



CLUSTERING

SYRACUSE UNIVERSITY
School of Information Studies

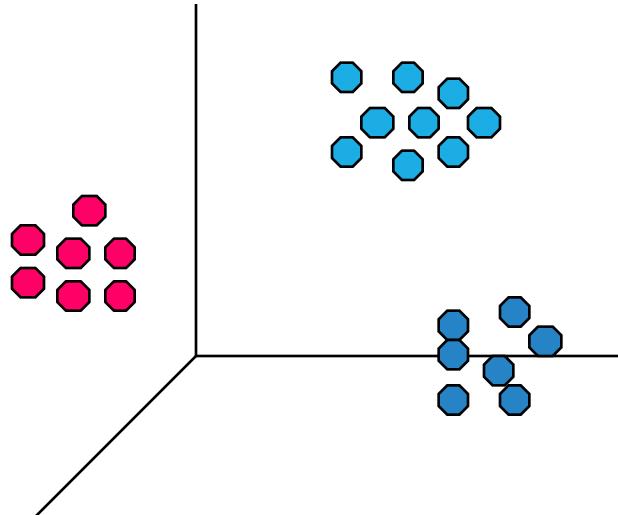
CLUSTERING

Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that:

Data points in one cluster are more similar to one another.

Data points in separate clusters are less similar to one another.

Intracluster distances are minimized.



Intercluster distances are maximized.

SIMILARITY MEASURES FOR CLUSTERING

Similarity measures:

Euclidean distance, if attributes are continuous

Other problem-specific measures

CLUSTERING APPLICATION 1

Market-customer segmentation

Goal: To find the “subgroups” among a large customer base

Approach:

Collect some “attributes” about the customers – their age, income, favorite brands, etc.

Calculate the “similarity” between the customers.

“Cluster” similar customers together.

WHAT DOES A CLUSTER MEAN?

Although a clustering algorithm can group or cluster similar customers together, it does not tell us what each cluster means.

Data analysts need to understand and interpret the meaning of clusters; e.g., one cluster may be interpreted as “tree huggers,” “money savers,” or “luxury fans.”

WHAT DOES A CLUSTER MEAN?

The screenshot shows a Google search results page for the query "alumni segmentation". The top navigation bar includes "All", "Images" (which is selected), "News", "Videos", "Shopping", "More", and "Search tools". On the right, there's a "View saved" button. Below the search bar, there are several search results:

- Three College Alumni Donor Segments**: A chart showing three segments: Champions (31%), Friends (36%), and Acquaintances (33%). Each segment has a brief description and a small profile photo.
- THE OMATIC SOLUTION FOR HIGHER EDUCATION ADVANCEMENT**: A diagram titled "INTEGRATED WITH THE RAISERS EDGE®" showing six stages: TARGET, PREDICT, IDENTIFY, ENGAGE, CAPTURE, and ANALYZE. Each stage has a corresponding icon and a brief description.
- Segment Your Alumni to Improve Engagement**: A slide with a world map and icons representing different segments. It includes a bar chart comparing "High vs. Average Income" and "High vs. Average Age".
- CASEVVI**: A slide featuring the text "BETTER TOGETHER" and "Developing Donor Segments: A Scientific and Social Approach to Effective Multi-Channel Communication Strategy". It includes a bar chart and a call to action: "Follow the Better Together Conference on Twitter #CASEV+VI".
- Trends Impacting Communication with Alumni/Donors**: A slide with a city skyline background and a bulleted list of trends: "Rise of the non-profits", "Non-funded written marketing plans", "Younger donors", "Female donors", and "Technology".
- Segmenting Alumni to Improve Engagement**: A slide with a book cover image and a "Get Your Free Copy" button. It includes a bulleted list: "Learn how to use data to effectively communicate with your alumni and help you plan better events.", "Average age = 45", "Average annual giving = \$76,052", "Working full-time = 61%", "Retired = 48%", and "Married = 53%".
- Who are Champions?**: A slide with a man sitting at a desk. It lists statistics: "32% Moved to their college in the last 12 months", "49% Worked for their college", "Total alum dollar = \$354", "Average size of donation = \$1,603", and "Total donations to all alumni in 2013 = \$1,603". It includes a quote from John about his alma mater.
- CONNECTED**, **ENGAGED**, **COMMITTED**: Three boxes describing levels of engagement:
 - CONNECTED**: "✓ Updated contact on website", "✓ Joined SM network(s)".
 - ENGAGED**: "✓ Contributes posts, tweets", "✓ attends an alumni event".
 - COMMITTED**: "✓ Makes donation", "✓ Refers and mentors students", "✓ Asks network to support school".

