

# ProbMap - A Probabilistic Approach for Mapping Large Document Collections

Thomas Hofmann

*Department of Computer Science, Brown University  
Box 1910, Providence, RI 02912, th@cs.brown.edu*

---

## Abstract

The visualization of large text databases and document collections is an important step towards more flexible and interactive types of information access and retrieval. This paper presents a probabilistic approach which combines a statistical, model-based analysis of a given set of document with a topological visualization principle. Our method can be utilized to derive *topic maps*, which represent topical information by characteristic keyword distributions arranged in a two-dimensional spatial layout. Combined with multi-resolution techniques this provides a three-dimensional space for interactive information navigation in large text collections.

*Key words:* information retrieval, data mining, machine learning, latent class models, data visualization, self-organizing map

---

## 1 Introduction

Despite of the great enthusiasm and excitement our time shows for all types of new media, it is indisputable that the most nuanced and sophisticated medium to express or communicate our thoughts is what Herder calls the ‘vehiculum of our thoughts and the content of all wisdom and knowledge’[7] – our language. Consequently, prodigious benefits could result from the enhanced circulation and propagation of recorded language by todays digital networks, which make abundant repositories of text documents such as electronic libraries available to a large public. Yet, the availability of large databases does not automatically imply easy access to relevant information. Retrieving information from a glut of nuisance data can be tedious and extremely time consuming. Hence, there is a high demand for intelligent tools and navigation aids that provide uncomplicated and fast access to information, preferably on different level of resolution and abstraction. From the viewpoint of human-computer interaction, *information visualization* plays an important role, mainly for two reasons:

First, the human visual system is capable of processing large amounts of data in parallel which allows high-bandwidth communication. Second, a two- or three-dimensional spatial organization of information is appealing, since it is in accord with our natural navigation skills.

Obviously, there are important issues of *how* to visualize data, *e.g.*, questions concerning the geometry of the visualization space, the use of colors and textures, the representation of structured objects like graphs and alike. Yet, there is also the complementary problem of how to map the data to a visual representation such that the relevant structure as well as important relationships between the data are preserved. The latter problem almost inevitably involves *data analysis*, for example, issues like data reduction, grouping and decomposition: Which aspects of the data do we want to visualize? Which properties and relations should be preserved? What is the relevant, information carrying signal, and what should be considered as noise? These problems are often solved manually, *i.e.*, domain experts provide a mapping from abstract properties of data objects to visible properties in their visual representation.

This paper presents a data-driven, *statistical* approach to visualizing large collections of text documents by generating two-dimensional map displays. It aims at a concise visualization of the topical content of a collection as well as a representation of conceptual and topical similarities between documents or aspects of documents in the form of *topic maps*<sup>1</sup>. The proposed method has two building blocks:

- (i) A technique called *probabilistic latent semantic analysis* [8,3] which models context-dependent word occurrences and performs a topical decomposition of the document collection.
- (ii) A principle of *topology preservation* [13] which allows to visualize the extracted information in a spatial or topological layout, for example, in the form of a two-dimensional map.

Herein, data analysis and visualization are not treated as separate procedural stages: a single objective function combines a statistical criterion with topological constraints to ensure visualization. From the viewpoint of data analysis, this coupling makes sense, whenever the final end is not the analysis per se, but the presentation and visualization of regularities and patterns to a user. As a general principle, the latter implies that the value of an analysis carried out by means of a machine learning algorithm depends on whether or not its results can be represented in a way which makes it amenable to human (visual) inspection. Obviously, it can be of great advantage, if this is taken into account as early as possible in the analysis and not in a *post hoc* manner.

---

<sup>1</sup> Our notion of topic maps differs from the definition of this term according to the international standard (ISO/IEC CD 13250). The latter specifies merely a syntax for representing knowledge, while our method is a data analysis technique.

The rest of the paper is organized as follows: Section 2 introduces a probabilistic method for latent semantic analysis, which is extended in Section 3 to incorporate topological constraints. Finally, Section 4 shows some exemplary results of multi-resolution maps extracted from document collections.

