Constant Time Search and Retrieval in Big Data, with Linear Time and Space Preprocessing, through Randomly Projected Piling and Sparse Ultrametric Coding

Fionn Murtagh, Pedro Contreras
International Conference
p-ADIC MATHEMATICAL PHYSICS AND ITS APPLICATIONS
p-ADICS.2015, September 2015
Topics

1. The conventional and a new approach in using random projection.
   • Conventional: random projection for optimal subspace mapping (based on Johnson-Lindenstrauss Lemma); and also Power Iteration Clustering (Lin & Cohen, Võhandu et al.)

2. Motivation and objectives: to exploit the remarkable simplicity of very high dimensional spaces.
   • This is through data piling and concentration in very high dimensions.

3. First steps towards sparse p-adic encoding of data.
• Previous work has included clustering redshift values (in astrophysics, redshift being related to the recession velocity, lookback time, 3rd dimension apart from RA and dec.), where redshifts were both photometric and spectrometric.

• Also this work included over 1000-dimensional chemoinformatics data, where chemical materials were “fingerprinted” in accordance with their molecular makeup.
In summary

• We use random projections of our data – of our cloud of points. When in high dimensions, such projections can even approximate a projection into a lower dimension orthonormal space. (This is what we want in PCA or CA, etc.)

• But we have pursued a new objective: to use random projection in order to approximate our data cloud such that it is rescaled well. What we mean by that: that its clustering properties are well respected – the interrelationships among points in our cloud of points.
In summary

• Having rescaled our data, based on random projections, we next show how we can simplify that mapping of our data cloud.

• Then we want to read off the clusters. We show that the Baire metric, that is also an ultrametric, is an excellent framework for this. (The Baire metric, as will be shown, is the “longest common prefix metric”.)
Applications in Search and Discovery

- First, agglomerative hierarchical clustering; or: “hierarchical encoding” of data.
- Ultrametric topology, Baire distance.
- Clustering of large data sets.
- Hierarchical clustering via Baire distance using SDSS (Sloan Digital Sky Survey) spectroscopic data.
- Hierarchical clustering via Baire distance using chemical compounds.
Next: the Baire (ultra)metric
An example of Baire distance for two numbers \((x \text{ and } y)\) using a precision of 3:

\[
x = 0.425
\]

\[
y = 0.427
\]

Baire distance between \(x\) and \(y\):

\[
\bar{d}_B (x, y) = 10^{-2}
\]

Base \((B)\) here is 10 (suitable for real values)

Precision here = \(|K| = 3\)

Exponent:

\(k-1 \text{ s.t. } \arg\min_k x_k \neq y_k\)
On the Baire (ultra)metric

- Baire space consists of countable infinite sequences with a metric defined in terms of the longest common prefix [A. Levy. Basic Set Theory, Dover, 1979 (reprinted 2002)]

- The longer the common prefix, the closer a pair of sequences.

- The Baire distance is an ultrametric distance. It follows that a hierarchy can be used to represent the relationships associated with it. Furthermore the hierarchy can be directly read from a linear scan of the data. (Hence: hierarchical hashing scheme.)

- We applied the Baire distance to: chemical compounds, spectrometric and photometric redshifts from the Sloan Digital Sky Survey (SDSS), and various other datasets.
• A subset was taken of approximately 0.5 million data points from the SDSS release 5.
• These were objects with RA and Dec (Right Ascension and Declination, and spectrometric redshift, and photometric redshift). Problem addressed: regress one redshift (spectro.) on the other (photo.).
• Baire approach used, and compared with k-means.
• 1.2 million chemical compounds, each characterized by 1052 boolean presence/absence values.
• Random projections used on normalized compound/attribute values.
• Baire approach used; also another approach based on restricting the precision of the normalized compound/attribute values.
We took a subset of approximately 0.5 million data points from the SDSS release 5 [reference: D’Abrusco et al]

- declination (Dec)
- right ascension (RA)
- spectrometric redshift
- photometric redshift.

