y and z) have higher partial significances than one item contained in A (item x). This suggests that partial significances may be used to 'reshape' a candidate cluster set, by replacing poorly-associated members with other, more strongly-associated items, thereby improving the overall cluster quality. Let us consider the situation where A has been reshaped to yield a new candidate set A'. To assess the quality of A', the average association can be computed between set A and relevant sets based at items of A', instead of at items of A as in the original first-order definitions of self-confidence and self-correlation, as follows:

$$\mathrm{SC}_1(A, A') \triangleq \frac{1}{|A'|} \sum_{v \in A'} \mathrm{CM}(A, \mathrm{Q}(v, |A|));$$

$$\mathrm{SR}_1(A, A') \triangleq \frac{1}{|A'|} \sum_{v \in A'} \mathrm{R}(A, \mathrm{Q}(v, |A|)).$$

Starting from either of these two association measures, based on the random relevant set hypothesis, one can derive the following significance score for the reshaped set A'. The details of the derivation are omitted, as they are very similar to those of equation (6).

$$Z_1(A, A') \quad = \quad \frac{1}{\sqrt{|A'|}} \sum_{v \in A'} Z_1(A, v). \tag{8}$$

An important implication of equation (8) is that for any fixed candidate size $|A'| = k$, the highest possible significance is attained by letting A' consist of those k items of S having the highest first-order partial significance values with respect to A.

Returning to the example of Figure 2, the reshaped candidate set $A' = (A \cup \{y, z\}) \setminus \{x\}$ has significance value $Z_1(A, A') = \frac{137}{56}\sqrt{33} \approx 37.18$, which is an improvement over the original significance score $Z_1(A, A) = Z_1(A) \approx 36.29$. It can be verified that A' attains the maximum significance score over all possible reshapings of A.

It should be noted that there seems to be no simple, meaningful way of modifying the definitions of second-order self-confidence and self-correlation to accommodate cluster reshaping — the individual confidences and correlations computed do not explicitly rely on the membership of set A, making it difficult to establish a relationship between A' and A.

2.8 Relationship to previous models

The directed confidence and first-order self-confidence measures of the RSC model are derived from the confidence and self-confidence measures introduced under the Patch Model (PM) of [11]. PM self-confidence is defined in terms of what appears as directed confidence under RSC, with all cluster candidates and relevant sets restricted to being neighborhoods defined according to a pairwise distance metric over the item set. PM also employs directed confidence as an item-to-cluster relevancy measure, a choice justified by the analysis of RSC first-order cluster query ranking. The main improvements of RSC over PM consist of the statistical framework presented in this section, including the concepts of set correlation, self-correlation, and significance testing, as well as the notions of second-order self-confidence, and cluster reshaping based on partial significance.

The origins of the directed confidence measure itself can be traced to the shared-neighbor merge criterion of Jarvis and Patrick [13] used in agglomerative clustering. The criterion states that two clusters can be merged if they contain equal-sized subclusters A and B such that $\mathrm{c}(A, B) \geq mk$, where k is the size of A and B, and $0 < m \leq \frac{1}{k}$ is a fixed merge threshold parameter. As is often the case with agglomerative clustering methods, this merge criterion can result in clusters composed of chains of subclusters having little or no association with one another. Other shared-neighbor merge criteria have been proposed in an attempt to compensate for this chaining effect [4, 9]. Although the RSC model ultimately depends on shared neighbor information for its cluster quality assessment, the clustering heuristics presented in the next section are not agglomerative, and do not suffer from chaining.

3 Clustering Under the RSC Model

Traditionally, a clustering is considered to be of high quality when pairs of items belonging to a common cluster are mutually well-associated, while pairs of items belonging to different clusters are well-differentiated. In the previous section, we saw how the RSC model can be used to assess the similarity and relative degree of internal association of any two cluster candidate item sets in isolation. We now turn our attention to the problem of generating a full data clustering based on this model.