## 2 Probabilistic Latent Semantic Analysis

### 2.1 Data Representation and Modeling

*Probabilistic Latent Semantic Analysis* (PLSA) [8,9] is a general method for statistical factor analysis of two-mode and count data which we apply here to learning from document collections. Suppose a collection of documents  $\mathcal{D} = \{d^1, \dots, d^I\}$  over some fixed vocabulary of words or terms  $\mathcal{W} = \{w^1, \dots, w^J\}$  is given, *i.e.*, each document consists of a sequence (vector) of tokens  $d^i = (d_1^i, \dots, d_{L(i)}^i)$ ,  $d_t^i \in \mathcal{W}$ , where  $L(i)$  denotes the length of the  $i$ -th document. We will make an assumption known as the ‘bag-of-words’ view, presupposing that conditioned on the identity of a particular document, words are generated independently. Formally, we assume that for a fixed document  $d^i$  tokens  $d_t^i$  are the outcome of repeated random trials  $D_t^i$  with sample space  $\mathcal{W}$  and a (document-specific) probability mass function  $P$ . The latter is represented in terms of a parameter vector  $\pi^i \in \mathbb{R}^J$  with coefficients  $\pi_j^i \equiv P(D_t^i = w^j)$ .

The bag-of-words view allows us to reduce the document collection to a rectangular matrix  $\mathbf{N}$  of word counts; entries  $n_j^i$  indicate how often a word  $w^j$  occurred in a document  $d^i$ ,  $n_j^i \equiv |\{d_t^i : d_t^i = w^j, 1 \leq t \leq L(i)\}|$ . For each document, the random vector of count variables  $n^i = (n_1^i, \dots, n_J^i)$  has a multinomial distribution

$$n^i \sim \text{Multinom}(L(i), \pi^i), \text{ where } \pi^i \geq 0 \text{ and } \sum_{j=1}^J \pi_j^i = 1, i=1, \dots, I. \quad (1)$$

The word counts summarized in  $\mathbf{N}$  are sufficient statistics and the maximum likelihood estimates of  $\pi^i$  can simply be expressed as

$$\hat{\pi}_j^i = n_j^i / L(i), \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (2)$$

In the context of information retrieval,  $\mathbf{N}$  is referred to as the *term-document matrix*, while in more general statistical terminology it is usually called a two-way cross-classification scheme or a two-dimensional *contingency table*. The term-document matrix is also the basis for the popular *vector space model*

[20] and it is known that  $\mathbf{N}$  will in many cases preserve most of the relevant information, *e.g.*, for tasks like text retrieval based on keywords, which makes it a reasonable starting point for our purposes.

The term-document matrix immediately reveals the problem of *data sparseness*, which is one of the problems latent semantic analysis aims to address. A typical matrix derived from short texts like news stories, book summaries, or paper abstracts may only have a tiny fraction of non-zero entries. This has consequences, in particular for methods that are evaluating similarities between documents by comparing or counting common terms. The main goal of PLSA in this context is to map documents and words to a more suitable representation in a *probabilistic latent semantic space* and to derive smoothed estimates for  $\pi^i$ . As the name suggests, the representation of documents and terms in this space is supposed to make semantic relations more explicit. PLSA is an attempt to achieve this goal in a purely data driven fashion without recourse to general linguistic knowledge.

## 2.2 Probabilistic Latent Semantic Analysis

In PLSA, dimension reduction is used to derive simultaneous estimates for all probability mass functions  $\pi^i$ ,  $i = 1, \dots, I$ . For some prespecified integer  $K \geq 1$ , let us introduce parameter vectors  $\phi^k \in [0; 1]^J$ ,  $k = 1, \dots, K$  and  $\tau^i \in [0; 1]^K$ ,  $i = 1, \dots, I$  which fulfill the following set of normalization constraints

$$\sum_{j=1}^J \phi_j^k = 1, \quad k = 1, \dots, K \quad \text{and} \quad \sum_{k=1}^K \tau_k^i = 1, \quad i = 1, \dots, I. \quad (3)$$

In terms of these parameters, the PLSA model can be defined as

$$\pi_j^i(\phi, \tau) = \sum_{k=1}^K \phi_j^k \tau_k^i, \quad i = 1, \dots, I, \quad j = 1, \dots, J. \quad (4)$$

The model in (4) stipulates a particular functional form for the probabilities  $\pi_j^i$  and effectively imposes algebraic constraints on the parameter vectors  $\pi^i$ : they have to be expressible in terms of convex combinations of  $K$  prototypical parameter vectors  $\phi^k$  which we call *factors*.

