Information Retrieval

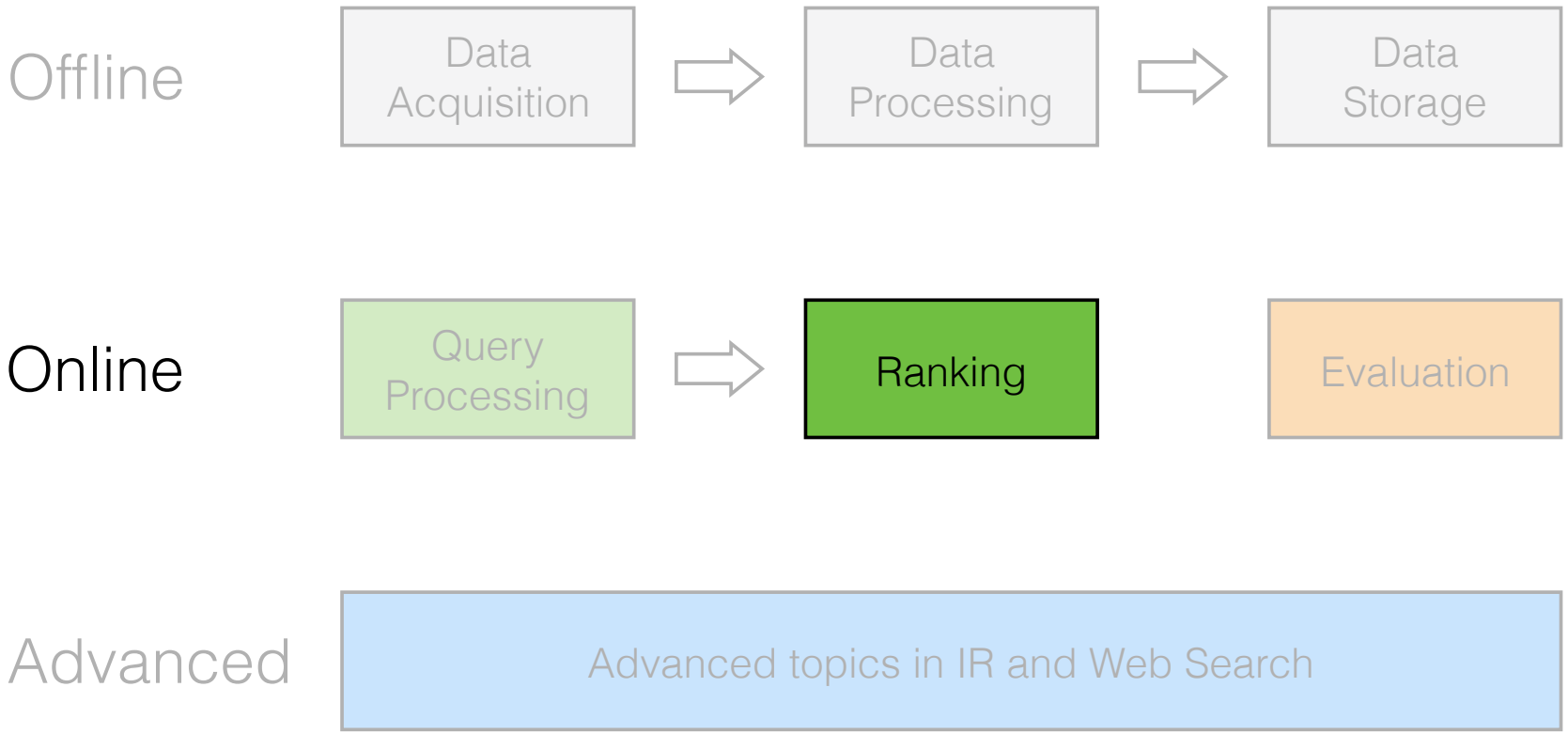
Semantic-based Retrieval

Ilya Markov
i.markov@uva.nl

University of Amsterdam

Outline

Ranking methods



Ranking methods

1. Content-based
   - Term-based
   - **Semantic**
2. Link-based (web search)
3. Learning to rank

Problems with term-based retrieval

1. Synonymy
2. Polysemy

Outline

Vector space model

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Anthony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Manning et al., "Introduction to Information Retrieval"