Three College Alumni Donor Segments

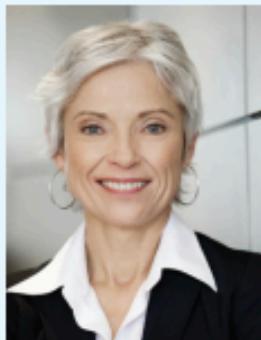


Champions

- Strongest advocates for the college.
- Value the professional and social benefits.
- Most likely to donate and the largest average donations.

Segment
Size

31%



Friends

- Proud graduates who regularly donate to the college.
- Much more committed to other philanthropies.
- Very satisfied with their lives.

36%



Acquaintances

- Had a passing relationship with their college.
- Minimal attachment as students and even less now.
- Provide little to no financial support.

33%

Who are Champions?



Average age = **45**

Average annual income = **\$76,052**

Working full-time = **61%**

Female = **48%**

Married = **53%**

32% Donated to their college in the last 12 months

49% Never donated to their college

\$1,769 Total alma mater donations since 2006

\$354 Average size of donation among donors

\$1,603 Total donations to all charities in 2010

John, what does your alma mater mean to you today?

I would not be who I am without my college's influence. I met many of my closest friends while a student and its numerous social opportunities remain an important part of my life. Professionally, I got my first job from a person who was a graduate of the college. The college continues to provide useful business contacts.

If you know me, you know I graduated from this college. Even if you don't know me, the logo on my jacket is a pretty good clue! When possible, I try to return for reunions and other important events. I take tremendous pride in the college's accomplishments and relish my association.

Supporting the college financially and by volunteering is a priority for me. Giving something back also feels good! I feel obligated to help the college because of all it has done for me. Its my duty!

Who are Friends?



Average age = **56**

Average annual income = **\$77,601**

Working full-time = **40%**

Female = **61%**

Married = **68%**

24% **Donated to their college in the last 12 months**

56% **Never donated to their college**

\$985 **Total alma mater donations since 2006**

\$197 **Average size of donation among donors**

\$2,750 **Total donations to all charities in 2010**

Susan, what does your alma mater mean to you today?

I am very proud of my college! Academically, it has always been a great school and I am fortunate to have attended. I am not the type of alumnus who wears college sweatshirts or puts decals on my car, but I certainly enjoy talking about the college when somebody asks. I rarely get back to campus, so the alumni magazine is a nice way to keep in touch.

I am very happy with my life and grateful to the college. I have no regrets for having attended my college. It is a great school. Nonetheless, I have not been involved with the college since my graduation. I am really not sure why, except my other interests take up all my time.

Yes, I regularly make modest donations to the college. It just seems like the right thing to do -- more of habit than a passion. Organizations providing food and health services are in greater need of my money and time.

Who are Acquaintances?



Average age = **51**

Average annual income = **\$69,935**

Working full-time = **49%**

Female = **59%**

Married = **54%**

5% Donated to their college in the last 12 months

86% Never donated to their college

\$226 Total alma mater donations since 2006

\$45 Average size of donation among donors

\$1,300 Total donations to all charities in 2010

Kate, what does your alma mater mean to you today?

Gosh, I really haven't thought much about my college since graduating 30 years ago. I wasn't a flag waving student and certainly have not become one as an alumnus! I didn't even attend commencement for my graduation. Honestly, I don't understand why people have strong feelings toward colleges. For me, college was a place where I earned my degree – no more, no less. I paid dearly for that degree, so why am I supposed to be grateful to them?

Yes, the college contacts me each year requesting a donation. I just say no and wait for their call next year when I say no again. I don't even read the alumni magazine they send. My annual refusal to give them money is my only contact with the college. Why do they keep calling? They should know by now that I am not going to give them anything. Calling me is a waste of their money and my time. I don't get it.