3.1 Scalability issues

As does the PM-based PatClust heuristic method of [11], the RSCbased clustering method presented in this section seeks to generate as many clusters as possible, subject to the following restrictions:

- All cluster candidate item sets should meet minimum threshold values of cluster quality.
- All pairs of cluster candidates should meet maximum threshold values on cluster similarity.

Under RSC, cluster quality can be measured as a function of significance, self-correlation, and/or self-confidence (first- or second-order), whereas only the first-order self-confidence measure is available under PM. Cluster similarity can be measured in terms of significance, correlation and or confidence under RSC, but only confidence under PM. Nevertheless, regardless of the measures used, under both models each cluster can be thought to compete for 'territory' within the space covered by the dataset. If a region of the data is sufficiently well-associated for a subset to meet or exceed the minimum threshold on cluster quality, then a cluster should be chosen to represent the region. However, if two or more highly-similar cluster candidates arise from within the region, then only one of the candidates should be retained.

The selection of cluster candidates can be viewed within the framework of the well-studied family of independent vertex set problems for graphs. Those cluster candidate item sets whose quality scores meet the minimum threshold are mapped onto vertices of a graph, with assigned weights equal to the quality scores. A vertex pair is joined by an edge wherever the inter-cluster similarity scores of the corresponding cluster candidates exceeds the maximum threshold. Clustering thus reduces to the problem of selecting a subset of graph vertices that maximizes some objective function involving such variables as subset size and vertex weights, subject to the restriction that no graph edge may have both of its endpoints selected. Even for the simple case where only the size of the subset is to be maximized (ignoring the vertex weights), we are left with the classical maximum independent vertex set problem, which is is known to be NP-hard [6].

The inherent hardness of the cluster candidate selection problem is not the only impediment to the scalability of RSC-based (and PM-based) clustering. In practice, it is not possible to consider all possible data subsets as cluster candidates; some form of restriction on the eligibility of cluster candidate item sets is needed. Also, calculating the significance of large candidate cluster sets is too expensive when the number of such sets is high, since the number of relevant sets involved is linear in the size of the candidate set, and the total size of the relevant sets is quadratic. For these reasons, practical applications of the RSC model thus unavoidably require the development of heuristics (as opposed to 'exact' techniques) whose design choices are driven by the needs of efficiency and scalability.

3.2 Scaling via sampling

The heuristic described in the remainder of this section, *Greedy Relevant Set Correlation* (GreedyRSC), serves as but one example of a practical application of the RSC clustering model. As GreedyRSC will turn out to rely heavily on cluster reshaping for computing final cluster memberships, henceforth only first-order formulations will be considered when discussing RSC significances, self-confidences and self-correlations.

The overall strategy and design choices of GreedyRSC resemble those of the PM-based PatClust heuristic method of [11]: both heuristics employ a greedy strategy for cluster selection whereby candidates with the highest quality are selected first, and any candidates found to be overly-similar to a previously-selected candidate are declared to be redundant, and then eliminated.

For the sake of efficiency and scalability, both PM and GreedyRSC incorporate the following additional heuristic design choices:

- Both methods avoid the quadratic cost of cluster quality evaluation by strictly limiting the size of all relevant sets considered to be at most some constant $b > 0$.

- Both allow the discovery of clusters of arbitrarily-large size by first computing small tentative clusters with respect to a range of data samples of varying sizes. PatClust uses these tentative clusters (called *patches*) as estimates of full-sized clusters, but does not provide the full contents of these clusters. GreedyRSC, on the other hand, treats the tentative clusters as *patterns* for the explicit generation of full-sized clusters, by reshaping the tentative clusters with respect to the full dataset using the techniques of Section 2.7.

- Both methods limit the number of candidate clusters considered by using only relevant sets of sample items as the eligible candidate patches or patterns.

The GreedyRSC method also seeks to reduce the total size and number of candidate cluster sets generated, by eliminating redundant patterns and cluster candidates at intermediate stages of the clustering process.

