UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

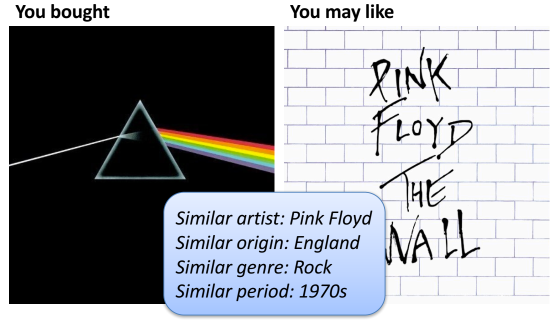
Recommender Systems

Topic Modeling

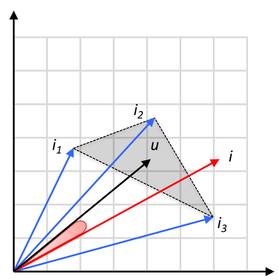
Rodrygo Santos
rodrygo@dcc.ufmg.br

Content-based recommendation

You bought **You may like**



Vector space representation



- Each item is a vector
  - One component for each term in the vocabulary
- Each user is a vector
  - Some combination of item vectors
- Prediction by similarity
  - Cosine of the angle between the user and item vectors

The curse of dimensionality

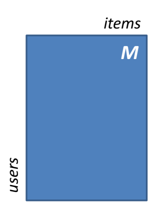
- The space of terms is very ***high-dimensional***!
- ***Problems***
  - **Efficiency**
    It will take longer to compute similarities
  - **Effectiveness**
    It will be harder to match similar ***concepts***

- Google Web N-grams [Franz and Brants, 2006]

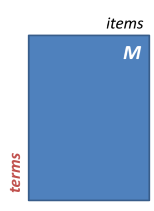
# tokens	1,024,908,267,229
# sentences	95,119,665,584
# 1-grams	**13,588,391**
# 2-grams	314,843,401
# 3-grams	977,069,902
# 4-grams	1,313,818,354
# 5-grams	1,176,470,663

The curse of dimensionality

Collaborative filtering

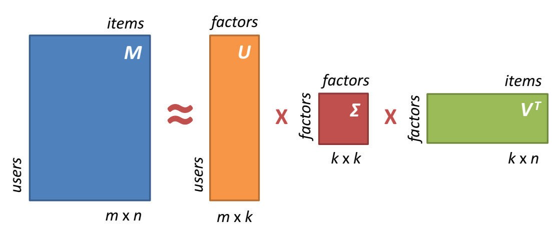


Content-based filtering



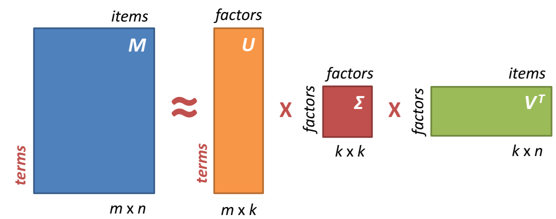
Latent semantic analysis

Collaborative filtering



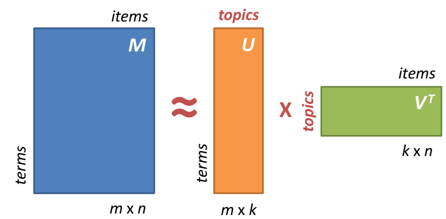
Latent semantic analysis

Content-based filtering



Latent topic modeling

Content-based filtering



Dimensionality reduction

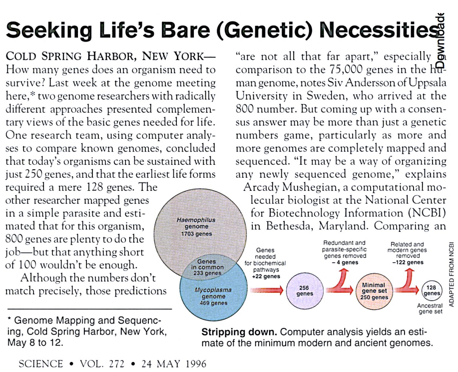
- TF-IDF [Luhn, IBM J. R&D 1957; Sparck-Jones, J. Doc. 1972; Salton and Buckley, IP&M 1988]
- LSI [Deerwester et al., ASIS 1988]
- pLSI [Hofmann, SIGIR 1999]
- LDA [Blei et al., JMLR 2003]

