UCIrvine | The Paul Merage School of Business

2022.11.24

BANA 273 Session 9

Prof. Vibs

The Paul Merage School of Business

University of California, Irvine

Agenda

- Term Project Presentations next week
- Upload presentation file to Canvas at least 1 hour before class

- Overview of

- Wiki for contributing final tions
  - https://docs.google.com/doc kkveDdileus5zJOg-LfU5siOZT8ObUR0GrsbF3iVE/edit?usp=sharing

Attribute Selection

- Weka – Correlation Based Feature (CFS) Selection
  - CfsSubset

- A good feat ontains features highly correlated with (pr) the class, yet uncorrelated with (not predictive of) each other.

- CFS is a fully automatic algorithm -- it does not require the user to specify any thresholds or the number of features to be selected, although both are simple to incorporate if desired

Other Methods

- Text Mining
- KNN
- Collaborativ
- Logistic Re
- Support Vector Machines
- Neural Nets
- Bagging
- Boosting

Why Text Mining?

- What can be discovered from text?
- Significant proportion of information of great potential value is stored in documents:
  - News stories pertaining to competition, customers & the business env
  - Technical re
  - Email communications with artners, and within the organization
  - Corporate documents embodying corporate knowledge and expertise
  - Legal documents --- automatic reasoning

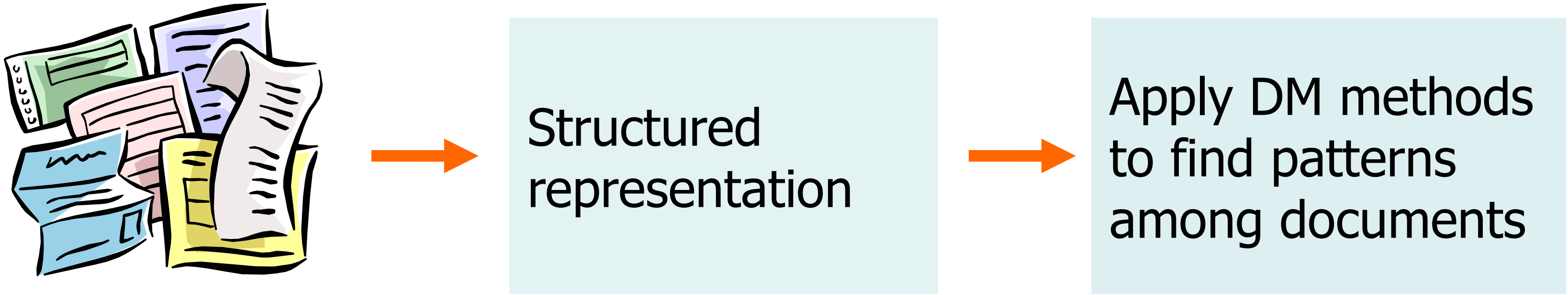
Opportunities

Finding patterns in text:

- Identify and track trends in industry - associations
  - What are my competitors doing?
  - What relevant products are being developed?
  - What are the cts?
- Identify emergi ocuments -cluster
  - Customer communications: cl s, each segment identifies a common theme su ints about a certain problem, or queries about product features.
- Automated categorization of e-mails (**Spam Filter!**), web pages, and news stories – classification

Structuring Textual Information

- Many methods designed to analyze structured data
- If documents can be represented by a set of attributes – can use existing data mining methods
- How to repre



Structured representation

Apply DM methods to find patterns among documents

Text Mining Concepts

- Document
- Token or term
- Corpus
- Bag of Words
- Stop word eli [illegible] wer case
- Term Frequency (TF)
- Inverse Document Frequency (IDF)
- TFIDF
- N-gram sequences
- Named entity extraction
- Topic models

Document Representation

- A document representation aims to capture what the document is about
- One possible approach:
  - □ Each row in the table represents a document
  - □ Attribute describes whether or not a term appears in the document

Example

	Camera	Digital	Memory	Pixel	…
Document 1	1	1	0	1	
Document 2	1	1	0	0	
…	…	…	…	…	

Document Representation using TF

- Term Frequency:
  - Attributes represent the frequency in which a term appears in the document
  - TF(t, d)