The use of sampling for RSC-based clustering can be intuitively justified as follows. Let C be a true (unknown) cluster of high quality, as evidenced by its meeting a high minimum first-order self-confidence threshold. A high quality score implies that the relevant sets of many cluster members are in mutual agreement to a high degree, so much so that if the set C were replaced by one of these relevant sets (call it $C' = \mathrm{Q}(q, |C|)$), that the remaining elements would likely still be in agreement with it. Restricting the relevant set items (including C') to a sample of the dataset still has the potential for discovering these agreements if the intersections between the relevant sets and the sample are sufficiently large. More precisely, we consider relevant sets of fixed size $t = \frac{m|C|}{n}$ with respect to a sample of size m taken from the full dataset (of size n), and focus our attention on $C'' = \mathrm{Q}''(q, t)$, where $\mathrm{Q}''(q, t)$ denotes the t items most relevant to q within the sample. The self-confidence value of C'', using relevant sets of size t drawn from the sample, serves as an estimate of the self-confidence value of C, using relevant sets of size $|C|$ drawn from the full dataset. In this fashion, C'' serves as a pattern from which the members of C can be estimated, by reshaping C'' with respect to the full set as described in Section 2.7.

If we are to obey the restriction that all relevant sets be limited in size to at most some constant, then in order to discover C, the sample sizes should be chosen so that for at least one sample, the value t falls into a constant-sized range. One way of covering all possible values of t (and thereby allowing the discovery of clusters of arbitrary size) is to create a hierarchy of subsets $H = \{S_0, S_1, \ldots, S_{h-1}\}$ by means of uniform random sampling, such that:

- S_0 is identical to S, and $S_i \subset S_{i-1}$ for all $0 < i \leq h - 1$;

- the number of samples h is chosen to be the largest integer such that $|S_{h-1}| > c$, for some constant $c > 0$;

- the size of S_i is equal to $\lfloor \frac{|S|}{2^i} \rfloor$ for all $0 \leq i \leq h - 1$;

- the pattern sizes t covered by sample i fall in the range $0 < a < t < b$, where a and b are chosen such that $b > 2a$.

This last condition ensures that all cluster sizes between a and $b2^{h-1}$ are covered by some pattern size with respect to at least one of the samples.

To support the sampling heuristic, for each sample S_i, we assume the existence of an oracle O_i that accepts any query item $q \in S$, and returns a ranked relevant set consisting of b items of S_i. The samples sets can optionally be selected and maintained by the oracles themselves.

As a final observation regarding the benefits of sampling, we note that a reasonable restriction on inter-cluster similarity implies that only one pattern need be retained for any given item-sample combination. For any item q, we have $\text{CM}(\text{Q}(q,a),\text{Q}(q,b)) = \sqrt{\frac{a}{b}}$, even when the relevant sets are drawn from a sample of the dataset. If a maximum threshold value χ is placed the on the allowable cosine value between any two clusters (including patterns), if $a \geq b\chi^2$, then at most one choice of pattern size can be made for any q with respect to any given sample. For example, the condition holds for the convenient choices $b = 4a$ and $\chi \leq 0.5$. In the overview of the GreedyRSC method below, we will assume that these parameters have been chosen so as to justify the retention of no more than one pattern per item-sample combination.

3.3 The GreedyRSC heuristic

1. For each sample set S_i, do the following:

   (a) *Relevant sets.*
   For each item $q \in S$, use oracle O_i to generate a relevant set $R_{q,i}$ for q with respect to the set S_i, such that $|R_{q,i}| = b$ for some constant $0 < b < c$.

   (b) *Inverted relevant sets.*
   Produce a collection of inverted relevant sets $I_{v,i}$, where $q \in I_{v,i}$ if and only if $v \in R_{q,i}$.