Singular value decomposition

- C is a $m \times n$ matrix (term-document)
- C can be decomposed as

$$C = U\Sigma V^T$$

- U is a $m \times m$ unitary matrix
- Σ is a diagonal $m \times n$ matrix with singular values
- V^T is a $n \times n$ unitary matrix

SVD example: original matrix

	d_1	d_2	d_3	d_4	d_5	d_6
ship	1	0	1	0	0	0
boat	0	1	0	0	0	0
ocean	1	1	0	0	0	0
voyage	1	0	0	1	1	0
trip	0	0	0	1	0	1

Manning et al., "Introduction to Information Retrieval"

SVD example: decomposition

	1	2	3	4	5
ship	−0.44	−0.30	0.57	0.58	0.25
boat	−0.13	−0.33	−0.59	0.00	0.73
ocean	−0.48	−0.51	−0.37	0.00	−0.61
voyage	−0.70	0.35	0.15	−0.58	0.16
trip	−0.26	0.65	−0.41	0.58	−0.09

$\times$

2.16	0.00	0.00	0.00	0.00
0.00	1.59	0.00	0.00	0.00
0.00	0.00	1.28	0.00	0.00
0.00	0.00	0.00	1.00	0.00
0.00	0.00	0.00	0.00	0.39

$\times$

	d_1	d_2	d_3	d_4	d_5	d_6
1	−0.75	−0.28	−0.20	−0.45	−0.33	−0.12
2	−0.29	−0.53	−0.19	0.63	0.22	0.41
3	0.28	−0.75	0.45	−0.20	0.12	−0.33
4	0.00	0.00	0.58	0.00	−0.58	0.58
5	−0.53	0.29	0.63	0.19	0.41	−0.22

Manning et al., "Introduction to Information Retrieval"

Low-rank approximation

$$C = U\Sigma V^T = \sum_{i=1}^{\min(m,n)} \sigma_i \vec{u}_i \vec{v}_i^T$$

$$\approx \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T = U_k \Sigma_k V_k^T$$

