Marc Teboulle
Charles Nicholas
Jacob Kogan

Clustering Large and High Dimensional Data

CIKM 2003
modern nonlinear optimization methods. This place particular attention on document clustering, and on the applications of
and an introduction to some recently developed clustering techniques. We
as more efficient, retrieval. The tutorial provides an overview of clustering,
suggests that document clustering should result in more effective, as well
documents tend to be relevant to the same requests (van Rijsbergen 1979)
document, known “cluster hypotheses”, for example, which says that closely associated
clustering has been a subject of interest in IR for many years. The well-
been applied in text mining and web page clustering, among others. Indeed,
formation retrieval, to name just a few. Within IR, clustering techniques have
design, data mining, bio-informatics (e.g. gene expression analysis), and in-
tering has been used in a variety of areas, including: computer vision, VLSI
having any background knowledge of the characteristics of the data. Clus-
data sets, and to identify abstract structures that might reside there, without
and available. Clustering techniques are used to discover natural groups in
large and often high dimensional data sets are now increasingly common

Abstract
Clustering software

Examples from IR

Hierarchical algorithms, e.g., single-link

Introduction and Basic concepts

Overview of this Tutorial
Clustering as an optimization problem

Partition algorithms, e.g., spherical

Non-hierarchical algorithms, e.g., k-means
Every computer scientist should know something about it!

- Entity
  - Terminology to lots of items, refer to them as a single (aggregate)

- Clustering is an important form of abstraction: instead of re-

- Disciplines
  - Practical applications abound, in computer science, and other

- Why study clustering?
2000, 2; 02, 2; 03, 1
SIGMOD 1990, 1; 85, 1; 90, 1; 91, 2; 92, 1; 96, 1; 98, 1; 99, 1;
CIKM 1993, 1; 95, 1; 98, 3; 99, 4; 2000, 4; 01, 2; 02, 5
Similar numbers from other conferences:

(Table on previous slide from Smelaton et al., SIGIR 2002)

Clustering is a hot area, in both the IR and database community,
to certain nodes. Topic, language, source, or other attributes, and then allocated.

In particular, in distributed systems, records can be clustered by

indexing and retrieval operations can be optimized.

If records can be clustered together in a sensible fashion, then

Why is clustering important in a database? or Information Re-

trieval?
cessing even comes up.
and what structure there is in it, before the issue of query pro-
Perhaps more important, clustering tells you about the corpus,
for many years.
How well the CH applies in practice has been a research question
the same requests."
states that closely associated documents tend to be relevant to
The "Cluster Hypotheses", proposed by van Rijsbergen in 1971,
Cluster Hypotheses
– amenable to parallelization

– scales as the dimensionality of points increases

– scales as the number of points increases

plexity

reasonable performance, in terms of time and space com-

•

it should produce "good" clusters, whatever that means

What do we want in a clustering algorithm?
Hierarchical algorithms produce a nested data set.

Clustering within clusters

Relatively slow

Graphical output via dendrogram, for example

Taxonomy of clustering algorithms
An example dendrogram from http://www.clustan.com
Hierarchical algorithm. Recursive application of a non-hierarchical algorithm results in a structure within clusters, if any, may not be apparent. •

can achieve linear time and space complexity. •

graphical output via colors on a CRT, for example. •

data

Non-hierarchical algorithms produce a simple partition of the taxonomy of clustering algorithms (continued).
(Cutting et al., SIGIR 92, SIGIR 93)

A divisive hierarchical algorithm is at the heart of scatter/gather.

Although agglomerative algorithms have received more attention, although

Hierarchical algorithms may be divisive, (top-down) or agglomer-

(bootstrap)
Hierarchical Agglomerative Clustering Methods (HACMS)

- Ward's method
- Group average link
- Complete link
- Single link
Tends to form "long straggly" clusters.

Easy to implement, and runs quickly once similarity matrix is computed.