Dec vs RA are shown
## Data – sample

<table>
<thead>
<tr>
<th>RA</th>
<th>DEC</th>
<th>spec. redshift</th>
<th>phot. redshift</th>
</tr>
</thead>
<tbody>
<tr>
<td>145.4339</td>
<td>0.56416792</td>
<td>0.14611299</td>
<td>0.15175095</td>
</tr>
<tr>
<td>145.42139</td>
<td>0.53370196</td>
<td>0.145909</td>
<td>0.17476539</td>
</tr>
<tr>
<td>145.6607</td>
<td>0.63385916</td>
<td>0.46691701</td>
<td>0.41157582</td>
</tr>
<tr>
<td>145.64568</td>
<td>0.50961215</td>
<td>0.15610801</td>
<td>0.18679948</td>
</tr>
<tr>
<td>145.73267</td>
<td>0.53404553</td>
<td>0.16425499</td>
<td>0.19580211</td>
</tr>
<tr>
<td>145.72943</td>
<td>0.12690687</td>
<td>0.03660919</td>
<td>0.06343859</td>
</tr>
<tr>
<td>145.74324</td>
<td>0.46347806</td>
<td>0.120695</td>
<td>0.13045037</td>
</tr>
</tbody>
</table>
• Motivation - regress \( z_{\text{spect}} \) on \( z_{\text{phot}} \)
• Furthermore: determine good quality mappings of \( z_{\text{spect}} \) onto \( z_{\text{phot}} \), and less good quality mappings
• I.e., cluster-wise nearest neighbour regression
• Note: cluster-wise not spatially (RA, Dec) but rather within the data itself
On the left we have $z_{\text{spec}}$ where three data peaks can be observed. On the right we have $z_{\text{phot}}$ where only one data peak can be seen.
• **82.8%** of \( z_{\text{spec}} \) and \( z_{\text{phot}} \) have at least 2 common prefix digits.
  
  • I.e. numbers of observations sharing 6, 5, 4, 3, 2 decimal digits.

• **We can find very efficiently where these 82.8% of the astronomical objects are.**

• **21.7%** of \( z_{\text{spec}} \) and \( z_{\text{phot}} \) have at least 3 common prefix digits.
  
  • I.e. numbers of observations sharing 6, 5, 4, 3 decimal digits.
• Next - another case study, using chemoinformatics - which is high dimensional.

• Since we are using digits of precision in our data (re)coding, how do we handle high dimensions?
Baire Distance Applied to Chemical Compounds
Matching of Chemical Structures

- Clustering of compounds based on chemical descriptors or chemical representations, in the pharmaceutical industry.
- Used for screening large corporate databases.
- Chemical warehouses are expanding due to mergers, acquisitions, and the synthetic explosion brought about by combinatorial chemistry.
Binary Fingerprints

Fixed length bit strings with encoding schemes
Daylight, MDL
BCI  *(We will be using this)*
Chemoinformatics clustering

• 1.2 million chemical compounds, each characterized by 1052 boolean presence/absence values.

• Firstly we note that precision of measurement leads to greater ultrametricity (i.e. the data are more hierarchical).

• From this we develop an algorithm for finding equivalence classes of specified precision chemicals. We call this: data “condensation”.

• Secondly, we use random projections of the 1052-dimensional space in order to find the Baire hierarchy. We find that clusters derived from this hierarchy are quite similar to k-means clustering outcomes.
Random projection and hashing

In fact random projection here works as a class of hashing function.

Hashing is much faster than alternative methods because it avoids the pairwise comparisons required for partitioning and classification.

If two points \((p, q)\) are close, they will have a very small \(|p-q|\) (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.
• Normalize chemical compounds by dividing each row by row sum (hence “profile” in Correspondence Analysis terms).
• Two clustering approaches studied:
• Limit precision of compound / attribute values. This has the effect of more compound values becoming the same for a given attribute. Through a heuristic (e.g. interval of row sum values), read off equivalence classes of 0-distance compounds, with restricted precision. Follow up if required with further analysis of these crude clusters. We call this “data condensation”. For 20000 compounds, 1052 attributes, a few minutes needed in R.
• Second approach: use random projections of the high dimensional data, and then use the Baire distance.
Summary Remarks on Search and Discovery

- We have a new way of inducing a hierarchy on data

- First viewpoint: encode the data hierarchically and essentially read off the clusters