CLUSTERING: APPLICATION 2

Document clustering

Goal: To find groups of documents that are similar to each other based on the important terms appearing in them

Approach: To identify frequently occurring terms in each document, form a similarity measure based on the frequencies of different terms; use it to cluster.

Gain: Search engines can organize search results by document clusters.

SEARCH ENGINE BASED ON DOCUMENT CLUSTERING

<http://search.carrot2.org/stable/search>

Search “Amazon” and see the returned results organized into clusters with labels.

The pioneer clustering-based search engine Vivisimo was acquired by IBM in 2012.

CARROT2 SEARCH ENGINE BASED ON DOCUMENT CLUSTERING

The screenshot shows the Carrot2 search engine interface. At the top, there is a navigation bar with links for eTools Web Search, Wiki, Jobs, PubMed, and PUT. Below the navigation bar is a search bar containing the query "amazon". To the right of the search bar are buttons for "Search" and "More options". On the left side, there is a sidebar with tabs for Folders, Circles, and FoamTree. Under the Folders tab, there is a list of topics: All Topics (97), Amazon.com (18), Amazon Web Services (9), Amazon Video (8), Customers (8), Prime (7), Relatório de Atividades do Fundo Amazônia (7), Amazon River (5), App (5), Featuring (5), Cloud Computing (4), and a link to "more | show all". The main content area displays the search results for "amazon". The results are titled "Top 97 results of about 97 for amazon". The first result is a link to Amazon.com: Online Shopping for Electronics, Apparel, Comp... (Online retailer of books, movies, music and games along with e... https://www.amazon.com/ [Ask, Goo, Google]). The second result is a link to Gold Box Deals | Today's Deals - Amazon.com (Shop Amazon's Gold Box for our Deal of the Day, Lightning De... https://www.amazon.com/gp/goldbox [Ask, Bing, Yahoo]). The third result is a link to Amazon.com - Wikipedia (Amazon.com, Inc often referred to as simply Amazon, is an A... https://en.wikipedia.org/wiki/Amazon.com [Goo, Google, Wil...)). The fourth result is a link to Amazon.com: Kindle E-readers: Kindle Store (Welcome to the Kindle e-reader store featuring the best device... https://www.amazon.com/Amazon-Kindle-Ereader-Family/b?ie=)). Each result entry includes a small icon and three icons for sharing or viewing.

Top 97 results of about 97 for amazon

- [Amazon.com: Online Shopping for Electronics, Apparel, Comp...](#)
Online retailer of books, movies, music and games along with e...
<https://www.amazon.com/> [Ask, Goo, Google]
- [Gold Box Deals | Today's Deals - Amazon.com](#)
Shop Amazon's Gold Box for our Deal of the Day, Lightning De...
<https://www.amazon.com/gp/goldbox> [Ask, Bing, Yahoo]
- [Amazon.com - Wikipedia](#)
Amazon.com, Inc often referred to as simply Amazon, is an A...
<https://en.wikipedia.org/wiki/Amazon.com> [Goo, Google, Wil...)
- [Amazon.com: Kindle E-readers: Kindle Store](#)
Welcome to the Kindle e-reader store featuring the best device...
<https://www.amazon.com/Amazon-Kindle-Ereader-Family/b?ie=>

CLASSIFICATION VS. CLUSTERING

Classification: Supervised learning

Clustering: Unsupervised learning

No training data

No predefined target variable

More suitable for exploratory analysis for data sets that we don't know much about

CAN A CLUSTERING MODEL DO CLASSIFICATION?

Yes, sometimes serves as the first step, to define the categories

E.g., after a customer base is clustered into three clusters, we can examine these clusters and “label” them in “tree huggers,” “savers,” and “luxury fans” categories.

Given a new customer, we can predict which cluster this customer belongs to, based on his or her similarity with the members in each cluster.

CLUSTERING FOR CLASSIFICATION

Clustering points: 3,204 Articles of *Los Angeles Times*

Similarity measure: How many words are common in these documents (after some word filtering)

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278