May impose upper cause the dimensionality i

	Camera	Digital		nt	…
Document 1	3	2	0	1	
Document 2	0	4	0	3	
…	…	…	…	…	

Inverse Document Frequency (IDF)

- But a term is mentioned more times in longer documents
- Therefore, use relative frequency (% of document): IDF(t) = 1 + l ocs containing t)

	Camera	Digital	Memory	Print	…
Document 1	3	2	1	2	
Document 2	1	1.4	1	3	
…	…	…	…	…	

Combining TF and IDF

- TFIDF(t, d) = TF(t, d) * IDF (t)
- Each row represents a document
- Each colum
- You can use tc. on this data

N-gram sequences

- “The quick brown fox jumps”
- 2-grams or bi-grams:
  - {quick, bro wn, brown-Fox, fox_jumps
  - You can see that the numb s can quickly get out of hand

Named entity extraction

- Example “Silicon Valley”, “LA Lakers”, “Merage School of Business”

Topic Models

Topic Models

Examining the Impact of Keyword Ambiguity on Search Advertising Performance: A Topic Model Approach, Gong, Abhishek and Li (MISQ 2018)

Text Mining Application 1: Association Rules

After proper representation, data mining techniques can be applied to text, e.g. association rules, clustering, classification.

Keyword-based Association Rules: treat keywords as items.

Microsoft → Antitrust

Document No.	Item 1	Item 2		
100	France	Iraq	US	
101	NASDAQ	NYSE	job	
102	Iraq	US	UK	
103	Microsoft	antitrust	OS	
104	Microsoft	Antitrust	windows	
…				

OR

	icrosoft	antitrust	France	…
		0	1	
		0	0	
102	0	0	0	
103	1	1	0	
104	1	1	0	
…				

Personalized Web Ad Delivery

- Objective:
  - Improve effectiveness of Web ads
  - Customize ad delivery so that ad corresponds to the context user is exploring
- Web content is dynamic → need automated ad placement

  - Example: Gmai

- Solution:
  - Represent each ad as a document w words.
  - For example: ad for hybrid car is re he following set of keyword: car, electric, environment, etc.
  - Then deliver ads to viewers of pages (i.e., documents) that resemble this description.

Link Structure Analysis to rank Web pages

- Traditional Information Retrieval methods only examine the appearance of relevant terms, and often fail to account for
  - The quality of the information in the retrieved documents.
  - The reliabili

- From the retrieved docume o rank authoritative documents higher

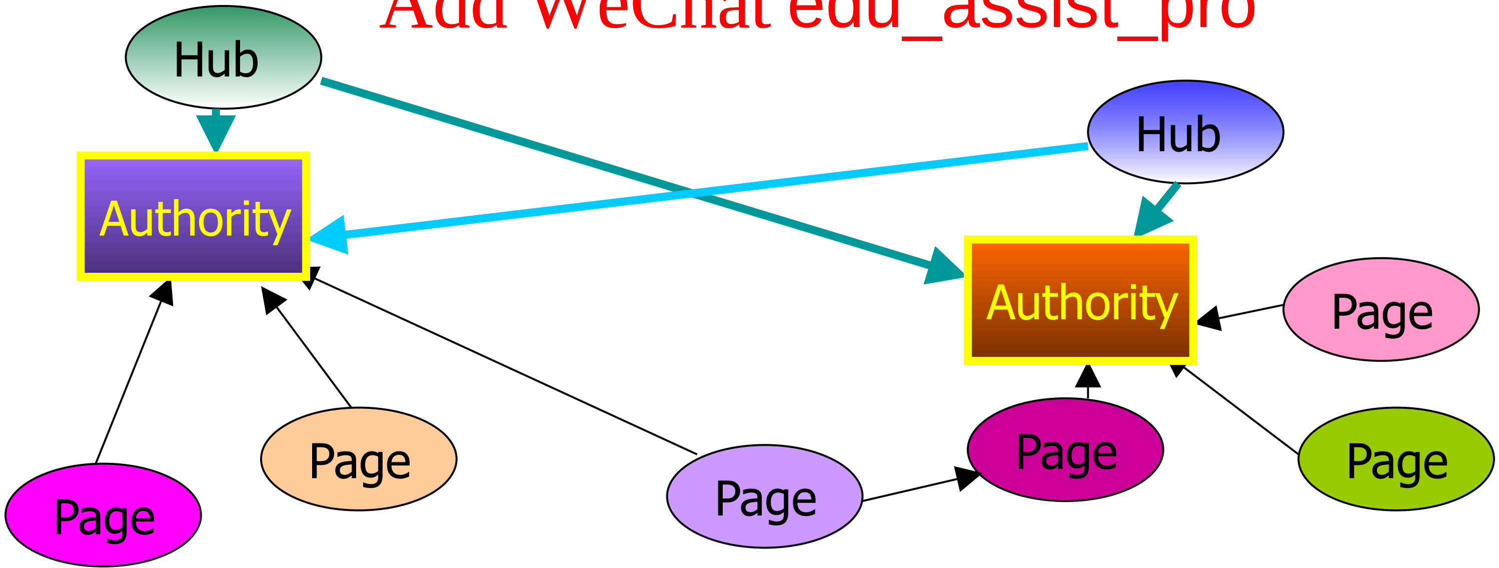
- Approach: Mining the Web’s link structure to identify authoritative web pages

