

# Web Retrieval

## Chapter 2 – Technical Basics

Steffen Staab, Sergej Sizov

<http://west.uni-koblenz.de>



### ▪ **Information Retrieval:**

- ◆ Soumen Chakrabarti: Mining the Web: Analysis of Hypertext and Semi-Structured Data, Morgan Kaufmann, 2002
- ◆ Christopher Manning, Prabhakar Raghavan, Hinrich Schütze: Introduction to Information Retrieval, Cambridge University Press, 2007
- ◆ Pierre Baldi, Paolo Frasconi, Padhraic Smyth: Modeling the Internet and the Web - Probabilistic Methods and Algorithms, Wiley & Sons, 2003.
- ◆ Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval, Addison-Wesley, 1999.
- ◆ Christopher D. Manning, Hinrich Schütze: Foundations of Statistical Natural Language Processing, MIT Press, 1999.
- ◆ David A. Grossman, Ophir Frieder: Information Retrieval: Algorithms and Heuristics, Springer, 2004.
- ◆ Ian H. Witten: Managing Gigabytes: Compressing and Indexing Documents and Images, Morgan Kaufmann, 1999.

### **Stochastics**

Larry Wasserman: All of Statistics, Springer, 2004.

George Casella, Roger L. Berger: Statistical Inference, Duxbury, 2002.

Arnold Allen: Probability, Statistics, and Queueing Theory with Computer Science Applications, Academic Press, 1990.

### **Machine Learning**

Richard O. Duda, Peter E. Hart, David G. Stork: Pattern Classification, Wiley&Sons, 2000.

Trevor Hastie, Robert Tibshirani, Jerome H. Friedman: Elements of Statistical Learning, Springer, 2001.

Tom M. Mitchell: Machine Learning, McGraw-Hill, 1997.

### **Tools and Programming**

Ian H. Witten, Eibe Frank: Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2005.

Erik Hatcher, Otis Gospodnetic: Lucene in Action, Manning Publications, 2004.

Tony Loton: Web Content Mining with Java, John Wiley&Sons, 2002.

### **important conferences on IR**

(see DBLP bibliography for full detail, <http://www.informatik.uni-trier.de/~ley/db/>)  
SIGIR, ECIR, CIKM, TREC, WWW, KDD, ICDM, ICML, ECML

### **online portals**

DBLP, Google Scholar, SiteSeer search engines

ACM, IEEE portals

Scientific mailing lists (e.g. DBWorld, AK-KDList, SIG-IRList, WebIR, DDLBETAtag, etc.)

### **evaluation initiatives:**

- Text Retrieval Conference (TREC), <http://trec.nist.gov>
- Cross-Language Evaluation Forum (CLEF), [www.clef-campaign.org](http://www.clef-campaign.org)
- Initiative for the Evaluation of XML Retrieval (INEX),  
<http://inex.is.informatik.uni-duisburg.de/>
- KDD Cup, <http://www.kdnuggets.com/datasets/kddcup.html>  
and <http://kdd05.lac.uic.edu/kddcup.html>
- Language-Independent Named-Entity Recognition,  
[www.cnts.ua.ac.be/conll2003/ner/](http://www.cnts.ua.ac.be/conll2003/ner/)

### **feel free to contact..**

a) lecturer, b) authors of publications, c) members of online communities  
and mailing lists

## Probability Theory & Stochastics

Events, Probabilities, Random Variables, Distributions,  
Basics from Information Theory, Markov chains

## Basics from Information Theory

Entropy, cross-entropy, information gain..

## Linear Algebra

Vectors and matrices, eigenvectors, common decompositions

# Part 1: Basics from probability theory and stochastics

A **probability space** is a triple  $(\Omega, E, P)$  with

- a set  $\Omega$  of elementary events (sample space),
- a family  $E$  (event space) of subsets of  $\Omega$  with  $\Omega \in E$  which is closed under  $\cap$ ,  $\cup$ , and  $-$  with a countable number of operands (with finite  $\Omega$  usually  $E=2^\Omega$ ), and
- a **probability measure  $P: E \rightarrow [0,1]$**  with  $P[\Omega]=1$  and  $P[\cup_i A_i] = \sum_i P[A_i]$  for countably many, pairwise disjoint  $A_i$

## Properties of P:

$$P[A] + P[\neg A] = 1$$

$$P[A \cup B] = P[A] + P[B] - P[A \cap B]$$

$$P[\emptyset] = 0 \text{ (null/impossible event)}$$

$$P[\Omega] = 1 \text{ (true/certain event)}$$

Two events  $A, B$  of a prob. space are **independent** if  $P[A \cap B] = P[A] P[B]$ .