Dendrogram: when the last two objects are joined, i.e. at the root of the tree, is based on the similarity of their most similar points. Stop single points (not yet in the same cluster. Similarity of objects or clusters, at each step, join the most similar pair of objects (clusters) or
% clear the old plots, if any
% if dimension is 2 or 3, we can do a plot

assert (m >= 2);

m = input('number of data points (vectors in R^n) to generate: ');

assert (n >= 2);

n = input('dimensionality of data space in (n >= 2): ');

--- %

The code shown here runs on Octave.

able on tutorial web site.

Refer to one page of code for SingleLink in handout. Also avail-
% keep the same plot for each pass
end
plot

plot [0:1:0] [0:1:0] d with points

plot [0:1:0] [1:0:0] d with points

set nkey

set size ratio 1

set size ratio 2

set size ratio 3

set size ratio 4

set size ratio 5

if n == 3

set nkey

plot the points in the matrix

end

plot

set parametric
delete
plot clear

plot

set nparametric

plot clear

if n == 2
\texttt{cluster} = \texttt{t};
\texttt{for} \texttt{t}=1: \texttt{w};
\texttt{cluster} = \texttt{zeros(t,w)};
\%
\texttt{initialize with each point in its own cluster}
\%
\texttt{make a vector which says which cluster a given point is in}
\%
\texttt{that are not yet in the same cluster}
\%
\texttt{at each step, join the most similar pair of objects (points or clusters)}
\%
\texttt{loop through the data}
\%
\texttt{endfor}
\%
\texttt{endfor}
\%
\texttt{for} \texttt{w} = 1: \texttt{t};
\texttt{for} \texttt{t}=1: \texttt{w};
\texttt{s} = \texttt{zeros(w,m)};
\%
\texttt{print} \texttt{message} \\
\%
\texttt{now build a distance matrix, using \texttt{euclidean distance} \\
\%
\texttt{end for} \\
\%
\texttt{hold on}
assert(cluster(c) == cluster(c));

if (cluster(c) == cluster(c) &&
    (cluster(c) <= n))

for i = 1:w:
    for j = 1:w:
        for k = 1:m:
            for l = 1:m:

                if (i == j) &&
                    (i == k) &&
                    (j == l)

                    printf("the two closest points are %d in cluster %d and %d in cluster %d\n", i, j, k, l);

        if (cluster(c) > closest) &&
            (closest == y) &&
            (j == y)

            closest = y;

        if (cluster(c) > closest) &&
            (closest == y) &&
            (j == y)

            closest = y;

    if (closest > closest) &&
        (closest == y) &&
        (j == y)

        closest = y;

assert(closest == closest);
for k=1:m

cluster(k) = newc;
if cluster(k) == oldc

for k=1:m

% closest? cluster
% connect the clusters by making each point in closest's cluster a member of

endfor
endfor
end

cluster = closest(cluster); oldc = closest(cluster);
newlink(2,:) = d(closest,2,new);
newlink(1,:) = d(closest,1,new);
newlink = zeros(2,n);
Code example available on tutorial website.

Tends to form compact clusters.

Performance is about the same.

Implementation is slightly more complicated than single-link, but

...joined, i.e., at the root of the dendrogram.

...least similar points. Stop when the last two objects are

...theirs cluster. Similarity of objects is based on the similarity of

At each step, join the most similar pair of objects not yet in the

Complete Link
more time, but significance less space. Furthermore, require a little
similarity computations on the fly, and therefore require a little
space in a naive implementation. Some algorithms allow for

To build the similarity matrix obviously requires \( O(n^2) \) time and

in each pass of the algorithm, similarity matrix will need to be modified as clusters are created.

of the similarity matrix depends on the data. Furthermore, the
terms in common have similarity zero. Therefore, the density
In document clustering, using cosine similarity, documents with
each pair of points in the data.
Hierarchical algorithms often need to know the similarity between

The Similarity Matrix
On random data, the two algorithms seem to produce very similar results.