- Alternative viewpoint: we can cluster information based on the longest common prefix

- We obtain a hierarchy that can be visualized as a tree

- We are hashing, in a hierarchical or multiscale way, our data

- We are targeting clustering in massive data sets

- The Baire method - we find - offers a fast alternative to k-means and a fortiori to traditional agglomerative hierarchical clustering

- At issue throughout this work: embedding of our data in an ultrametric topology
• Quite a different starting point:
• Using Apache Lucene and Solr for indexing, storage, and query support
• The following slide is showing where we used 152,998 cooking recipes, with 101,060 unique words in them.
247 attributes, 8 of them shown

cake
desserts
cheese
mexican
italian
seafood
chinese
salads

Find:

152298 results found in 9 ms Page 1 of 15300

"21" Club Rice Pudding More Like This
Id: mm000001102.txt
F1 (xcoord): -0.7341409 F2 (ycoord): -0.09961348
Recipe: Categories: Dessert Yield: 10 Servings 1 qt Milk 1 pt Heavy cream 1/2 ts Salt 1 Vanilla bean 3/4 c Long-grained rice 1 c Granulated sugar 1 Egg yolks 1 1/2 c Whipped cream Raisins (optional) From: Bobb1744 at sol.com Date: Tue, 23 Apr 1996 13:28:27 -0400 Recipe By: Aunt Sally's In a heavy saucepan, combine the milk, cream, salt, vanilla bean and 3/4 cup of the sugar and bring to a boil. Stirring well, add the rice. Allow the mixture to simmer gently, covered, for 3 1/2 hours over a very low flame, until rice is soft. Remove from the heat and cool slightly. Remove the vanilla bean. Blending well, stir in the remaining 1/4 cup of sugar and the egg yolks. Allow to cool a bit more. Preheat the broiler. Stir in all but 2 tablespoons of the whipped cream; pour the mixture into individual crocks or a souffle dish. (Raisins may be placed in the bottom of the dishes, if desired.) After spreading the remaining whipped cream in a thin layer over the top, place the crocks or dish under the broiler until the pudding is lightly browned. Chill before serving. MC-RECIPE at MASTERCOOK.COM MASTERCOOK RECIPES LIST SERVER MC-RECIPE DIGEST V1 No.55 From the MasterCook recipe list. Downloaded from Glen's MM Recipe Archive, http://www.erols.com/mosey. 

"A Greeting" More Like This
Id: mm000001103.txt
F1 (xcoord): 0.6679912 F2 (ycoord): -0.01193458
Recipe: Categories: None Yield: 1 Servings Recipe by: steeleman at execpc.com VLF Vegetarian Recipes This is a kind of very-low-fat vegetarian starter's kit, containing recipes and other information (category "text") collected from a variety of computer sources, and our own experiences. The category marked "family approved" are recipes we use regularly, that we all like. The others were collected, but for the most part have not been tried. For a great collection of recipes, we recommend michelle dick's www.fatfree.com, and the Usenet newsgroup alt.food.fat-free. This type of diet was recommended by our cardiologist to help reverse Merle's heart disease. Whether it is working for that is not yet known, but we both DO feel healthier eating this way. Also, we find the food very satisfying and filling, but low-calorie enough to allow for significant weightloss. And eating this way is getting easier all the time. If you walk SLOWLY through the grocery store each week, you will see new "FAT-FREE" food items each week, increasing your choices. Mende and Deb Steelman steeleman at execpc.com 10/2/94 File ftn://ftp.scribner.co.uk/new/food/vegmaster/narratives/mmfatfrr.txt
From random projection to the Baire hierarchical clustering

- Selection of 10,317 funding proposals, out of set of 34,352, were indexed in Apache Solr. Their similarities were determined, using Solr’s MLT (“more like this”) score. (This uses weights for fields in the proposal documents, and is analogous to a chi squared, or tf-idf-based similarity.) (tf-idf: term frequency – inverse document frequency)

- We used a very sparse similarity matrix of dimensions 10317 x 34252.
- Through random projection, we obtained a unidimensional scaling of the 10317 proposals.