A finite set of events  $A = \{A_1, \dots, A_n\}$  is **independent** if for every subset  $S \subseteq A$  the equation  $P[\bigcap_{A_i \in S} A_i] = \prod_{A_i \in S} P[A_i]$  holds.

The **conditional probability**  $P[A | B]$  of  $A$  under the condition (hypothesis)  $B$  is defined as: 
$$P[A | B] = \frac{P[A \cap B]}{P[B]}$$

Event  $A$  is **conditionally independent** of  $B$  given  $C$  if  $P[A | BC] = P[A | C]$ .

## Total probability theorem:

For a partitioning of  $\Omega$  into events  $B_1, \dots, B_n$ :

$$P[A] = \sum_{i=1}^n P[A | B_i] P[B_i]$$

**Bayes' theorem:** 
$$P[A | B] = \frac{P[A \cap B]}{P[B]} = \frac{P[B | A] P[A]}{P[B]}$$

$P[A|B]$  is called *posterior probability*

$P[A]$  is called *prior probability*

## Example: Probabilistic Retrieval with Term Independence

## Ranking Proportional to Relevance Odds

$$\mathit{sim}(d, q) = O(R | d) = \frac{P[R | d]}{P[\neg R | d]}$$

odds for relevance  
(ratio of relevant documents)

$$= \frac{P[d | R] \times P[R]}{P[d | \neg R] \times P[\neg R]}$$

Bayes' theorem

$$\sim \frac{P[d | R]}{P[d | \neg R]} = \prod_i \frac{P[X_i | R]}{P[X_i | \neg R]}$$

independence or  
linked dependence

$$\mathit{sim}(d, q)' = \log \prod_{i \in q} \frac{P[X_i | R]}{P[X_i | \neg R]}$$

$X_i = 1$  if  $d$  includes  
 $i$ -th term, 0 otherwise

$$= \sum_{i \in q} \log P[X_i | R] - \log P[X_i | \neg R]$$

estimate: 
$$P[d \in c_k | d \text{ has } \vec{X}] = \frac{P[d \text{ has } \vec{X} | d \in c_k] P[d \in c_k]}{P[d \text{ has } \vec{X}]}$$

$$\sim P[X | d \in c_k] P[d \in c_k]$$

$$= \prod_{i=1}^m P[X_i | d \in c_k] P[d \in c_k]$$

with feature independence  
or linked dependence:

$$\frac{P[X | d \in c_k]}{P[X | d \notin c_k]} = \prod_i \frac{P[X_i | d \in c_k]}{P[X_i | d \notin c_k]}$$

$$= \prod_{i=1}^m p_{ik}^{X_i} (1 - p_{ik})^{1 - X_i} p_k$$

with empirically estimated  
 $p_{ik} = P[X_i = 1 | c_k]$ ,  $p_k = P[c_k]$

$$\Rightarrow \log P[c_k | d] \sim \sum_{i=1}^m X_i \log \frac{p_{ik}}{(1 - p_{ik})} + \sum_{i=1}^m \log(1 - p_{ik}) + \log p_k$$

for binary classification with odds rather than probs for simplification

A **random variable (RV)**  $X$  on the prob. space  $(\Omega, E, P)$  is a function  $X: \Omega \rightarrow M$  with  $M \subseteq \mathbb{R}$  s.t.  $\{e \mid X(e) \leq x\} \in E$  for all  $x \in M$  ( $X$  is measurable).

$F_X: M \rightarrow [0,1]$  with  $F_X(x) = P[X \leq x]$  is the **(cumulative) distribution function (cdf)** of  $X$ .

With countable set  $M$  the function  $f_X: M \rightarrow [0,1]$  with  $f_X(x) = P[X = x]$  is called the **(probability) density function (pdf)** of  $X$ ;  
in general  $f_X(x)$  is  $F'_X(x)$ .

For a random variable  $X$  with distribution function  $F$ , the inverse function  $F^{-1}(q) := \inf\{x \mid F(x) > q\}$  for  $q \in [0,1]$  is called **quantile function** of  $X$ .  
(0.5 quantile (50<sup>th</sup> percentile) is called median)

Random variables with countable  $M$  are called **discrete**,  
otherwise they are called **continuous**.