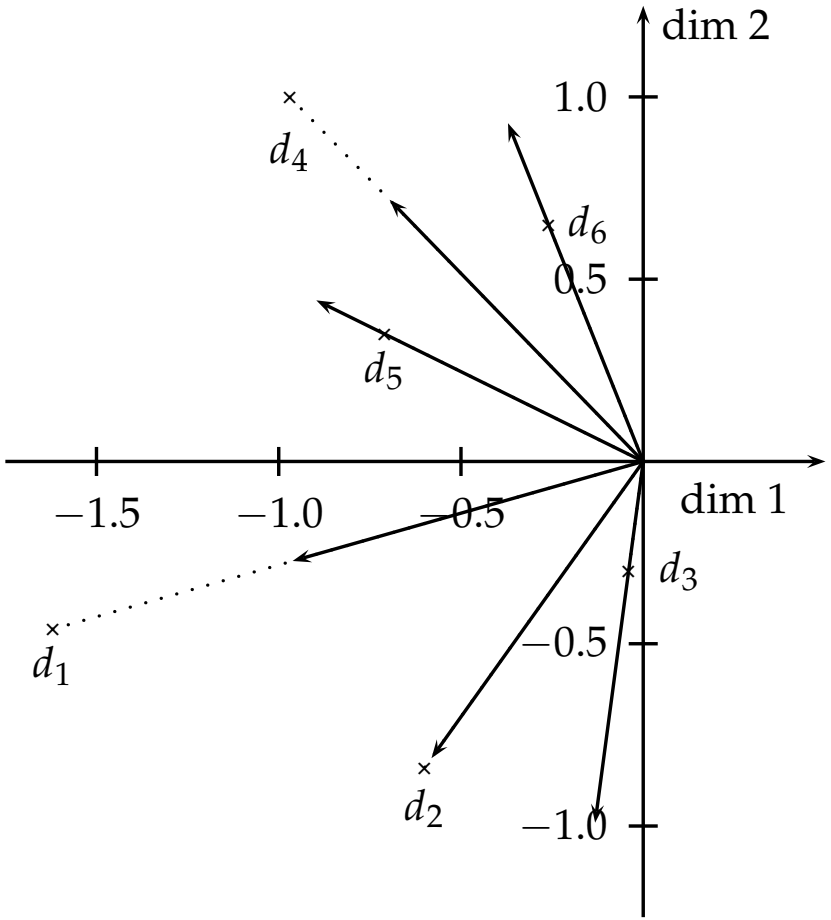
LSI/LSA example: low-rank approximation

$$\begin{bmatrix} 2.16 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.59 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \end{bmatrix} \times$$

	d_1	d_2	d_3	d_4	d_5	d_6
1	-1.62	-0.60	-0.44	-0.97	-0.70	-0.26
2	-0.46	-0.84	-0.30	1.00	0.35	0.65

Manning et al., "Introduction to Information Retrieval"

LSI/LSA example: vector space



Manning et al., "Introduction to Information Retrieval"

Latent semantic indexing/analysis

$$
\begin{array}{ccccccc}
 & C & & U_k & & \Sigma_k & V_k^T \\
 & (\mathbf{d}_j) & & & & & (\hat{\mathbf{d}}_j) \\
 & \downarrow & & & & & \downarrow \\
(\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,n} \end{bmatrix} & = (\hat{\mathbf{t}}_i^T) \rightarrow & \begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} & \dots & \begin{bmatrix} \\ \mathbf{u}_k \\ \\ \end{bmatrix} \end{bmatrix} & \cdot \begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_k \end{bmatrix} & \cdot \begin{bmatrix} \begin{bmatrix} & \mathbf{v}_1 & \end{bmatrix} \\ \vdots \\ \begin{bmatrix} & \mathbf{v}_k & \end{bmatrix} \end{bmatrix}
\end{array}
$$

$$d_j = U_k \Sigma_k \hat{d}_j \Longrightarrow \hat{d}_j = \Sigma_k^{-1} U_k^T d_j$$

https://en.wikipedia.org/wiki/Latent_semantic_analysis

Outline

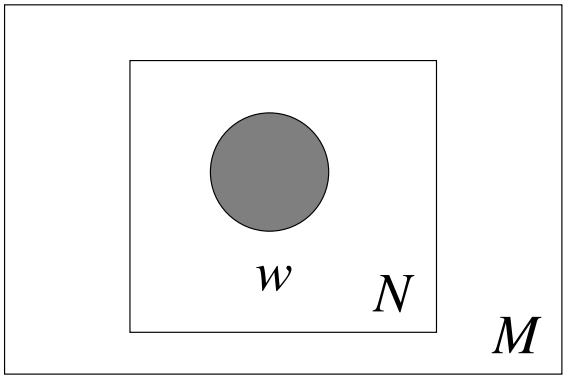
Unigram language model

Model M_1		Model M_2	
the	0.2	the	0.15
a	0.1	a	0.12
frog	0.01	frog	0.0002
toad	0.01	toad	0.0001
said	0.03	said	0.03
likes	0.02	likes	0.04
that	0.04	that	0.04
dog	0.005	dog	0.01
cat	0.003	cat	0.015
monkey	0.001	monkey	0.002
...	...	...	...

$$P(t) = P(t \mid M)$$

Manning et al., "Introduction to Information Retrieval"

Unigram language model



Blei et al., "Latent Dirichlet Allocation"