- In the following the mean of 99 random projections was used.
- The projection values were rescaled to the interval 0,1.
- Layer 1 clusters: the same first digit.
- Layer 2 clusters: given the same first digit, having the same second digit.
- Layer 3 clusters: given the same first two digits, having the same third digit.
- And so on.
- This is a regular 10-way tree.
Abscissa: 10118 documents sorted by random projection value. Ordinate: 8 digits comprising random projection value.
Layer 1: 8 clusters are very evident. Layer 2: there are 87 clusters (maximum is 100). Layer 3: here 671 clusters (maximum is 1000).
Low dimensional goodness of fit to our data, versus linear rescaling

• Conventional use of random projections:
  • Project data into lower dimension subspace, of dimension > 1.
  • Aim is to have proximity relations respected in the low dimensional fit to the high dimensional cloud of points.

• In our work, we seek a consensus one-dimensional mapping of the data, that represents relative proximity.

• Two following slides: Our aim is relative clustering properties. Cf. the now conventional use of the Johnson-Lindenstrauss lemma.
F. Critchley and W. Heiser, "Hierarchical trees can be perfectly scaled in one dimension", Journal of Classification, 5, 5-20, 1988.
Dimensionality reduction by random projection. To facilitate clustering in high dimensions

\[ F(x) : \mathbb{R}^d \rightarrow \mathbb{R}^k \]

Reduced dimensionality: \( k \ll d \)

Below:
Johnson-Lindenstrauss Lemma
Distance changes by a fraction \( 1 \pm \varepsilon \)

Lemma 1. For any \( 0 < \varepsilon < 1 \) and any integer \( n \), let \( k \) be a positive integer such that

\[ k \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \ln n. \]

Then for any set \( V \) of any points in \( \mathbb{R}^d \), there is a map \( f : \mathbb{R}^d \rightarrow \mathbb{R}^k \) such that for all \( u, v \in V \),

\[ (1 - \varepsilon) \| u - v \|^2 \leq \| f(u) - f(v) \|^2 \leq (1 + \varepsilon) \| u - v \|^2. \]

Furthermore, this map can be found in randomized polynomial time.

• In random projection matrix, each column is of unit norm. Values are 0-mean Gaussian. So – random Gaussian vectors.

• Reduced dimensionality space is not guaranteed to be in orthonormal coordinate system.

• Distortion of the variances/covariances relative to orthogonality of the random projections has approximate variance 2/m where m is low dimensionality.

• For sufficient m, orthonormal system is mapped into a near-orthonormal system.

• Kaski cites Hecht-Nielsen: the number of almost orthogonal directions in a coordinate system, that is determined at random in a high dimensional space, is very much greater than the number of orthogonal directions.
• Conventional random projections: random vectors that are iid 0-mean Gaussian. This is only necessary condition for preserving pairwise distances (Li, Hastie, Church, Proc. 12th ACM SIGKDD, 2006).

• Other work has used 0 mean, 1 variance, 4th moment =3. Also elements of random projection matrix from {-1,0,1} with different (symmetric in sign) probabilities.

• It is acknowledged that: “a uniform distribution is easier to generate than normals, but the analysis is more difficult”.
Our non-conventional approach to random projections

- Uniform \([0,1)\) valued vectors in the random projection matrix.
- Projections are rescaled to be in \([0,1)\), i.e. closed/open interval.
- Take mean (over random projections) of projected values.
- It is known from the central limit theorem, and the concentration, or data piling, effect of high dimensional data clouds, that: pairwise distances become equidistant, and orientation tends to be uniformly distributed.
- We find also: norms of the target space axes are Gaussian. (That is, before taking the mean of the projections.) As typifies sparsified data, the norms of the points themselves are negative exponential, or power law, distributed.
Scaling followed by clustering

- Correlation between most projection vectors $> 0.99$. We also found very high correlation between first principal component loadings and the mean random projection ($> 0.999999$).
- Our objective is less to determine or model cluster properties as they are in very high dimensions, than it is to extract useful analytics by “re-representing” the data. That is to say, we are having our data coded (or encoded) in a different way.
Summary Remarks on Reading Baire Distance Properties from the (Mean) Random Projected Values

- We have a new way of inducing a hierarchy on data
- First viewpoint: encode the data hierarchically and essentially read off the clusters
- Alternative viewpoint: we can cluster information based on the longest common prefix
- We obtain a hierarchy that can be visualized as a tree
- We are hashing, in a hierarchical or multiscale way, our data
- We are targeting clustering in massive data sets
- The Baire method - we find - offers a fast alternative to k-means and a fortiori to traditional agglomerative hierarchical clustering
- At issue throughout this work: embedding of our data in an ultrametric topology
• Visualization of Baire hierarchy.