Identify Authoritative Web Pages

- The Web includes pages and hyperlinks
- A lot of information is in the structure of web page linkages. Hyperlinks contain rich latent human information
  - An author cr another page -- can be viewed as endorseme
  - The collective endorsement o ge by different authors can help discover authoritative pages
- Google uses link structure of the Web to rank documents (PageRank)

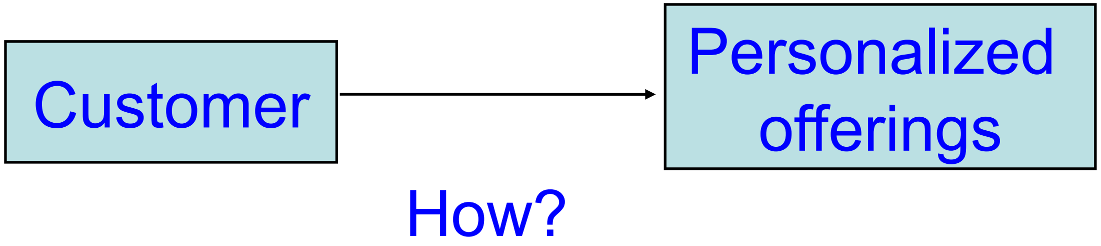
Using Hubs to identify Authoritative Web Pages

- A hub is a page pointing to many good authorities.
  - E.g., a web page pointing to many good sources of information on business intelligence
- A hub may not be an authority, and have very few links pointing to it.
  - Yet a link fro than a link from a regular page
- An authority is good hubs



Personalization

Personalization/customization tailors certain offerings by providers to consumers based on knowledge about them with cert



Classifier: Logistic Regression

- This is not a regression
- Uses logistic function and hinge loss function

K Nearest Neighbor (KNN)

K-Nearest Neighbor can be used for classification/prediction tasks.

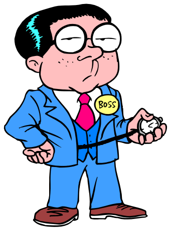
Step 1: Using a chosen distance metric, compute the distance between the new example and all past examples.

Step 2: Choose the *k* past examples that are closest to the new example.

Step 3: Work out the earest neighbors - the predominant class is your predict w example. i.e. classification is done by *majority vot* rest neighbors. For prediction problem with numeric target variable, the (weighted) average of the k nearest neighbors is used as the predicted target value.

How do we determine our neighbors?

- Each example is represented with a set of numerical attributes



John:
Age=35
Income=95K
No

Rachel:
Age=41
Income=215K
No. of credit cards=2

- “Closeness” is defined in ter *clidean* distance between two examp
  - □ The Euclidean distance between X=(x_1, x_2, x_3,…x_n) and Y =(y_1,y_2, y_3,…y_n) is defined as:

$$D(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

K-Nearest Neighbor Classifier

Example : 3-Nearest Neighbors

Customer	Age	Income	No. credit cards	Response
John	35	35K	3	No
Rachel	22			Yes
Hannah	63			No
Tom	59	170K	1	No
Nellie	25	40K	4	Yes
David	37	50K	2	?