Latent Dirichlet allocation (LDA)

"Imagine searching and exploring documents based on the themes that run through them. Rather than finding documents through keyword search alone, we might first find the theme that we are interested in, and then examine the documents related to that theme."

[Blei, CACM 2012]

An example



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Latent topics

human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations
"genetics"	**"evolution"**	**"disease"**	**"computers"**

Generative modeling

- LDA is a generative model
  - It models the process of generating words
- Say you want a document with *n* words
  - Assume there are *k* known topics
  - Choose the document's distribution over topics
  - For each of the *n* words to be generated
    - Choose a topic from the document's topic distribution
    - Choose a word from the chosen topic

Generative modeling

- Generating *n* words
  - Choose the document's topic distribution
  - For each of the *n* words to be generated
    - Choose a topic from the document's distribution
    - Choose a word from the chosen document topic

Generative modeling

- Generating *n* words
  - **Choose the document's topic distribution**
  - For each of the *n* words to be generated
    - Choose a topic from the document's distribution
    - Choose a word from the chosen document topic

50% genetics
30% evolution
15% disease
5% computers

Generative modeling

- Generating *n* words
  - Choose the document's topic distribution
  - For each of the *n* words to be generated
    - **Choose a topic from the document's distribution**
    - Choose a word from the chosen document topic

50% genetics
30% evolution
15% disease
5% computers

"evolution"
evolution
evolutionary
species
organisms
life
origin

Generative modeling

- Generating *n* words
  - Choose the document's topic distribution
  - For each of the *n* words to be generated
    - Choose a topic from the document's distribution
    - **Choose a word from the chosen document topic**

50% genetics
30% evolution
15% disease
5% computers

"evolution"
evolution
evolutionary
species
organisms
life
origin

origin

Generative modeling

- Generating *n* words
  - Choose the document's topic distribution
  - For each of the *n* words to be generated
    - **Choose a topic from the document's distribution**
    - Choose a word from the chosen document topic

50% genetics
30% evolution
15% disease
5% computers

"genetics"
human
genome
dna
genetic
genes
sequence

origin

Generative modeling

- Generating *n* words
  - Choose the document's topic distribution
  - For each of the *n* words to be generated
    - Choose a topic from the document's distribution
    - **Choose a word from the chosen document topic**

50% genetics
30% evolution
15% disease
5% computers

"genetics"

human
genome
dna
genetic
genes
sequence

origin human

Generative modeling

- Generating *n* words
  - Choose the document's topic distribution
  - For each of the *n* words to be generated
    - **Choose a topic from the document's distribution**
    - Choose a word from the chosen document topic

50% genetics
30% evolution
15% disease
5% computers

"computers"

computer
models
information
data
computers
system

origin human

Generative modeling

- Generating *n* words
  - Choose the document's topic distribution
  - For each of the *n* words to be generated
    - Choose a topic from the document's distribution
    - **Choose a word from the chosen document topic**

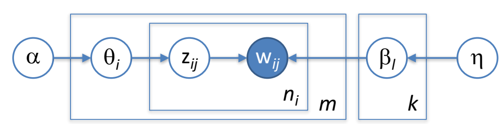
50% genetics
30% evolution
15% disease
5% computers