PLSA can be interpreted in terms of a latent class or mixture model: With each token  $d_t^i$  a latent class variable  $Z_t^i$  over an event space  $\mathcal{Z} = \{z^1, \dots, z^K\}$  is associated. Formally, these variables model repeated multinomial trials with probability mass function  $\tau_k^i \equiv P(Z_t^i = z^k)$ ; hence a different probability law

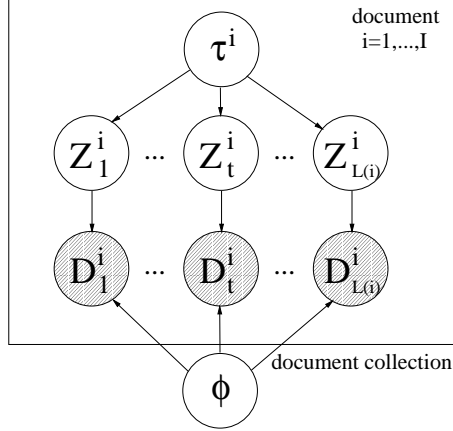


Fig. 1. Graphical model representation of Probabilistic Latent Semantic Analysis.

parameterized by  $\tau^i$  applies to the variables associated with different documents. The parameters  $\phi^k$  then define conditional probabilities over  $\mathcal{W}$  for each latent class  $z^k$ ,  $\phi_j^k \equiv P(D_t^i = w^j | Z_t^i = z^k)$ . The generative model implicit in (4) can thus be formulated as follows:

```

for  $i = 1, \dots, I$ 
  for  $t = 1, \dots, L(i)$ 
    randomly generate a state  $z^k \in \mathcal{Z}$  with probability  $\tau_k^i$ ,
       $Z_t^i \leftarrow z^k$ 
    randomly generate a word  $w^j$  with probability  $\phi_j^k$ ,
       $D_t^i \leftarrow w^j$ 
  end
end

```

A graphical model representation of PLSA is depicted in Fig. 1. The variables in the frame are replicated for  $i = 1, \dots, I$  with an appropriate choice of  $L(i)$ .

The intention pursued by the introduction of latent variables is to model *topics* such that each possible state  $z^k$  would ideally represent one topic and the parameters  $\tau_k^i$  would indicate to what extent a document  $d^i$  deals with each particular topic. In this view, the  $\phi^k$  parameters are supposed to model a topic-specific word distribution, *i.e.*, words that characterize a topic  $z^k$  well should have large probabilities  $\phi_j^k$ .

### 2.3 Parameter Estimation

In order fit the model in (4), we follow the statistical standard procedure and perform maximum likelihood estimation with the Expectation Maximization (EM) algorithm [4,21]. Our goal is to maximize the log-likelihood

$$\mathcal{L}(\phi, \tau; \mathbf{N}) = \sum_{i=1}^I \sum_{j=1}^J n_j^i \log \pi_j^i(\phi, \tau) \quad (5)$$

with respect to  $\phi = (\phi^1, \dots, \phi^K)$  and  $\tau = (\tau^1, \dots, \tau^I)$ . EM maximizes  $\mathcal{L}$  locally by alternating two steps: (i) an expectation (E) step where posterior probabilities for the unobserved (latent) variables  $Z_t^i$  are computed based on the current estimates of the parameters<sup>2</sup>, (ii) a maximization (M) step, where parameters are updated based on the posterior probabilities computed in the E-step.