Code is available on tutorial web site. (Demo Octave, if time permits)

Comparing single link vs. complete link

Compared to
Implementation very much like complete link.

is based on average similarity of component points.

like single link and complete link, except that similarity of objects
archy tends to produce homogeneous clusters and a symmetric hierarchy.

Smallest increase in total intra-cluster sum of squares that at each step, the two objects being joined result in the like the single link, complete link, and group average link, except Ward's Method.
Dijkstra's algorithm includes SLINK and Prim. These algorithms have known space and time complexities of $O(n^2)$, i.e., $O(u^2)$ time and $O(n)$ space, for which several optimal algorithms have been made to various HACMs, especially single-link improvements have been made to various HACMs, especially single-link improvements have been made to various HACMs.

The Lance-Williams dissimilarity update formula lets an algorithm calculate the dissimilarity between newly formed clusters.

(Complexity $O(n \log n)$)

Calculating a similarity matrix.

All rely on calculating (perhaps on-the-fly) and perhaps recall calculating the similarity between two clusters.

The algorithms are fairly similar, differing mostly in how they evaluate the HACMs.
e.g. incremental k-means

Redistribution lets documents move from one cluster to another.

batch k-means

Single pass is simple, but sensitive to order of input data. e.g.

Algorithms for Non-hierarchical methods
Explore the Corpus: Find clusters in order to better understand what documents (or topics or languages) are available.

Results Presentation: Process query $q$ as usual, but cluster results. Some search engines do this, but not very many.

Query Processing: Given a query $q$, work down the tree, taking branches where similarity with the query is greatest.

Decrease collection is a single document, or if the similarity starts to decrease until some stopping criterion is met, e.g. stop.

Use of clustering in IR.
ber of clusters, which were then labeled by inspection.

Similarity threshold was adjusted to produce a desirable num-

glomerative method.

Document vectors were clustered using a hierarchical, ag-

and similarity matrix calculated.

Stopwords were removed, and terms stemmed and weighted,

abstracts from 853 SIGIR papers were collected.

Smeaton's SIGIR paper study
tween clusters. The variance within clusters and maximizes the variance be-
across all 12 sensory variables. [the algorithm] minimizes ter when they have broadly the same taste characteristics. The clustering software grouped whiskies into the same class.

• The clustering software grouped whiskies into the same class.

• and the reviewers’ ratings were noted.

• Reviews of 88 different single malt whiskies were collected, from http://www.whiskyclassified.com/

• Twelve characteristics of whisky were identified.

What about the whisky example?
Since system is interactive, speed of clustering is critical.

The process is user-driven, divisive (i.e., top-down) and hierarchical.

Those clusters are gathered, clustered, and the process repeats.

More clusters for further study.

Short summaries are presented to user, who selects one or more clusters.

Data is divided into a small number of clusters own right, apart from retrieval.

Key idea: clustering can be effective as an access method in its Scatter/Gather
Twice.

Cluster centers may shift, so repeat assignment once or twice.

Distance

Assign each document to one of the clusters, e.g., by closest

Find k "centers" in the sample using e.g. group average link

documents

k is the number of clusters, and n is the total number of

\( \sqrt{\sum_{i=1}^{k} d_n} \)

where

Take a random sample of the data of size \( \sqrt{kn} \)

Bucketshot Algorithm
Clustering centers may shift, so repeat assignment once or twice.

- Assign each document to one of these k clusters, as in bucket.
- Repeat, until only k groups remain.


group algorithms. Treat these N/m groups as individuals, and
cluster the data in each of the buckets using some cluster.

Cluster the data in each of the buckets using some cluster.

With m < k.

- Divide the document set into N/m groups of fixed size m.

Partitioning Algorithm
better than \( O(\sqrt{n}) \) for \( n = 100000 \) which is a million times

Total cost is still \( O(\sqrt{n}) \) for \( n = 100 \), a factor of 10.

10 cluster centers are what we wanted. Cost: \( O(100^2) \), with a

Finding the centers of these 10 sets, and the resulting

Treat these 100000 “centers” as individuals, divide them into 10

Ink cost: \( O(100^2) \), with a factor of 10.