• Means of 99 random projections.
• Abscissa: the 10118 (non-empty) documents are sorted (by random projection value).
• Ordinate: each of 8 digits comprising random projection values.
• Traditional clustering: use pairwise distances, determine clustering structure (hierarchy or optimization of criterion). Often: then a partition is determined.

• Here we build a series of partitions. Then the hierarchy is determined from them.
Next: Storage efficiency

Fit $m$-adic data by $p$-adic, $p < m$, through stepwise approximation.
Fit m-adic data by p-adic, $p < m$, through stepwise approximation

- Take 10-adic (decimal) array. Values, $v \in \mathbb{Z}_{10}$.
- Determine the minimum number of sequential values (e.g. 0 and 1, 1 and 2, 2 and 3, etc.).
  - At a given digit precision level,
  - consider successive pairs of values,
  - that have the same parent value.
  - Call the minimum value pair (over all distinct value pairs, over all digit precision levels) $w_1, w_2$. Define $w_2 = w_1 + 1$
- Update the set $\mathbb{Z}_{m-1} \leftarrow \mathbb{Z}_m$ by:
  - $\forall w \geq w_2$, $w \leftarrow w - 1$.
- Clearly this includes: $w_2 \leftarrow w_1$

When $p$ is prime, $\mathbb{Z}_p$ is a finite field, or Galois field.
5-adic approximation of Baire array display

Find:
Quality of $m$-adic approximation, has “elbow” in curve at $m = 5$, or conventionally, $P = 5$, 5-adic representation.
Next: Relationship with Lin & Cohen’s Power Iteration Clustering
Power Iteration Clustering (Lin & Cohen)

- F Lin, W Cohen, Power iteration clustering, Prof. 27th Intl. Conf. on Machine Learning, ICML (2010, Haifa, Israel), Omnipress, 655-662.
- A Lohk, O Tilk, L Võhandu, How to create order in large closed subsets of WordNet-type dictionaries, Eesti Rakenduslingvistika Ühingu aastaraamat, 9, 149-160.

(Following figures from latter.)
PIC applied to normalised rows, sum = 1 (unit mass), to iteratively determine dominant eigenvector. Through early halting of convergence, a constant, trivial, first eigenvector is avoided. Good clustering properties: Lin & Cohen use k-means on the PIC output.

**Procedure Power Iteration (A):**

1. Choose random nonzero vector $b_0 \neq 0$; $t = 0$;
2. Calculate $b_{t+1} = \frac{Ab_t}{\|Ab_t\|}$;
3. $t = t + 1$ and repeat step 2.

**Figure 4.** Description of power iteration eigenvalue algorithm
To avoid local optima, we have used, e.g. 60, 70, 80, 90, ... iterations. We also normalize using $L_\infty$ metric (Chebyshev) rather than $L_1$ so as to map the scaling onto $[0,1]$

**Procedure PIC($W, \epsilon$):**

1. Choose random vector $b_0 \neq 0 \land b_0 \neq c1; \quad d_0 = 1$
2. Calculate $b_{t+1} = \frac{Wb_t}{\|Wb_t\|_1}; \quad d_{t+1} = |b_{t+1} - b_t|$
3. If $\|d_{t+1} - d_t\|_\infty > \epsilon$, then $t = t + 1$ and repeat step 2; otherwise output $b_t$.

**Figure 5.** Description of the main subroutine of PIC algorithm
Our current directions in this work

(1) Benefits and operational practice of sparse p-adic coding for linear time (hierarchical) clustering and linear space storage.

(2) Correspondence analysis in very high dimensional and large data volume spaces.

(3) One area of application: applying J Habermas’s *Theory of Communicative Action* to areas of Big Data where behavioural or activity profiles can be contrasted with aggregate, collective or mean process outcomes.