   (c) *Pattern generation.*
   Let $R_{q,i,t} \subseteq R_{q,i}$ denote the relevant set consisting of the t highest-ranked items of $R_{q,i}$, for any $0 < t \leq b$. Compute the value of t that maximizes the significance score $Z_1(R_{q,i,t})$ over all $a \leq t \leq b$. Let $P_{q,i}$ be the set at which the maximum is attained. If $a < |P_{q,i}| < b$ and if the significance score meets the minimum threshold value, then designate $P_{q,i}$ as the pattern of q with respect to sample S_i (otherwise, q is not assigned a pattern with respect to S_i).

   (d) *Redundant pattern elimination.*
   Iterate through the patterns of S_i in decreasing order of significance. For pattern $P_{v,i}$, use the inverted relevant sets $I_{*,i}$ to determine all other lower-ranked patterns sharing items with $P_{v,i}$. If the inter-set significance score $Z_1(P_{v,i}, P_{w,i})$ exceeds the maximum threshold value, then delete $P_{w,i}$.

   (e) *Pattern reshaping.*
   For every surviving pattern $P_{v,i}$, use the inverted relevant sets $I_{*,i}$ to determine those items w from the full dataset for which R_{w,i,p_v} shares members with $P_{v,i}$, where p_v denotes the size of $P_{v,i}$. Sort these items in decreasing order of their item-to-set significance with $P_{v,i}$, namely $Z_1(P_{v,i}, R_{w,i,p_v})$. Let $C_{v,i,t}$ denote the set consisting of the t highest-ranked items in this ordering. Reshape the pattern into a cluster candidate set by computing the value of t that maximizes the significance score $Z_1(P_{v,i}, C_{v,i,t})$; let $C_{v,i}$ denote this cluster candidate.

   (f) *Redundant cluster candidate elimination.*
   Iterate through the cluster candidates in decreasing order of the significance $Z_1(P_{v,i}, C_{v,i})$. For candidate $C_{v,i}$, use inverted cluster membership lists to determine all other lower-ranked candidates sharing items with $C_{v,i}$. If the inter-set significance score $Z_1(C_{v,i}, C_{w,i})$ exceeds the maximum threshold value, then delete $C_{w,i}$.

2. *Integration across samples.*
   Sort all surviving cluster candidates produced across all samples S_i, in decreasing order of the significance scores $Z_1(P_{v,i}, C_{v,i})$. For candidate $C_{v,i}$, use inverted cluster membership lists to determine all other lower-ranked candidates sharing items with $C_{v,i}$. If the inter-set significance score $Z_1(C_{v,i}, C_{w,i})$ exceeds the maximum threshold value, then delete $C_{w,i}$.

3.4 Complexity analysis

Over all executions of step 1(a), a query to the oracle is made for each item of S with respect to each sample set S_i. In general, if $\phi(i)$ represents the average cost of the queries taken over set S_i, the total cost is proportional to $n \sum_{i=0}^{h-1} \phi(i)$. If the oracle is implemented as a distance-based ranking using sequential search, the total number of distances computed would be no more than $O(n^2)$. However, if fast approximate search structures are used to limit the cost of an individual query, a lower complexity can be realized. For example, limiting the average query time $\phi(i)$ to be $O(b + h - i)$ results in an overall cost of $O(bn \log n + n \log^2 n)$ distances computed.

Producing the inverted relevant sets in step 1(b) requires a total of $O(bn \log^2 n)$ operations. For each item, with respect to each sample, determining the candidate pattern size in step 1(c) requires $O(b^2)$ operations, for a total of $O(b^2 n \log n)$.

The elimination of redundant patterns in step 1(d) requires the intersection to be computed between $P_{v,i}$ and every other pattern containing at least one member of $P_{v,i}$, as determined using the inverted lists for the members of P_{v_i}. If $\psi_{w,i}$ is the size of the inverted member list for item $w \in S_i$, then the total number of contributions to intersections that can be ascribed to w is no more than $\psi_{w,i}^2$. Summing these contributions over all items of S_i, and noting that the average inverted list size is bounded by b, we obtain $\sum_{w \in S} \psi_{w,i}^2 \leq (b^2 + \sigma_i^2)n$, where σ_i^2 is the variance of the sizes of the inverted member lists of members of S_i. Letting $\sigma^2 = \frac{1}{h} \sum_{0 \leq i < h} \sigma_i^2$ be the average of these variances, we can bound the total cost of this step by $O((b^2 + \sigma^2)n \log n)$.