Finding the “center” of these sets, using e.g. group average

Make \( N/w = 1000 \) document sets, 100 documents in each set.

Centers, let \( m = 100 \). Example: start with \( N = 100000 \) documents. To choose \( k = 10 \)
for documents than for widgets :) 
Stratified vs. random sampling is perhaps even more important 
custers according to cutting et al, Fractionation seems to make better 
results. Buckshot can be run several times, in the hopes of getting better 
buckshot. Complexity is the same, but doesn’t have the randomness of 
Fractionation takes longer than buckshot, although asymptotic 
Fractionation vs. Buckshot
is greater than the minimum for that level of the hierarchy.

Each "satellite" document and the "star" at the cluster's center
exhibit a subgraph, where the distance between
possibly more than one (and

Find star-shaped subgraphs such that each document is one (and

threshold between pairs of documents.

Each level in the hierarchy is determined by a minimum similarity

A hierarchical technique for browsing an information collection.

Star Clusters - another IR application of clustering
5. Assess output, if needed (and it usually is - CKN)

4. Data abstraction, if needed

3. Clustering algorithm

2. Define a similarity measure

1. Choose a representation of the data

Components of a Clustering Task:

From Jain’s 1999 Survey on Data Clustering
Transformation e.g. Cartesian to polar coordinates for points in $\mathbb{R}^2$ •

Dimension reduction •

Feature selection •

Data representation •

37
spaces, then cluster. Lots of recent work on projecting data into lower dimensional spaces. What happens?

Terms in common, but term frequencies are identical. How could terms in the vector space representation misses, e.g., no similarities that the vector space may have. Documents “live” in high-dimensional spaces, and they may have.

High-dimensionality can be an issue.

Speaking of Dimension Reduction...
Simulation of data sets have been developed. (Jain, pg. 272)

For semi-structured data, e.g. text with nominal or ordinal val-

\[
\text{(I)} \quad \frac{\sum_{I=1}^{T} \sum_{j=1}^{T} \text{weight}_i \text{weight}_j}{\sum_{I=1}^{T} \text{weight}_i} = \frac{\lambda D_D S}{\lambda S_U}
\]

Similarity measures

For document clustering, one can use Cosine, Dice, or Jaccard

Similarity measures
satisfy the Triangle Inequality, but may still produce good results.

However, measures such as Mutual Neighbor Distance do not

\[ d(p, q) \leq d(p, c) + d(c, q) \]

and the triangle inequality symmetry: \( d(p, q) = d(q, p) \)

strict positivity of \( d \)

A metric space is a space in which a distance metric has the

sent as points on a hyper-sphere.

citation, including (length-normalized) documents when repres

Euclidean distance is another reasonable choice in many appli-

Similarly vs distance
Image processing papers use subjective assessment of test image similarity.

total intra-cluster similarity while minimizing total inter-cluster similarity as an optimization, e.g., find the cluster centers that maximize from an objective standpoint, cluster quality is usually expressed.

How does one measure the quality of a clustering?

Assessing the output: what is a "good" clustering?
Start Matlab and then type "kmeansdemo" in command win-
dow.

Get the package clusters from Matlab's site using clusters package in Matlab.
to get started, select "maintain help" from the help menu

using matlab, you can create type "help" to get started, type "help doc" for more information.
Pivoting refers to looking at variables across clusters, or looking at clusters across variables. Pivoting varying number of clusters, and pivoting of data. Need a Windows platform. Click on Clustan Graphics 6 icon. Using Clustan
10 \Windows\cluster sports.mat

cmd.exe From cluster's matrices directory.
Need a Windows (or Unix) platform. Open a run window, run
Using Cluto
http://www.octave.org

http://www-users.cs.umn.edu/~karypis/cluto/index.html

http://www.cluster.com


http://www.cs.umbc.edu/~richters/software.html

Related Software