For discrete random variables the density function is also referred to as the **probability mass function**.

- **Bernoulli** distribution with parameter  $p$ :  $P[X = x] = p^x (1-p)^{1-x}$   
for  $x \in \{0, 1\}$

- **Uniform** distribution over  $\{1, 2, \dots, m\}$ :

$$P[X = k] = f_X(k) = \frac{1}{m} \text{ for } 1 \leq k \leq m$$

- **Binomial** distribution (coin toss  $n$  times repeated;  $X$ : #heads):

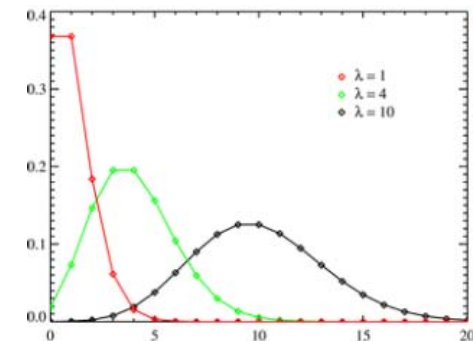
$$P[X = k] = f_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- **Poisson** distribution (with rate  $\lambda$ ):

$$P[X = k] = f_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

- **Geometric** distribution (#coin tosses until first head):

$$P[X = k] = f_X(k) = (1-p)^{k-1} p$$

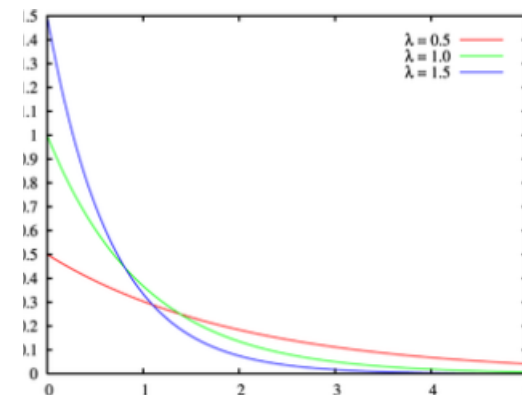


- **Uniform** distribution in the interval  $[a,b]$

$$f_X(x) = \frac{1}{b-a} \quad \text{for } a \leq x \leq b \quad (0 \text{ otherwise})$$

- **Exponential** distribution (z.B. time until next event of a Poisson process) with rate  $\lambda = \lim_{\Delta t \rightarrow 0} (\# \text{ events in } \Delta t) / \Delta t$ :

$$f_X(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0 \quad (0 \text{ otherwise})$$



- **Hyperexponential** distribution:  $f_X(x) = p\lambda_1 e^{-\lambda_1 x} + (1-p)\lambda_2 e^{-\lambda_2 x}$

Let  $X_1, \dots, X_m$  be random variables over the same prob. space with domains  $\text{dom}(X_1), \dots, \text{dom}(X_m)$ .

The *joint distribution* of  $X_1, \dots, X_m$  has a density function

$$f_{X_1, \dots, X_m}(x_1, \dots, x_m)$$

$$\text{with } \sum_{x_1 \in \text{dom}(X_1)} \dots \sum_{x_m \in \text{dom}(X_m)} f_{X_1, \dots, X_m}(x_1, \dots, x_m) = 1$$

$$\text{or } \int_{\text{dom}(X_1)} \dots \int_{\text{dom}(X_m)} f_{X_1, \dots, X_m}(x_1, \dots, x_m) dx_m \dots dx_1 = 1$$

The *marginal distribution* of  $X_i$  in the joint distribution of  $X_1, \dots, X_m$  has the density function

$$\sum_{x_1} \dots \sum_{x_{i-1}} \sum_{x_{i+1}} \dots \sum_{x_m} f_{X_1, \dots, X_m}(x_1, \dots, x_m) \text{ or}$$

$$\int_{X_1} \dots \int_{X_{i-1}} \int_{X_{i+1}} \dots \int_{X_m} f_{X_1, \dots, X_m}(x_1, \dots, x_m) dx_m \dots dx_{i+1} dx_{i-1} \dots dx_1$$

For a discrete random variable  $X$  with density  $f_X$

$$E[X] = \sum_{k \in M} k f_X(k) \quad \text{is the } \textit{expectation value (mean)} \text{ of } X$$

$$E[X^i] = \sum_{k \in M} k^i f_X(k) \quad \text{is the } \textit{i-th moment} \text{ of } X$$