The cluster reshaping step 1(e) is performed by finding all patterns $P_{w,i}$ intersecting $P_{v,i}$, computing their correlations with $P_{v,i}$, and then sorting the correlations. The bound on the cost of eliminating redundant patterns in step 1(e) also applies to this step, except for the additional work of sorting the accumulated correlations. The total number of items to be sorted for each sample S_i is at most bn, the total size of all member lists. The total cost of sorting correlations over all samples is thus $O(bn \log(bn) \log n)$. Since $\log b$ is of order $o(\log n)$, this simplifies to $O(bn \log^2 n)$.

The cost of eliminating redundant cluster candidates in step 1(f) can be accounted for in a similar manner as for patterns in step 1(d), with clusters $C_{v,i}$ taken in place of patterns $P_{v,i}$. Here, let $\xi_{v,i}$ be the size of the inverted cluster membership list associated with v at the time of execution of step 1(f) for sample S_i. Letting τ_i^2 be the variance of the values of $\xi_{v,i}$ over all $v \in S_i$, and noting that the average inverted list size remains bounded by b, we observe that the cost for sample S_i is of order $O((b^2 + \tau_i^2)n)$. Letting $\tau^2 = \frac{1}{h} \sum_{0 \leq i < h} \tau_i^2$ be the average of the variances over all samples, we obtain a bound for the total cost of this step in $O((b^2 + \tau^2)n \log n)$. The bounds for steps 1(e) and 1(f) also apply to the final candidate pruning performed in step 2.

Overall, disregarding the preprocessing time required for computing relevant sets, the execution time for GreedyRSC is bounded by $O((b^2 + \sigma^2 + \tau^2)n \log n + bn \log^2 n)$. Since the standard deviations σ_i and τ_i are typically of the order of the mean b in practice, σ and τ can also be estimated as roughly $\tilde{O}(b)$, for an overall cost bound of $\tilde{O}(b^2 n \log n + bn \log^2 n)$. The observed cost is dominated by the computation of relevant sets in step 1(a), and the first phase of redundant cluster candidate elimination in step 1(f).

4 Experimental Results

The GreedyRSC heuristic was implemented and tested on a dataset consisting of 378,659 sparse feature vectors on 40,000 attributes, each vector representing a bacterial open reading frame (protein sequence). Each vector value was derived from a BLAST similarity score of the sequence with respect to a reference sequence drawn from the same set, according to the methods outlined in [2, 15]. Values below a certain minimum threshold were zeroed out, producing an average of roughly 125 non-zero attributes per vector.

The role of the query oracle was played by a SASH approximate similarity search structure, using the vector angle as the pairwise similarity measure. The SASH was chosen due to its ability to handle data of extremely high dimensionality directly, without recourse to dimensional reduction techniques. The node degree of the SASH was set to 16. The SASH query performance was then tuned to a speedup of roughly 30 times over sequential search, for a recall rate of approximately 75%. The maximum pattern size was set to $b = 100$. For more details on the SASH search structure and its uses, see [12].

For the implementation, a cluster candidate C was selected only if it met minimum thresholds on two parameters: self-confidence and normalized squared set significance. The normalized squared set significance is obtained from the set significance $Z_1(C)$ or $Z_1(C, C')$ by dividing by $\sqrt{|S_i| - 1}$ and then squaring the result; here, S_i is the sample from which the cluster pattern derives. For the purposes of comparing the significance of clusters deriving from the same sample, or for cluster reshaping, the outcome when using normalized squared significance is the same as for the original first-order set significance. However, the normalized squared significance is interesting in that it equals $|C|$ whenever the self-confidence of C equals one. Setting a normalized squared significance threshold of k is thus able to produce clusters of size as small as k, provided that the relevant sets of their items are in perfect agreement. In the experiments, the minimum threshold value was chosen as 4. The clusters were also required to achieve a minimum self-confidence score of 0.4.