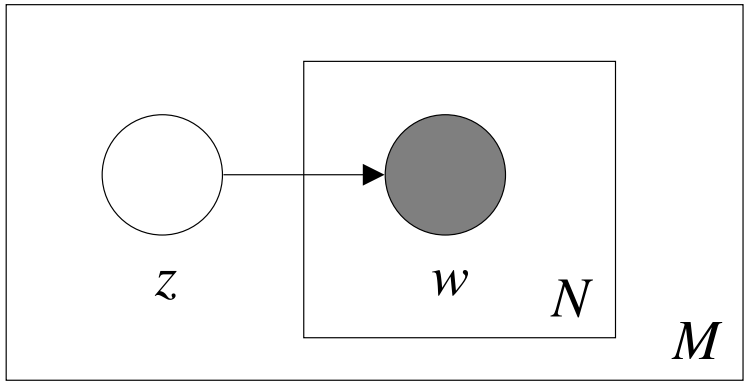
Mixture of unigrams

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

$$P(t) = \sum_z P(t \mid z)P(z)$$

Blei et al., "Latent Dirichlet Allocation"

Mixture of unigrams



Blei et al., "Latent Dirichlet Allocation"

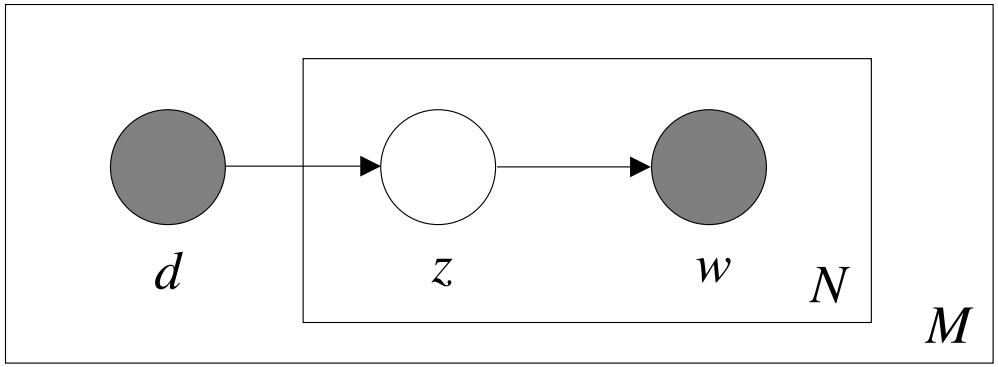
Probabilistic LSA

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

$$P(t \mid d) = \sum_z P(t \mid z) P(z \mid d)$$

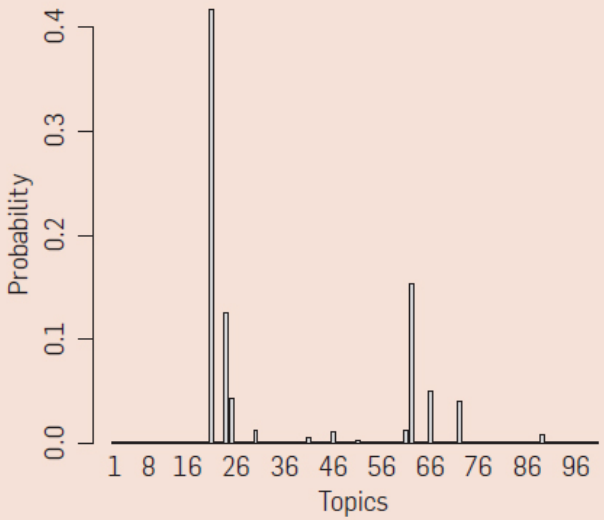
Blei et al., "Latent Dirichlet Allocation"

Probabilistic LSA



Blei et al., "Latent Dirichlet Allocation"