For the E-step one takes the prior probabilities  $P(Z_t^i = z^k) = \tau_k^i$  as a starting point and applies Bayes' rule to compute the posterior for all latent class variables

$$P(Z_t^i = z^k | D_t^i = w^j; \tau, \phi) = \frac{\tau_k^i \phi_j^k}{\sum_{l=1}^K \tau_l^i \phi_l^j}. \quad (6)$$

We will use the shorthand notation  $P(z^k | w^j; \tau^i, \phi) \equiv P(Z_t^i = z^k | D_t^i = w^j; \tau, \phi)$ , since posterior probabilities for different  $Z_{t_1}^i$  and  $Z_{t_2}^i$  are equal whenever the corresponding tokens are the same.

In order to derive the M-step equations one has to maximize the so-called expected complete data log-likelihood  $\mathcal{Q}$  with respect to  $\phi$  and  $\tau$  [11].  $\mathcal{Q}$  is given by

$$\mathcal{Q}(\phi, \tau; \bar{\phi}, \bar{\tau}) = \sum_{i=1}^I \sum_{j=1}^J n_j^i \sum_{k=1}^K P(z^k | w^j; \bar{\tau}^i, \bar{\phi}) [\log \tau_k^i + \log \phi_j^k]. \quad (7)$$

Taking the normalization constraints into account, the stationary conditions are given by

$$\phi_j^k = \frac{\sum_{i=1}^I n_j^i P(z^k | w^j; \bar{\tau}^i, \bar{\phi})}{\sum_{j'=1}^J \sum_{i=1}^I n_{j'}^i P(z^k | w^{j'}; \bar{\tau}^i, \bar{\phi})}, \quad \tau_k^i = \frac{\sum_{j=1}^J n_j^i P(z^k | w^j; \bar{\tau}^i, \bar{\phi})}{\sum_{j=1}^J n_j^i}. \quad (8)$$

Alternating (6) and (8) initialized from randomized starting conditions results in a procedure which will converge to a local maximum of the log-likelihood in (5).<sup>3</sup>

<sup>2</sup> Equivalently, this step calculates the expected sufficient statistics of the complete data model.

<sup>3</sup> Strictly speaking, the PLSA model is only identifiable from the complete data, but not from the incomplete data. A thorough treatment of this issue is beyond the scope of this paper (cf. [5]), but we would like to note that the lack of identifiability is due

image processing	speech recognition	video coding
image segment	speaker	video
texture	speech recognition	sequence
color	signal	motion
tissue	train	frame
brain	hmm	scene
slice	segment	segment
cluster	source	shot
mri	speaker	image
volume	segment	cluster
	sound	visual

Bosnia	Iraq	Rwanda
un	iraq	refugees
bosnian	iraqi	aid
serbs	sanctions	rwanda
bosnia	kuwait	relief
serb	un	people
sarajevo	council	camps
nato	gulf	zaire
peacekeepers	saddam	camp
nations	baghdad	food
peace	hussein	rwandan

Fig. 2. (a) The three latent factors to most likely generate the word ‘segment’, derived from a  $K = 128$  PLSA of the CLUSTER document collection. The displayed terms are the ones with the highest class-conditional probabilities  $\phi_j^k$ . (b) Three factors to most likely generate the word ‘UN’ from a 128 factor decomposition of the TDT1 corpus. (The headline descriptions were not generated automatically.)

#### 2.4 Example: Analysis of Word Usage with PLSA

Let us briefly discuss an elucidating example application of PLSA at this point. For illustrative purposes, we have run PLSA with  $K = 128$  factors on two datasets: (i) CLUSTER, a collection of paper abstracts on clustering and (ii) the TDT1 collection (cf. Section 4 for details).