Cluster similarity was assessed by means of normalized inter-set significance (that is, the set correlation). A maximum threshold of 0.5 was applied.

The GreedyRSC implementation produced 8910 clusters for the protein sequence dataset, an average of one cluster for roughly every 42.5 sequences. Figure 4 shows a plot of the sizes of the result clusters, sorted from largest to smallest. In the same diagram, the normalized squared significances of the clusters are also plotted. Most of the clusters follow the Zipf distribution (the exceptions being the very largest clusters produced). The C++ implementation required approximately 18 hours of computation on a 3.0GHz single-processor workstation running Windows XP, as well as roughly 1.6Gb of main memory and 22Gb of disk space.

As an example of the clusters produced by GreedyRSC, the contents of the cluster of median size are listed in Figure 3. With respect to the cluster, the rank and normalized significance (correlation) of each member sequence is given, together with the cumulative normalized squared significance that would result if the cluster boundary were drawn just after the sequence in question. Finally, annotations are given for the sequences where available. Note that the function of many of the sequences in the cluster are unknown, although the top-ranked cluster members are otherwise in agreement. Clusters for which some of the sequences have unknown functions, but the remainder have a common functionality, are particularly interesting in that the unknown functionalities can be predicted in light of the known functionalities.

5 Discussion and Conclusion

We conclude with a discussion of potential applications and other issues involving the RSC model that merit further investigation.

5.1 Classification

As a byproduct of the GreedyRSC cluster reshaping step, every item whose relevant set intersects a cluster pattern is listed as a potential member of the cluster; the reshaping step determines whether or not the item is accepted as a cluster member. Regardless of whether or not the item is accepted, the significance of its relationship with the cluster is computed. This information can be saved with each

Cluster size:	20
Normalized squared significance:	13.7042
Self-confidence:	0.827778
Pattern size:	9
Sample size:	378,659

Rank	*Corr*	*NSqSig*	*Annotation*
1	1.000	1.000	molybdopterin oxidoreductase, molybdopterin binding subunit — Molybdopterin oxidoreductase — Pyrobaculum aerophilum IM2
2	1.000	2.000	tetrathionate reductase subunit A — Molybdopterin oxidoreductase — Salmonella enterica subsp. enterica serovar Typhi CT18
3	1.000	3.000	TtrA — Molybdopterin oxidoreductase — Pasteurella multocida PM70
4	1.000	4.000	tetrathionate reductase subunit A — Molybdopterin oxidoreductase — Salmonella enterica subsp. enterica serovar Typhi Ty2
5	1.000	5.000	hypothetical protein — Enterococcus faecalis V583
6	1.000	6.000	hypothetical protein — Molybdopterin oxidoreductase — Aeropyrum pernix K1
7	1.000	7.000	putative tetrathionate reductase, subunit A — Molybdopterin oxidoreductase — Vibrio parahaemolyticus O3:K6 RIMD 2210633 chromosome 1
8	1.000	8.000	tetrathionate reductase complex, subunit A — Molybdopterin oxidoreductase — Salmonella typhimurium LT2
9	1.000	9.000	molybdopterin oxidoreductase, molybdopterin binding subunit, putative — Molybdopterin oxidoreductase — Archaeoglobus fulgidus DSM 4304
10	0.889	9.779	hypothetical protein — Sulfolobus tokodaii 7
11	0.778	10.343	hypothetical protein — Leptospira interrogans lai 56601 chromosome 1
12	0.778	10.915	hypothetical protein — Escherichia coli O157:H7 RIMD 0509952
13	0.778	11.491	hypothetical protein — Zn-finger, C2H2 type — Sulfolobus solfataricus P2
14	0.778	12.071	hypothetical protein — Escherichia coli O157:H7 EDL933
15	0.667	12.452	Oxydoreductase, putative — Sulfolobus solfataricus P2
16	0.667	12.840	hypothetical protein — Protein of unknown function DUF192 — Agrobacterium tumefaciens C58 (Cereon) circular chromosome
17	0.667	13.235	hypothetical protein — Protein of unknown function DUF192 — Agrobacterium tumefaciens C58 (U. Washington) circular chromosome
18	0.556	13.443	50S ribosomal protein L36 — Ribosomal protein L36 — Blue (type 1) copper domain — Tropheryma whipplei TW08/27
19	0.556	13.661	hypothetical protein — Sulfolobus tokodaii 7
20	0.444	13.704	putative isomerase — Salmonella enterica subsp. enterica serovar Typhi Ty2
21	0.222	13.404	hypothetical protein — Neisseria meningitidis MC58
22	0.222	13.136	hypothetical protein — Streptococcus agalactiae 2603V/R
23	0.222	12.896	Unknown — Streptococcus agalactiae NEM316
24	0.222	12.679	hypothetical protein — Neisseria meningitidis MC58
25	0.222	12.484	hypothetical protein — Streptococcus pyogenes MGAS315