"computers"

computer
models
information
data
computers
system

origin human models

In plate notation

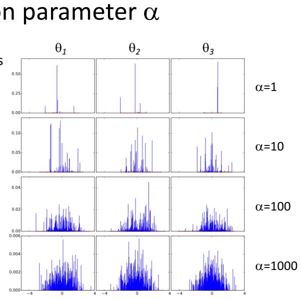


- θ_i: topic distribution of document *i* (of *m* documents)
  - α: parameter of the Dirichlet prior
- β_l: word distribution of topic *l* (of *k* topics)
  - η: parameter of the Dirichlet prior
- w_{ij}: *j*-th word in document *i*
- z_{ij}: chosen topic of word w_{ij}

Dirichlet distribution

- A "distribution of distributions" with concentration parameter α

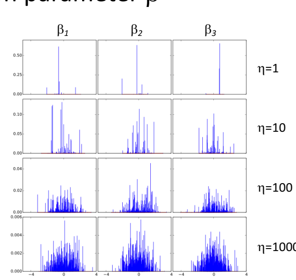
Document models as topic distros



Dirichlet distribution

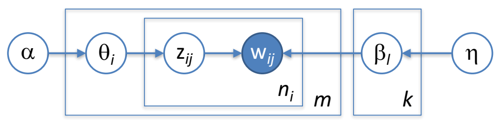
- A "distribution of distributions" with concentration parameter β

Topic models as word distros



In plate notation

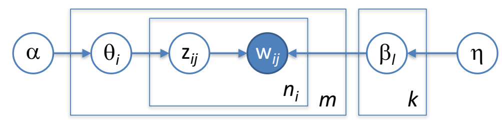
UFMG COMPUTER SCIENCE



- Choose $\theta_i \sim \mathrm{Dir}(\alpha)$ for $i \in \{1, \ldots, m\}$
- Choose $\beta_l \sim \mathrm{Dir}(\eta)$ for $l \in \{1, \ldots, k\}$
- For each document $i \in \{1, \ldots, m\}$
  - For each position $j \in \{1, \ldots, n_i\}$
    - Choose a topic $z_{ij} \sim \mathrm{Mult}(\theta_i)$
    - Choose a word $w_{ij} \sim \mathrm{Mult}(\beta_{z_{ij}})$

Mathematically

UFMG COMPUTER SCIENCE

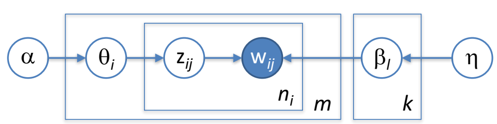


- Equivalent to the following joint distribution

$$p(\beta_{1:k}, \theta_{1:m}, z_{1:m}, w_{1:m})$$
$$= \prod_{l=1}^{k} p(\beta_l \mid \eta) \prod_{i=1}^{m} p(\theta_i \mid \alpha) \left(\prod_{j=1}^{n_i} p(z_{ij} \mid \theta_i) p(w_{ij} \mid \beta_{1:k}, z_{ij}) \right)$$

Reversing the logic

UFMG COMPUTER SCIENCE



- In reality, we don't know the topics
  - Or, equivalently, the θ_i and β_l distributions
- We actually know the documents
  - ***How to uncover the hidden topic structure?***

Posterior inference

UFMG COMPUTER SCIENCE

- How to compute the distribution of the topic structure given the observed documents?
  - Aka the posterior distribution

$$p(\beta_{1:k}, \theta_{1:m}, z_{1:m} \mid w_{1:m}) = \frac{p(\beta_{1:k}, \theta_{1:m}, z_{1:m}, w_{1:m})}{p(w_{1:m})}$$

- Problem: computing the marginal $p(w_{1:m})$
  - ***Intractable:*** would require examining every possible instantiation of the hidden variables

Approximate inference

UFMG COMPUTER SCIENCE

- Sampling algorithms
  - Iteratively collect samples from the posterior
    - e.g., Gibbs sampling
- Variational algorithms
  - Posit a parameterized family of distributions over the hidden structure, search for the best one
    - Easily handle millions of documents
    - Can accommodate streaming textual collections

More on this?

UFMG COMPUTER SCIENCE

- Related courses
  - Probabilistic graphical models
  - Bayesian inference

How to leverage topics?