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \quad \text{is the } \textit{variance} \text{ of } X$$

For a continuous random variable  $X$  with density  $f_X$

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx \quad \text{is the } \textit{expectation value} \text{ of } X$$

$$E[X^i] = \int_{-\infty}^{+\infty} x^i f_X(x) dx \quad \text{is the } \textit{i-th moment} \text{ of } X$$

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2 \quad \text{is the } \textit{variance} \text{ of } X$$

Theorem: Expectation values are additive:  $E[X + Y] = E[X] + E[Y]$   
 (distributions are not)

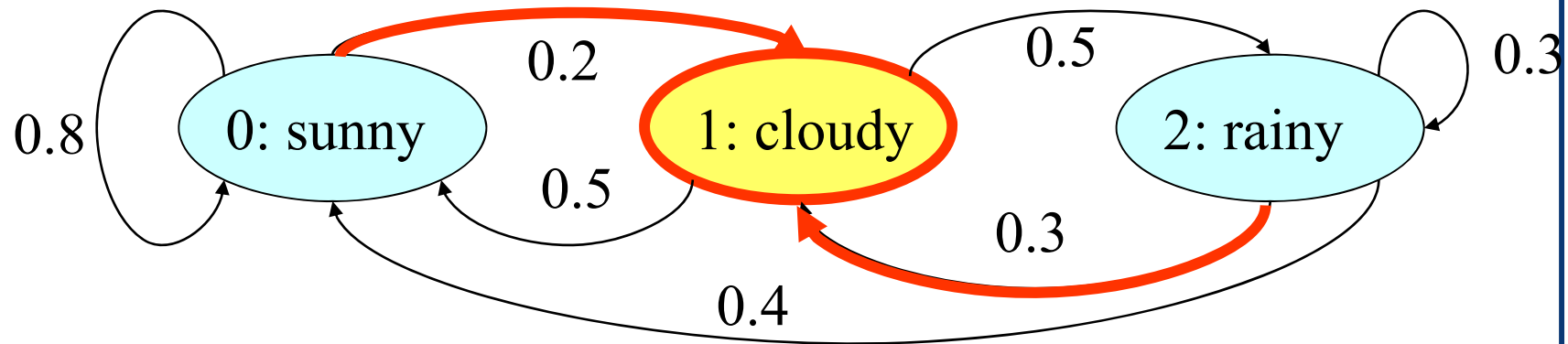
*Covariance* of random variables  $X_i$  and  $X_j$ :

$$\mathit{Cov}(X_i, X_j) := E[(X_i - E[X_i])(X_j - E[X_j])]$$

$$\mathit{Var}(X_i) = \mathit{Cov}(X_i, X_i) = E[X^2] - E[X]^2$$

*Correlation coefficient* of  $X_i$  and  $X_j$

$$\rho(X_i, X_j) := \frac{\mathit{Cov}(X_i, X_j)}{\sqrt{\mathit{Var}(X_i)} \sqrt{\mathit{Var}(X_j)}}$$



$$p_0 = 0.8 p_0 + 0.5 p_1 + 0.4 p_2$$

$$p_1 = 0.2 p_0 + 0.3 p_2$$

$$p_2 = 0.5 p_1 + 0.3 p_2$$

$$p_0 + p_1 + p_2 = 1$$

$$\Rightarrow p_0 \approx 0.657, p_1 = 0.2, p_2 \approx 0.143$$

state set: finite or infinite

time: discrete or continuous

state transition prob's:  $p_{ij}$

state prob's in step  $t$ :  $p_i^{(t)} = P[S(t)=i]$

Markov property:  $P[S(t)=i \mid S(0), \dots, S(t-1)] = P[S(t)=i \mid S(t-1)]$

interested in **stationary state probabilities**:

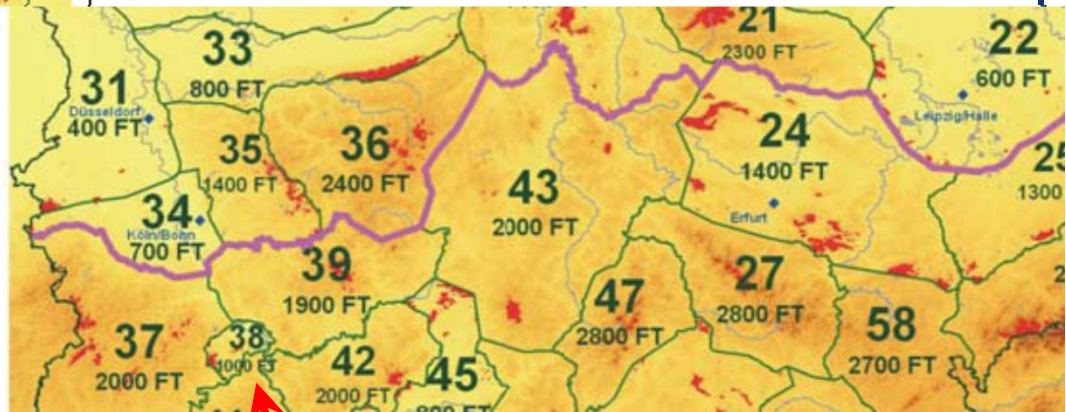
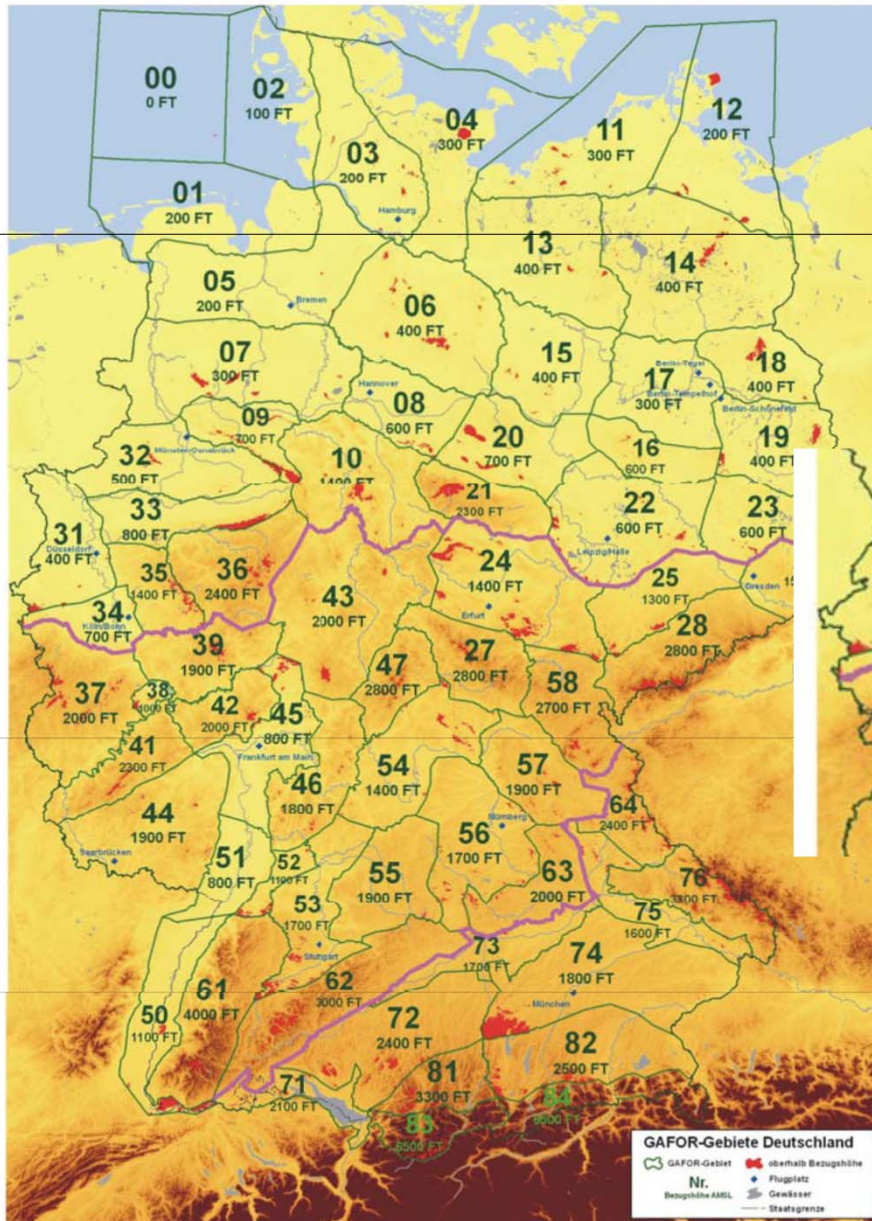
$$p_j := \lim_{t \rightarrow \infty} p_j^{(t)} = \lim_{t \rightarrow \infty} \sum_k p_k^{(t-1)} p_{kj}$$

$$p_j = \sum_k p_k p_{kj}$$

$$\sum_j p_j = 1$$

guaranteed to exist for irreducible, aperiodic, finite Markov chains

# Example: GAFOR Regions for Germany



38 = Neuwieder  
Becken !

