Methods for Intelligent Systems

Lecture Notes on Clustering (I)
2009-2010

Davide Eynard

eynard@elet.polimi.it

Department of Electronics and Information

Politecnico di Milano

Some Course Info

- Lectures given by:
  - Davide Eynard (Teaching Assistant)
    http://www.dei.polimi.it/people/eynard
    eynard@elet.polimi.it
- Course Material on Clustering
  - These lecture notes
  - Papers and tutorials (check *Bibliography* at the end)
- Web Links
  - http://del.icio.us/clust2008 (anyone can read links, without the need to log in)
  - more recent links inside these slides

Course Schedule [*Tentative*]

Date	Topic
11/03/2010	Clustering: Introduction
18/03/2010	Clustering: K-means & Hierarchical
25/03/2010	Clustering: Fuzzy, Gaussian & SOM
08/04/2010	Clustering: PDDP & Vector Space Model
15/04/2010	Clustering: Limits, DBSCAN & Jarvis-Patrick
29/04/2010	Clustering: Evaluation Measures

Clustering: a definition

"The process of organizing objects into *groups* whose members are *similar in some way*"

J.A. Hartigan, 1975

"An algorithm by which objects are grouped in *classes*, so that intra-class similarity is maximized and inter-class similarity is minimized"

J. Han and M. Kamber, 2000

Clustering: a definition

- Clustering is an *unsupervised learning* algorithm
  - Remember? "**Exploit regularities** in the inputs to **build a representation** that can be used for reasoning or prediction"
- Particular attention to
  - *groups/classes* (vs *outliers*)
  - *distance/similarity*
- What makes a good clustering?
  - No (independent) best criterion
  - ***data reduction*** (find representatives for homogeneous groups)
  - ***natural data types*** (describe unknown properties of natural clusters)
  - ***useful data classes*** (find useful and suitable groupings)
  - ***outlier detection*** (find unusual data objects)

(Some) Applications of Clustering

- Market research
  - find groups of customers with similar behavior for targeted advertising
- Biology
  - classification of plants and animals given their features
- Insurance, telephone companies
  - group customers with similar behavior
  - identify frauds
- On the Web:
  - document classification
  - cluster Web log data to discover groups of similar access patterns
  - recommendation systems ("If you liked this, you might also like that")

Example: Clustering CDs

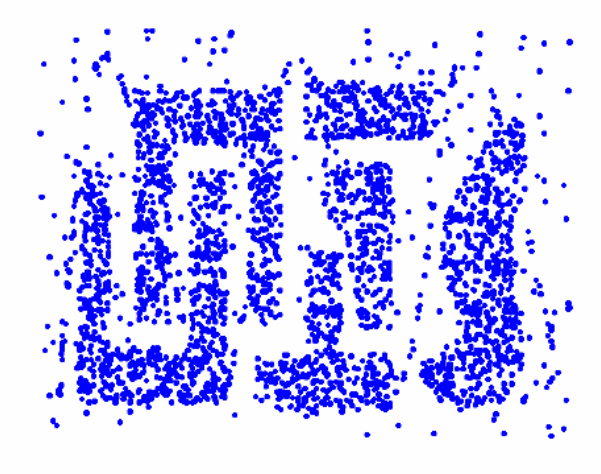
- Intuitively: music divides into categories, and customers prefer a few categories
  - But what are categories really?
- Represent a CD by the customers who bought it
- Similar CDs have similar sets of customers, and vice-versa
- Think of a space with one dimension for each customer
  - Values in a dimension may be 0 or 1 only
- A CD's point in the space is $(x_1, x_2, \ldots, x_k)$, where $x_i = 1 \; iff$ the i^{th} customer bought the CD
  - Compare with the "correlated items" matrix (rows=customers, columns=CDs)

Requirements

- Scalability
- Dealing with different types of attributes
- Discovering clusters with arbitrary shapes
- Minimal requirements for domain knowledge to determine input parameters
- Ability to deal with noise and outliers
- Insensitivity to the order of input records
- High dimensionality
- Interpretability and usability

Question

What if we had a dataset like this?