Figure 3: Details of the median-size protein sequence dataset cluster. *Rank* denotes the rank of the protein sequence in the cluster; *Corr* denotes the correlation (significance) between the sequence and the cluster; *NSqSig* denotes the normalized squared set significance of the cluster if the boundary were to be drawn after the current protein sequence; *Annotation* denotes annotation information available for the protein sequence.

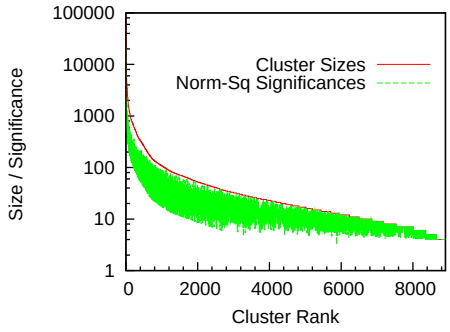

Figure 4: Plot of cluster size versus cluster rank, together with normalized squared significance. The maximum cluster size is 69859, the minimum size is 4. 8910 clusters were produced.

item as it is generated, to produce a list of associated clusters for the item together with the significances of the associations.

Consequently, the GreedyRSC can be adapted to provide a classification of the dataset items, in which each item is assigned to the cluster to which its item-to-cluster significance is highest. The classification technique is in some sense similar to nearest-neighbor classification (see [3] for a general reference), with set correlations playing the role of distances.

5.2 Postprocessing of query results

The reliance of RSC-based clustering on relevant set information for dataset items makes it particularly-well suited for the clustering of query result sets for information retrieval applications. If relevant sets have been precomputed and stored for every item in the database, large query results can be organized into clusters before presentation to the user, allowing greater ease of comprehension.

Alternatively, items from the result set can be classified in advance with respect to a pre-existing dataset clustering by means of item-to-cluster significances, as stated above.

5.3 Cluster mapping

In the final step of the GreedyRSC heuristic, the surviving clusters arising from each of the data samples are combined, with over-similar cluster pairs having one of the two clusters deleted. As a byproduct, all cluster pairs having inter-set significance scores above a minimum threshold can be recorded. Taken together, these relationships give rise to a graph of significant cluster relationships, in which each cluster is represented by a node of the graph, and each edge indicates a significant inter-cluster relationship. The notion of a cluster mapping was originally proposed in [11] in the context of PM-based clustering, and examples of such graphs can be found there.

5.4 Integration of clustering results

The RSC model provides a means for assessing the significance of the relationship between any two candidate cluster sets. As such, it can be used to assess the interrelationships between different clusterings of the same data set, particularly when two or more clusterings are to be integrated, even when the clusterings are produced by different methods. When a pair of clusters is found to have inter-set significance above a maximum threshold value, the cluster having smaller set significance can be deleted, or the contents of the two clusters can be merged. The clustering integration technique could potentially be used to track significant changes to datasets that may occur over time as a result of updates, to assess the impact of different similarity measures on classification and search performance, or to establish relationships between data clusters and the groupings of a pre-existing taxonomy.