A **stochastic process** is a family of random variables  $\{X(t) \mid t \in T\}$ .

$T$  is called parameter space, and the domain  $M$  of  $X(t)$  is called state space.  $T$  and  $M$  can be discrete or continuous.

A stochastic process is called **Markov process** if for every choice of  $t_1, \dots, t_{n+1}$  from the parameter space and every choice of  $x_1, \dots, x_{n+1}$  from the state space the following holds:

$$\begin{aligned} &P [ X(t_{n+1}) = x_{n+1} \mid X(t_1) = x_1 \wedge X(t_2) = x_2 \wedge \dots \wedge X(t_n) = x_n ] \\ &= P [ X(t_{n+1}) = x_{n+1} \mid X(t_n) = x_n ] \end{aligned}$$

A Markov process with discrete state space is called **Markov chain**.

A canonical choice of the state space are the natural numbers.

Notation for Markov chains with discrete parameter space:

$X_n$  rather than  $X(t_n)$  with  $n = 0, 1, 2, \dots$

The Markov chain  $X_n$  with discrete parameter space is

**homogeneous** if the transition probabilities  $p_{ij} := P[X_{n+1} = j \mid X_n = i]$  are independent of  $n$

**irreducible** if every state is reachable from every other state with positive probability:

$$\sum_{n=1}^{\infty} P[X_n = j \mid X_0 = i] > 0 \quad \text{for all } i, j$$

**aperiodic** if every state  $i$  has period 1, where the period of  $i$  is the greatest common divisor of all (recurrence) values  $n$  for which

$$P[X_n = i \wedge X_k \neq i \text{ for } k = 1, \dots, n-1 \mid X_0 = i] > 0$$

The Markov chain  $X_n$  with discrete parameter space is

**positive recurrent** if for every state  $i$  the recurrence probability is 1 and the mean recurrence time is finite:

$$\sum_{n=1}^{\infty} P[X_n = i \wedge X_k \neq i \text{ for } k = 1, \dots, n-1 | X_0 = i] = 1$$

$$\sum_{n=1}^{\infty} n P[X_n = i \wedge X_k \neq i \text{ for } k = 1, \dots, n-1 | X_0 = i] < \infty$$

**ergodic** if it is homogeneous, irreducible, aperiodic, and positive recurrent.

For the **n-step transition probabilities**

$p_{ij}^{(n)} := P[X_n = j | X_0 = i]$  the following holds:

$$p_{ij}^{(n)} = \sum_k p_{ik}^{(n-1)} p_{kj} \text{ with } p_{ij}^{(1)} := p_{ik}$$

$$= \sum_k p_{ik}^{(n-l)} p_{kj}^{(l)} \text{ for } 1 \leq l \leq n-1$$

in matrix notation:  $P^{(n)} = P^n$

For the **state probabilities after n steps**

$\pi_j^{(n)} := P[X_n = j]$  the following holds:

$$\pi_j^{(n)} = \sum_i \pi_i^{(0)} p_{ij}^{(n)} \text{ with initial state probabilities } \pi_i^{(0)}$$

in matrix notation:  $\Pi^{(n)} = \Pi^{(0)} P^{(n)}$  *(Chapman-Kolmogorov equation)*

Every homogeneous, irreducible, aperiodic Markov chain with a finite number of states is positive recurrent and ergodic.

For every ergodic Markov chain there exist **stationary state probabilities**

$$\pi_j := \lim_{n \rightarrow \infty} \pi_j^{(n)}$$

These are independent of  $\Pi^{(0)}$

and are the solutions of the following system of linear equations:

$$\pi_j = \sum_i \pi_i p_{ij} \quad \text{for all } j \quad (\text{balance equations})$$

$$\sum_j \pi_j = 1$$

in matrix notation:  $\Pi = \Pi P$   
 (with  $1 \times n$  row vector  $\Pi$ )  $\Pi \vec{1} = 1$

- Part 2: Basics of Information Theory

Let  $f(x)$  be the probability (or relative frequency) of the  $x$ -th symbol in some text  $d$ . The **entropy** of the text (or the underlying prob. distribution  $f$ ) is:

$$H(d) = \sum_x f(x) \log_2 \frac{1}{f(x)}$$

$H(d)$  is a lower bound for the bits per symbol needed with optimal coding (compression).