Problems

There are a number of problems with clustering. Among them:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of *distance* (for distance-based clustering);
- if an obvious distance measure doesn't exist we must define it, which is not always easy, especially in multi-dimensional spaces;
- the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

Clustering Algorithms Classification

- Exclusive vs Overlapping
- Hierarchical vs Flat
- Top-down vs Bottom-up
- Deterministic vs Probabilistic
- Data: symbols or numbers

Similarity through distance

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Simplest case: one numeric attribute A
  - $Distance(X, Y) = A(X) - A(Y)$
- Several numeric attributes
  - $Distance(X, Y)$ = Euclidean distance between X and Y
- Nominal attributes
  - Distance is set to 1 if values are different, 0 if they are equal
- Are all attributes equally important?
  - Weighting the attributes might be necessary

Distance Measures

Two major classes of distance measure:

- Euclidean
  - A Euclidean space has some number of real-valued dimensions and "dense" points
  - There is a notion of *average* of two points
  - A Euclidean distance is based on the locations of points in such a space
- Non-Euclidean
  - A Non-Euclidean distance is based on properties of points, but not on their *location* in a space

Distance Measures

Axioms of a Distance Measure:

- d is a *distance measure* if it is a function from pairs of points to reals such that:
  1. $d(x, y) \geq 0$
  2. $d(x, y) = 0$ iff $x = y$
  3. $d(x, y) = d(y, x)$
  4. $d(x, y) \leq d(x, z) + d(z, y)$ (triangle inequality)

Distances for numeric attributes

- **Minkowski distance**:

$$d_{ij} = \sqrt[q]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^q}$$

  - where $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ are two p-dimensional data objects, and q is a positive integer

- if $q = 1$, d is **Manhattan distance**:

$$d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$$

Distances for numeric attributes

- **Minkowski distance**:

$$d_{ij} = \sqrt[q]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^q}$$

  - where $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$ are two p-dimensional data objects, and q is a positive integer

- if $q = 2$, d is **Euclidean distance**:

$$d_{ij} = \sqrt[2]{\sum_{k=1}^{n} |x_{ik} - x_{jk}|^2}$$

For those of you who’re interested in a clustering project, two words on data retrieval...

Data Retrieval

- Many of the algos you’ll study will be ok for data clustering...
  - ... but you don’t have data to start with!
- Clustering is only a technique: how can you use it?
  - You first need to collect data
  - And to do this, you first need to understand how the Web works
- Some Web basics will help you in different ways
  - You’ll be able to extract data for your clustering algos
  - You’ll learn how to use the Web for your own advantage, find what you want more easily, have only interesting stuff delivered to you

... also, it’s quite funny ;)