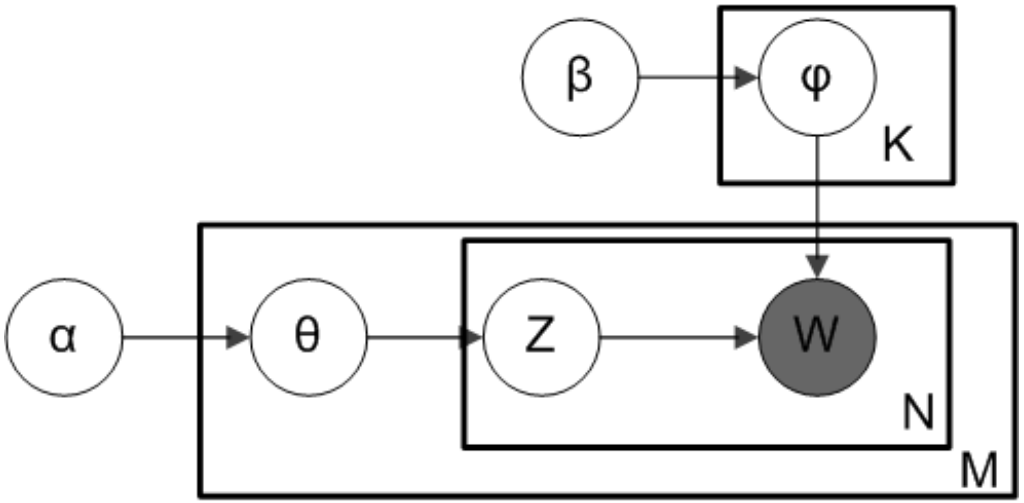
Latent Dirichlet allocation



"Genetics"	"Evolution"	"Disease"	"Computers"
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

Blei, "Probabilistic topic models"

Latent Dirichlet allocation



$$P(t \mid d) = \sum_{z=1}^{K} P(t \mid z, \phi) P(z \mid d, \theta)$$

https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Outline

IR conferences

- ACM Conference on Research and Development in Information Retrieval (SIGIR)
- ACM Conference on Information Knowledge and Management (CIKM)
- ACM Conference on Web Search and Data Mining (WSDM)
- European Conference on Information Retrieval (ECIR)
- ACM International Conference on Theory of Information Retrieval (ICTIR)

IR journals

- ACM Transactions o Information Systems (TOIS)
- Information Retrieval Journal (IRJ)
- Information Processing and Management (IPM)