5.5 Clustering in the presence of constraints

The reliance of RSC-based clustering on relevant sets instead of neighborhoods or similarity measures allows for clustering subject to additional constraints, provided that the constraints can be reflected in the relevancy rankings of individual items. For example, if it is known that items v and w should always appear together in any clustering, the relevancy rankings of items could conceivably be adjusted so as to encourage this. Traditional distance measures, due to their continuous and spatial natures, generally cannot easily be adjusted to handle special cases in the way that discrete relevancy rankings can. A clustering method that can take additional constraints into account have already been developed based on fuzzy k-means [7].

Acknowledgments

Thanks go to Yasumasa Shigemoto of the DNA Data Bank of Japan (DDBJ) for preparing the protein sequence vector set.

References

[1] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, *Proc. 20th VLDB Conference*, Santiago, Chile, 1994, pp. 487–499.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, A basic local alignment search tool, *J. Molecular Biology* 215:403–410, 1990.

[3] R. O. Duda, P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd ed., Wiley Interscience, New York, NY, USA, 2001.

[4] L. Ertöz, M. Steinbach and V. Kumar, A new shared nearest neighbor clustering algorithm and its applications, *Proc. Workshop on Clustering High Dimensional Data and its Applications* (in conjunction with *2nd SIAM International Conference on Data Mining* (ICDM 2002)), Arlington, VA, USA, 2002, pp. 105–115.

[5] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, *Proc. 2nd International Conference on Knowledge Discovery and Data Mining* (KDD), Portland, OR, USA, 1996, pp. 226–231.

[6] M. R. Garey and D. S. Johnson, *Computers and Intractability : A Guide to the Theory of NP-Completeness*, W. H. Freeman & Co., New York, USA, 1979.

[7] N. Grira, M. Crucianu and N. Boujemaa, Active semi-supervised fuzzy clustering for image database categorization, *Proc. 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, Singapore, 2005, pp. 9–16.

[8] S. Guha, R. Rastogi and K. Shim, CURE: an efficient cluster algorithm for large databases, *Proc. ACM SIGMOD Conference on Management of Data*, New York, USA, 1998, pp. 73–84.

[9] S. Guha, R. Rastogi and K. Shim, ROCK: a robust clustering algorithm for categorical attributes, *Information Systems* 25(5):345–366, 2000.

[10] J. Han and M. Kamber, *Data Mining: Concepts and Techniques* (2nd ed.), Morgan Kaufmann, San Francisco, CA, USA, 2006.

[11] M. E. Houle, Navigating massive data sets via local clustering, *Proc. 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (KDD), Washington DC, USA, 2003, pp. 547–552.

[12] M. E. Houle and J. Sakuma, Fast approximate similarity search in extremely high-dimensional data sets, *Proc. 21st IEEE International Conference on Data Engineering* (ICDE), Tokyo, Japan, 2005, pp. 619–630.

[13] R. A. Jarvis and E. A. Patrick, Clustering using a similarity measure based on shared nearest neighbors, *IEEE Transactions on Computers* C-22(11):1025–1034, November 1973.

[14] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, John Wiley & Sons, New York, USA, 1990.

[15] L. Liao and W. S. Noble, Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships, *J. Computational Biology* 10(6):857–868, 2003.

[16] J. McQueen, Some methods for classification and analysis of multivariate observations, *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[17] W. Wang, J. Yang and R. Muntz, Efficient and effective clustering methods for spatial data mining, *Proc. 23rd VLDB Conference*, Athens, Greece, 1997, pp. 186–195.

[18] T. Zhang, R. Ramakrishnan and M. Livny, BIRCH: an efficient data clustering method for very large databases, *Proc. ACM SIGMOD Conference on Management of Data*, Montréal, Canada, 1996, pp. 103–114.