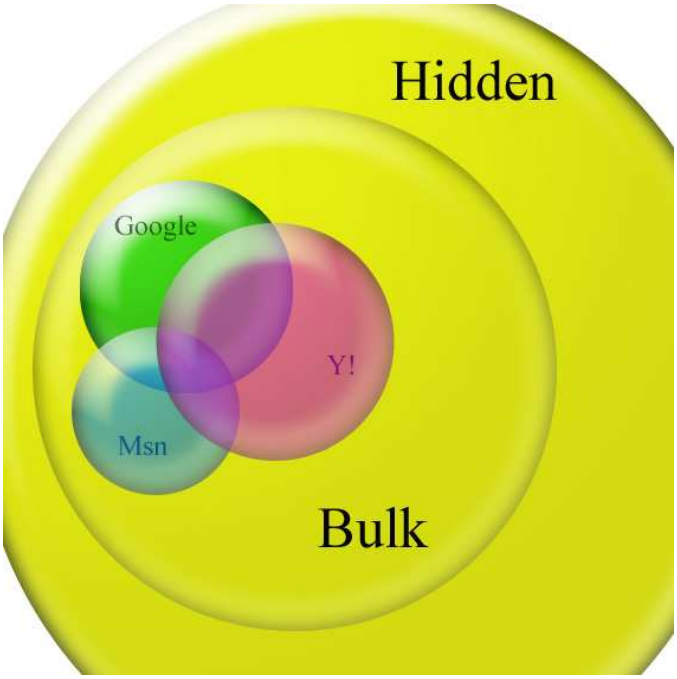
Why Text Clustering

Information on the Web increases every day:

- more than 25 billion indexed by google...
- ... and more than 25 billion are not indexed by any search engine!
- thousands of new pages every day
- thanks to blogs and forums, number increases more and more

The Structure of the Web

(courtesy of http://www.searchlores.org)



So, what?

We saw the Web is

- Not only the Web (irc, ftp, usenet, etc.)
- Not completely covered by search engines
- Growing very quickly

How can we actually *find* information on the Internet?

- Approach the problem from a "normal user" perspective
- Adapt it for a "search engine" point of view

How do you browse today?

Default browser (IE for Windoze) means you don’t have the chance to customize many things. As a result, you see only what others want you to see:

- images (often banners)
- popups
- tons of unuseful HTML code (which you don’t see, but you have to download anyway)

How do you browse today?

You see only what others want you to see, in the way they want you to see it:

- with "active", non accessible contents
- inside fixed-size windows
- following predefined paths (such as in Autogrills, every site has its own *noce di pepe*)

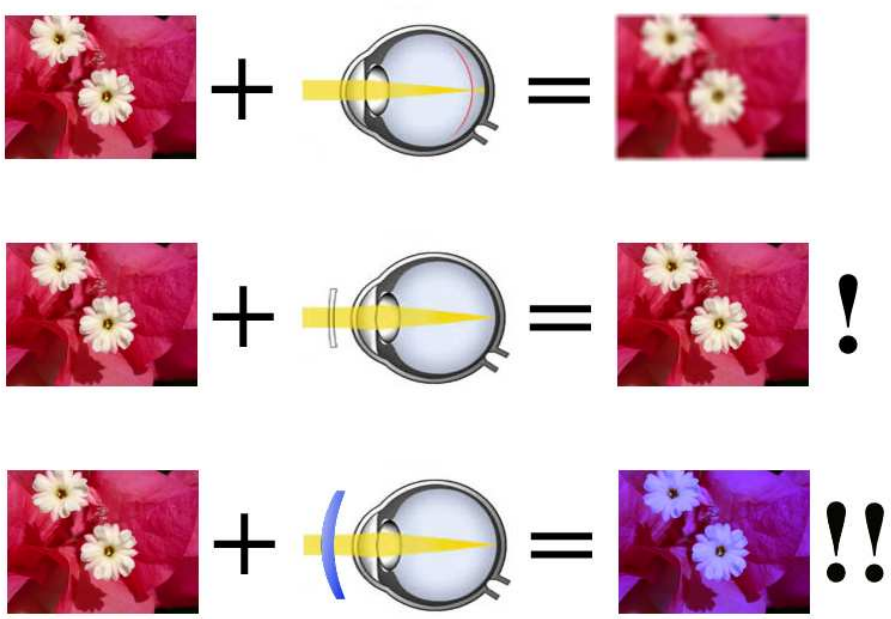
But well, this is what we're given, right?

How does it work, instead?