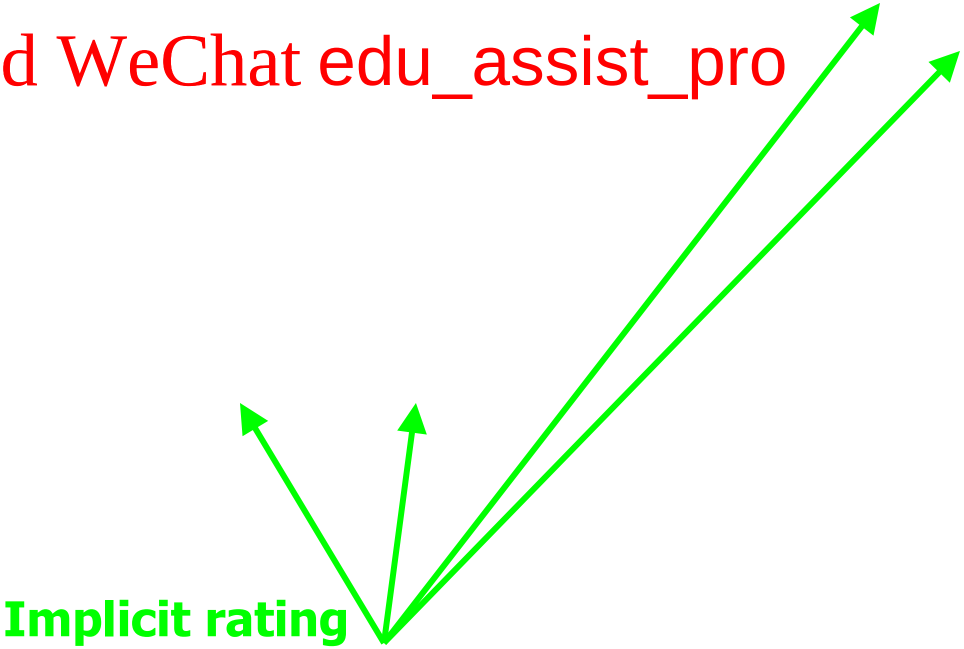
Collaborative Filtering:
Finding like-minded people

- One seeks recommendations about movies, restaurants, books etc. from people with similar tastes
- Automate t -mouth" by which people recommen or services to one another.

Collaborative Filtering

- Starts with a history of people’s personal preferences
- Uses a **distance function** – people who like the same things are “close”
- Determine a n k closest data points). We w dations from this neighborhood only.
  - Typically k is between 20 and 50
- Uses **“votes”** which are weighted by distances, so close neighbor votes count more

Example: amazon.com



Implicit rating

Artificial Neural Networks

- An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. -- Wikipedia

- A neural netw … onnected group of artificial neurons, and it proc … mation using a connectionist approach to co … -- Wikipedia

- Neural Nets learn complex functions Y=f(X) from data.

- ANN can approximate any function (e.g. logistic regression, linear regression).

Components of Neural Nets

- Neural Nets are composed of
  - Nodes, and
  - Arcs
- Each arc specifies a weight.
- Each node (other than the input nodes) contains a Transfer Function which converts its inputs to outputs. The input to a node is the weighted sum of

Recommender Systems

- Collaborative Filtering

- Content Bas
  - Use docum cription (tags)
  - Create user profile with we ferent tags
  - Example Books: Genre, Author, Length, Pictures etc.

- Knowledge Based Recommendation
  - When we do not have history of purchases (Camera)
  - Examine customer needs and match to product features

Bagging

- Combining predictions by voting/averaging
  - Each model receives equal weight
- "Idealized" version:

  - Sample sev
    (instead of

t of size n)

  - Build a classifier for each t

  - Combine the classifiers' predictions

Bagging classifiers

Model generation

```
Let n be the number of instances in the training data
For each of t iterations:
      Sample n instances from training set
            (with replacement)
      Apply learning algorithm to the sample
      Store r
```

Classification

```
For each of the t models:
      Predict class of instance using model
Return class that is predicted most often
```

Boosting

- Also uses voting/averaging
- Weights models according to performance
- Several vari
  - Read text f

Link Analysis is used for …

A: Identifying similar consumers for product recommendations

B: Highly non-

C: Replicating

D: Determining which web s uments are more authoritative and credible.

E: None of the above

Next Session

- Project Presentations
  - All Students must attend
  - Please uplo on Canvas at least 1 hou