Surveys

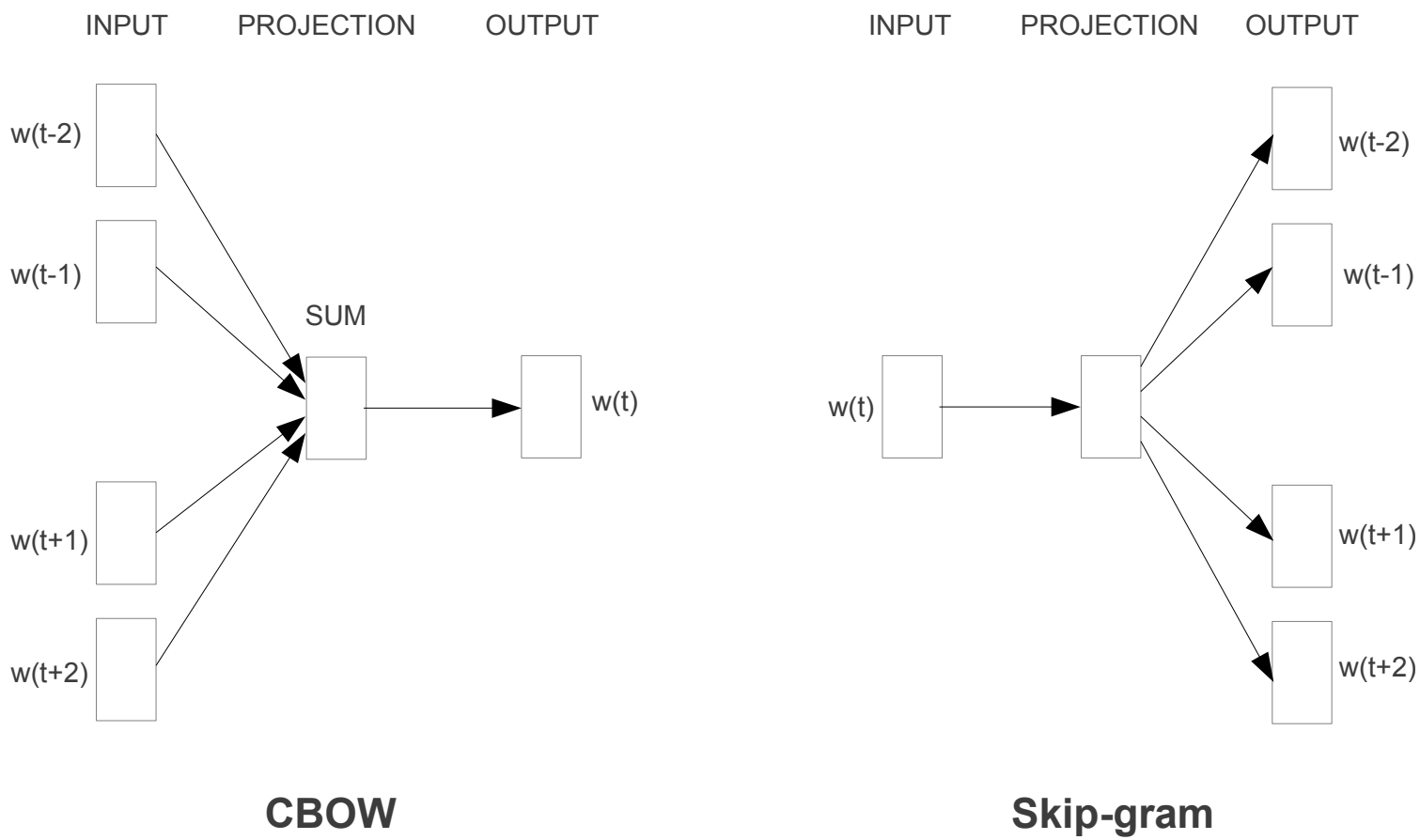
- Foundations and Trends in Information Retrieval (FnTIR)
- Synthesis Lectures on Information Concepts, Retrieval, and Services by Morgan&Claypool Publishers

CIKM 2017

- Evaluation: offline and online
- Ranking: deep learning, online LTR, short text retrieval
- Infrastructure: query processing, index compression, efficiency
- Temporal data: news, online learning, stream mining
- User-related: crowdsourcing, user behavior, user characteristics, privacy
- Structured: graphs, networks, events and entities
- Classics: clustering, classification, summarization
- Deep learning
- Recommender systems, collaborative filtering
- Health analytics