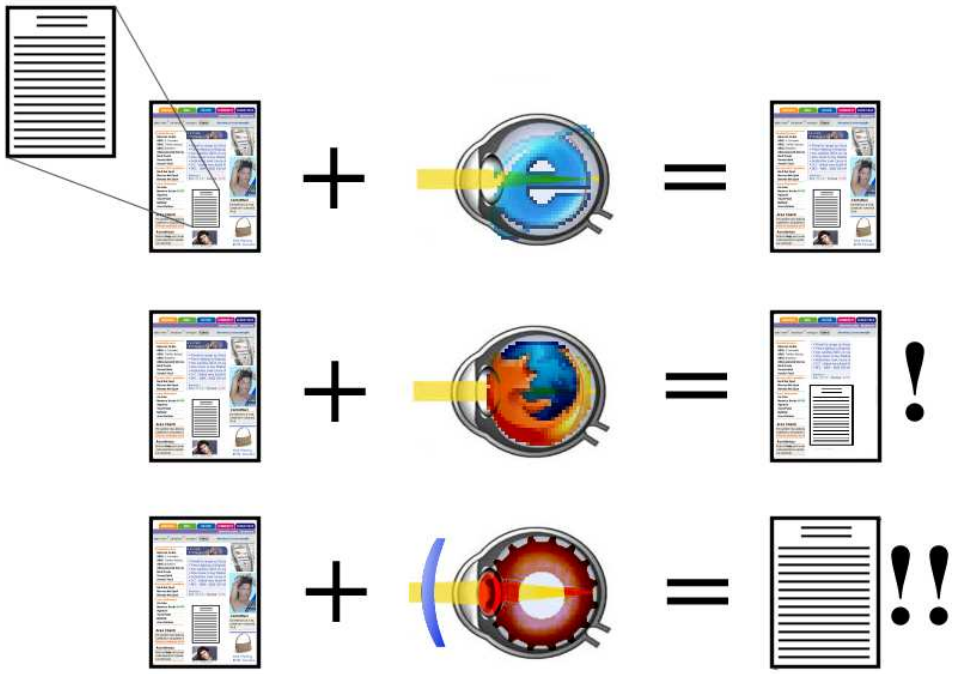
WRONG! A PC is not a TV, you can make it do whatever you want, such as:

- downloading only what you want
- showing Web pages the way you like
- collecting and analyzing data for you

Real Glasses



Power Glasses



Techniques and technologies

What are the techniques a user could use to get the best out of the Web?

- Some basic ones
  - alternative browsers
  - leechers and teleporters
  - spiders and scrapers
  - proxy-like softwares
- Some advanced ones
  - learn oneliners with curl, wget, lynx
  - learn how to search
  - learn how to extract data from Web pages
  - learn how to search inside extracted data

Learn how to search

Learning how to search, you'll also learn something more about how current Search Engines work, and get some ideas for your SE:

- Word-based searches (the "star" example)
- The "index of" trick (and how spammers exploited it, making things harder to find)
- Try different search engines (object-specific search)
- Try *clustering* search engines
  - What does *clustering* mean in this case?
- Try folksonomies
- Try blogs and forums

For more info about search strategies, give a look at this (warning: it contains LOTS of examples!)

Bot Basics

(or: learn how to extract data from Web pages)

- What is a bot?
- What should a bot do for us?
  1. visit a website, following links
  2. extract useful information
  3. work on data (or just save them to allow another app to use them)
- How can I create a bot?
  - tradeoff between performances and complexity
  - any programming language is fine (the faster the better)
  - check you have the libraries you need (http, text parsing etc.)

Recognizing Web Patterns

- Patterns in presentation/browsing
- Patterns within a website/a class of websites
- Tools: your brain ;-)

In both cases, automatically generated code helps much

Browsing with bots

Your bots will have to:

- download Web pages
- follow or collect links which satisfy particular conditions (on the tagged text or on the link itself), until a particular depth or forever
- fill forms (!)

Web Technologies

There are some things you should know to make a well-behaving bot:

- HTTP
  - GET and POST
  - Referer
  - UserAgent
  - Cookie
  - Proxy
- HTML
  - Form
  - Dynamically generated code

I suppose you already know at least the basics of these technologies. If you don’t, you can give a look at this tutorial.