For two prob. distributions  $f(x)$  and  $g(x)$  the **relative entropy (Kullback-Leibler divergence)** of  $f$  to  $g$  is

$$D(f \parallel g) := \sum_x f(x) \log \frac{f(x)}{g(x)}$$

Relative entropy is a measure for the (dis-)similarity of two probability or frequency distributions.

It corresponds to the average number of additional bits needed for coding information (events) with distribution  $f$  when using an optimal code for distribution  $g$ .

The **cross entropy** of  $f(x)$  to  $g(x)$  is:

$$H(f, g) := H(f) + D(f \parallel g) = - \sum_x f(x) \log g(x)$$

- Text is sequence of symbols (with specific frequencies)
- Symbols can be
  - letters or other characters from some alphabet  $\Sigma$
  - strings of fixed length (e.g. trigrams)
  - or words, bits, syllables, phrases, etc.

## *Limits of compression:*

Let  $p_i$  be the probability (or relative frequency)  
of the  $i$ -th symbol in text  $d$

Then the *entropy* of the text: 
$$H(d) = \sum_i p_i \log_2 \frac{1}{p_i}$$
is a *lower bound* for the average number of bits per symbol  
in any compression (e.g. Huffman codes)

## Note:

compression schemes such as *Ziv-Lempel* (used in zip)  
are better because they consider context beyond single symbols;  
with appropriately generalized notions of entropy  
the lower-bound theorem does still hold

## Part 3: Using Vectors and Matrices

## Vector space model revisited..

Terms characterize documents (term-document matrix)

But also: documents characterize terms..

User/resource/tag relationships in folksonomies

Alternate document/term representation forms also possible:

- e.g. concept based (using thesaurus like WordNet)

- e.g. with rich background document corpus (using Wikipedia in ESA)

But matrix representation is also frequently used for

Representing connections in graph models (adjacency matrix)

Representing transitions in stochastic models (probabilistic matrix)

.. and many other IR relevant applications and methods

A set  $S$  of vectors is called **linearly independent** if no  $x \in S$  can be written as a linear combination of other vectors in  $S$ .

The **rank** of matrix  $A$  is the maximal number of linearly independent row or column vectors.

A **basis** of an  $n \times n$  matrix  $A$  is a set  $S$  of linearly independent row or column vectors such that all rows or columns are linear combinations of vectors from  $S$ .

A set  $S$  of  $n \times 1$  vectors is an **orthonormal basis** if for all  $x, y \in S$ :

$$\|x\|_2 := \sqrt{\sum_{i=1}^n x_i^2} = 1 = \|y\|_2 \quad \text{and} \quad x \cdot y = 0$$

Let  $A$  be a real-valued  $n \times n$  matrix,  $x$  a real-valued  $n \times 1$  vector, and  $\lambda$  a real-valued scalar. Solutions  $x \neq 0$  and  $\lambda$  of the equation  $Ax = \lambda x$  are called an **Eigenvector** and **Eigenvalue** of  $A$ .

Eigenvectors of  $A$  are vectors whose direction is preserved by the linear transformation described by  $A$ .

The Eigenvalues of  $A$  are the roots (Nullstellen) of the characteristic polynomial  $f(\lambda)$  of  $A$ :

$$f(\lambda) = |A - \lambda I| = 0$$

with the determinant (developing the  $i$ -th row):

$$|A| = \sum_{j=1}^n (-1)^{i+j} a_{ij} |A^{(ij)}| \quad \text{where matrix } A^{(ij)} \text{ is derived from } A \text{ by removing the } i\text{-th row and the } j\text{-th column}$$

The real-valued  $n \times n$  matrix  $A$  is **symmetric** if  $a_{ij} = a_{ji}$  for all  $i, j$ .

$A$  is **positive definite** if for all  $n \times 1$  vectors  $x \neq 0$ :  $x^T \times A \times x > 0$ .

If  $A$  is symmetric then all Eigenvalues of  $A$  are real.

If  $A$  is symmetric and positive definite then all Eigenvalues are positive.