Outline

Word2vec

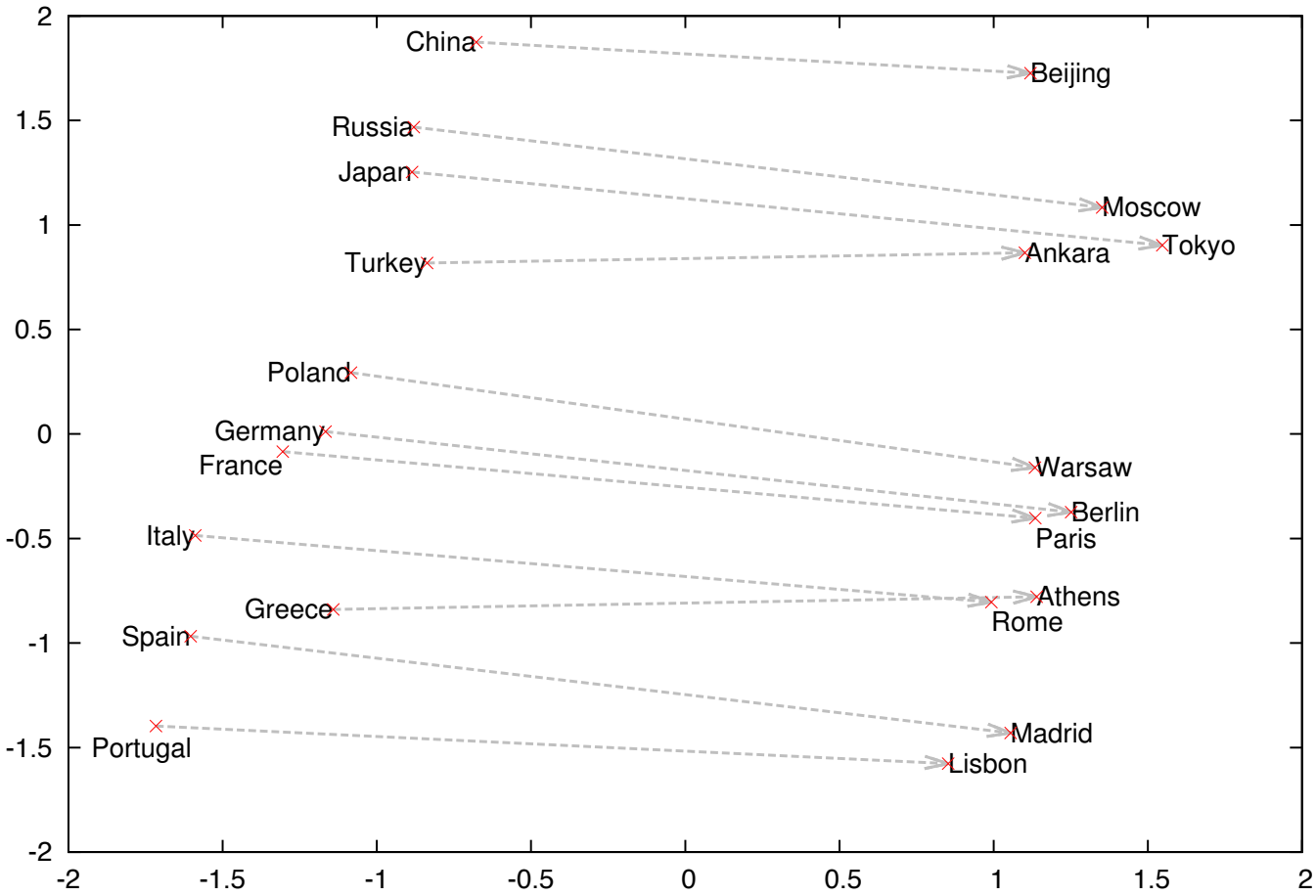


Mikolov et al., "Efficient Estimation of Word Representations in Vector Space"

Word2vec algorithm

1. Choose the length of embeddings
2. Initialize embeddings randomly
3. Update embeddings using gradient descent by optimizing CBOW or Skip-gram