A complete example

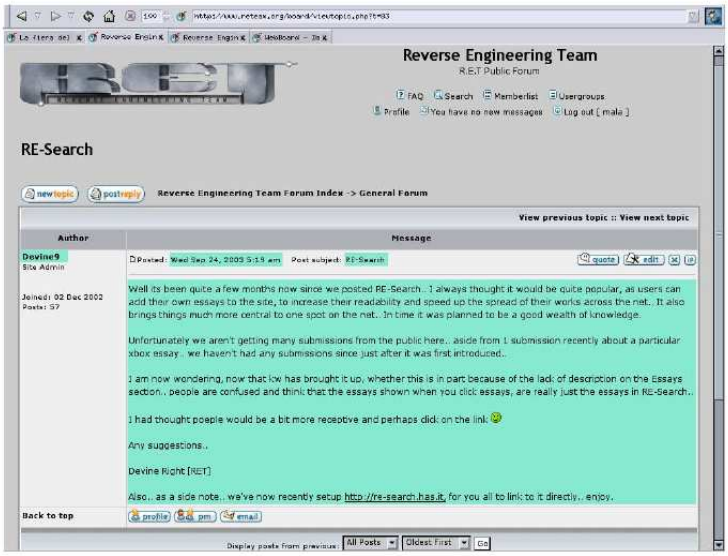
TWO (The Working Offline forum reader) is an old project of mine, which could help you understand how all these technologies work together.

- A component downloads pages from Web forums
- Another one extracts information from them
- Finally, data is normalized and saved inside the DB

Of course it's free (as in freedom): you can download it from http://two.sf.net

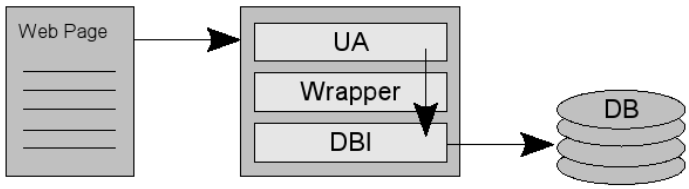
The idea

Figure 2.1: To view this page in your browser you have to download about 140KB. The interesting (highlighted) data are less than 1.4KB, that is a 1% ratio!



TWO structure

Figure 4.2: Input component's structure.



TWO performances

```
------------------------------
Test started at       22:49:00
Test finished at      00:28:00
------------------------------
Total test time       01:39:00
------------------------------



------------------------------
Downloaded pages          2245
Saved messages           13693
Bytes count           94967139
------------------------------



------------------------------
DB size before test (KB) 3288
DB size after test (KB) 11180
------------------------------
Total data size.(KB).... 7892
------------------------------



------------------------------
Forum data size (KB)     92741
TWO's data size (KB)      7892
------------------------------
Saved space.(KB).........84849
Saved space (perc).........91%
------------------------------
```

```
Pages count: 2244
Bytes count: 94943907
Messages count: 13692
Saving message #17986
----
GET http://board.anticrack.de/viewtopic.php?t=2491
User-Agent: Two/0.01
Cookie: phpbb2mysql_sid=d68ccd982732c219e55cff36ec5fa44f;
Cookie2: $Version="1"



Pages count: 2245
Bytes count: 94967139
Messages count: 13693
Saving message #17985
mala@kami:~/prj/last$
```

```
kami:/var/lib/mysql# du two
3288    two
kami:/var/lib/mysql# du two
11180   two
kami:/var/lib/mysql#
```

Bibliography

- "Metodologie per Sistemi Intelligenti" course - Clustering Tutorial Slides by P.L. Lanzi
- "Data mining" course - Clustering, Part I Tutorial slides by J.D. Ullman
- Satnam Alag: "Collective Intelligence in Action" (Manning, 2009)
- http://www.searchlores.org
- http://davide.eynard.it/malawiki/PowerBrowsing
- As usual, more info on del.icio.us