UFMG COMPUTER SCIENCE

- Vector space model
  - $p(i|u) = \cos(\theta_u, \theta_i)$
- Item likelihood model
  - $p(i|u) = \prod_{w \in i} p(w|\theta_u)^{\mathrm{tf}_{wu}}$
- Unified likelihood model
  - $p(i|u) = -KL(\theta_u || \theta_i)$

$$= -p(w|\theta_u) \log \frac{p(w|\theta_u)}{p(w|\theta_i)}$$

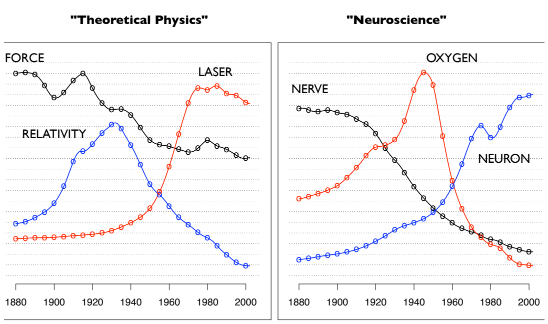
LDA variants

UFMG COMPUTER SCIENCE

- Syntactic topic model
  - A word or its topic is influenced by syntax
- Correlated topic model, hierarchical topic model
  - Some topics resemble other topics
- Polylingual topic model
  - Different languages, same topic mixtures
- Relational topic model
  - Exploiting link structure
- Dynamic topic model
  - Topics are time-dependent

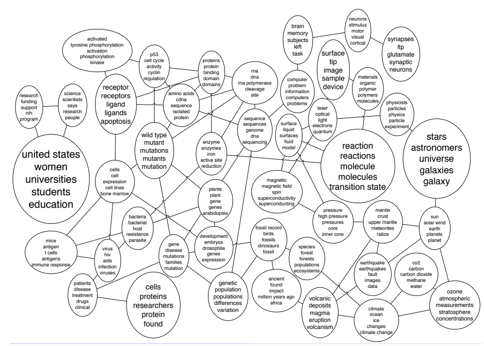
Modeling evolution

UFMG COMPUTER SCIENCE



Modeling correlations

UFMG COMPUTER SCIENCE



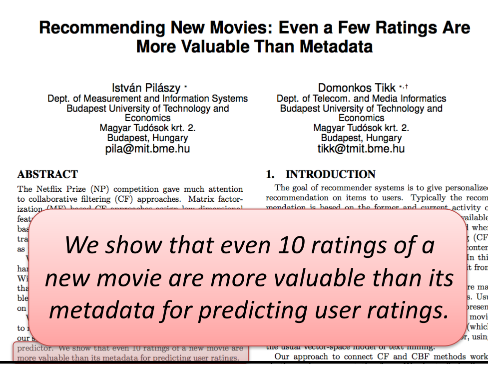
Summary

UFMG COMPUTER SCIENCE

- Content-based recommender systems are effective in many difficult scenarios
  - Cold-start items, basket analysis
- Build upon a history of research in IR
  - How to represent users and items
  - How to match users and items
- Still an active research area
  - How to go beyond a raw content representation?

A word of caution

UFMG COMPUTER SCIENCE


Recommending New Movies: Even a Few Ratings Are More Valuable Than Metadata

István Pilászy
Dept. of Measurement and Information Systems
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
pila@mit.bme.hu

Domonkos Tikk
Dept. of Telecom. and Media Informatics
Budapest University of Technology and Economics
Magyar Tudósok krt. 2.
Budapest, Hungary
tikk@tmit.bme.hu

ABSTRACT

The Netflix Prize (NP) competition gave much attention to collaborative filtering (CF) approaches. Matrix factorization ...

1. INTRODUCTION

The goal of recommender systems is to give personalized recommendation on items to users. Typically the recommendation is based on the former and current activity of ...

We show that even 10 ratings of a new movie are more valuable than its metadata for predicting user ratings.