Word2vec example



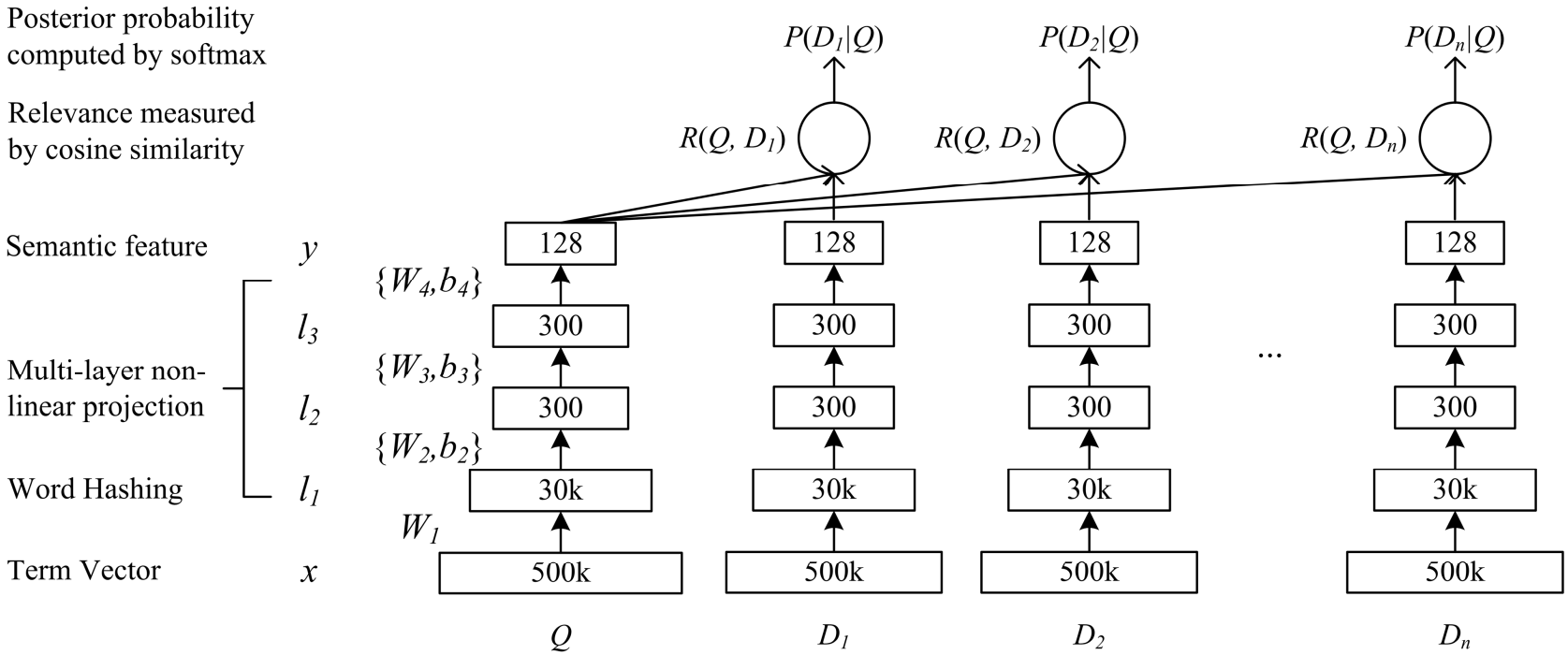
Mikolov et al., "Distributed Representations of Words and Phrases and their Compositionality"

Word2vec for retrieval

- Average word embeddings (centroids) for queries and documents
- Cosine similarity
- Works best for short documents

Outline

Deep structured semantic model



Huang et al. "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data"

Experimental comparison

#	Models	NDCG@1	NDCG@3	NDCG@10
1	TF-IDF	0.319	0.382	0.462
2	BM25	0.308	0.373	0.455
3	WTM	0.332	0.400	0.478
4	LSA	0.298	0.372	0.455
5	PLSA	0.295	0.371	0.456
6	DAE	0.310	0.377	0.459
7	BLTM-PR	0.337	0.403	0.480
8	DPM	0.329	0.401	0.479
9	DNN	0.342	0.410	0.486
10	L-WH linear	0.357	0.422	0.495
11	L-WH non-linear	0.357	0.421	0.494
12	**L-WH DNN**	**0.362**	**0.425**	**0.498**

- DNN – no semantic hashing
- L-WH linear – semantic hashing, NO non-linear activation functions
- L-WH non-linear – semantic hashing, non-linear activation functions
- L-WH DNN – full network

Huang et al. "Learning Deep Structured Semantic Models for Web Search using Clickthrough Data"

Outline

Semantic retrieval summary

- Latent semantic indexing/analysis
- Topic modeling (pLSA, LDA)
- Words embeddings (word2vec)
- Neural networks (DSSM)

Materials

- Manning et al., Chapter 18
- Blei et al.
  Latent Dirichlet Allocation
  Journal of Machine Learning Research, 2003
- Mikolov et al.
  Distributed Representations of Words and Phrases and their Compositionality
  Advances in neural information processing systems, 2013
- Huang et al.
  Learning Deep Structured Semantic Models for Web Search using Clickthrough Data
  Proceedings of CIKM, 2013

Ranking methods

1. Content-based
   - Term-based
   - Semantic
2. **Link-based (web search)**
3. Learning to rank