As a particularly interesting term in the CLUSTER domain we have selected the word ‘segment’. Fig. 2 (a) shows the most probable words of three out of the 128 factors which have the highest probability to generate the term ‘segment’. This sketchy characterization reveals very meaningful sub-domains: The first factor deals with image processing, where ‘segment’ refers to a region in an image. The second factor describes speech recognition where ‘segment’

---

to the redundancy in the specification of a  $(K - 1)$ -dimensional affine subspace by  $K$  ‘points’  $\phi^k$ . As a consequence, the log-likelihood will have ridges of local maxima and the EM algorithm will converge to a point on a ridge. There are various ways to guarantee identifiability: In our experiments, we have utilized a technique called tempered EM [8] that favors solutions  $(\phi, \tau)$  that yield a higher entropy for the joint posterior probabilities of the latent class variables. Interestingly, the ProbMap does not suffer from this problem, since it introduces additional couplings between the different  $\phi^k$ ’s.

refers to a phonetic unit of an acoustic signal such as a phoneme. The third factor deals with video coding, where ‘segment’ is used in the context of motion segmentation in image sequences. The factors thus seem to capture relevant topics in the domain under consideration.

Three factors from the decomposition of the TDT1 collections with a high probability for the term ‘UN’ are displayed in Fig. 2 (b). The vocabulary clearly characterizes news stories related to certain incidents in the period of 1994/1995 covered by the TDT1 collection. The first factor deals with the war in Bosnia, the second with UN sanctions against Iraq, and the third with Rwanda. These examples show that the factors extracted by PLSA might also correspond to *events*. Dependent on the training collection and the specific domain the notion of ‘topic’ has thus to be taken in a broader sense.

### 2.5 What is Missing?

From the example in Fig. 2 one can see that the factors learned by PLSA provide a fairly concise description of *topics* or *events*, which can potentially be utilized for interactive retrieval and navigation. However, there is one major drawback: assuming that for large text collections one would like to perform PLSA with a latent space dimensionality of the order of several hundreds or even thousands, it seems inappropriate to expect that the user will examine all factors in search for relevant documents and topics of interest. Of course, one may ask the user to provide additional keywords to narrow the search, but this is only an *ad hoc* remedy. What is really missing in PLSA is a relationship between the different factors. Suppose for concreteness a relevant topic represented by some  $\phi^k$  has been identified by the user; the index  $k$  of the factor does not provide any information about whether or not another factor with index  $k'$  could be related and potentially relevant as well, since the numbering of factors is arbitrary.

The *ProbMap* is a generalization of PLSA, which extends the model in a way that enables it to capture additional information about the relationships between topics. In the case of a two-dimensional map, this results in a spatial arrangement of topics on a two-dimensional grid, a format which may support different types of visualization and navigation. Other topologies can be obtained by exactly the same mechanism described in the sequel.

### 3 Topological PLSA: The ProbMap Method

#### 3.1 The Self-Organizing Map

In order to extend the PLSA model in the described way, we make use of a principle that was originally proposed in the seminal work of Kohonen on Self-Organizing Maps (SOM) [13,14]. While the formulation of the algorithm in [13] was heuristic and mainly motivated in a biological setting, several authors have subsequently proposed modifications which have stressed an information theoretic foundation of the SOM and pointed out the relation to vector quantization for noisy communication channels (cf. [16,2,10]). Moreover, it has been noticed [1] that the topology-preserving properties of the SOM are independent of the vectorial representation most research on the SOM has been focusing on.

#### 3.2 The ProbMap

The key step in the ProbMap approach is to introduce a set of additional latent variables  $Y_t^i$ ,  $i = 1, \dots, I$ ,  $t = 1, \dots, L(i)$  with the same state space  $\mathcal{Z}$  as the  $Z_t^i$ . The rationale behind this is to define an additional layer of randomness, where a state  $z^k \in \mathcal{Z}$  is switched to a state  $z^l$  with some probability  $\alpha_l^k$ . By choosing these transition probabilities adequately, a topological organization of factors will be favored. The ProbMap defines the following generative process for words, which differs from the PLSA by an additional step (marked with \*):

```
for  $i = 1, \dots, I$ 
  for  $t = 1, \dots, L(i)$ 
    randomly generate a state  $z^k \in \mathcal{Z}$  with probability  $\tau_k^i$ ,
     $Y_t^i \leftarrow z^k$ 
    *randomly generate a new state  $z^l \in \mathcal{Z}$  with probability  $\alpha_l^k$ ,
     $Z_t^i \leftarrow z^l$ 
    randomly generate a word  $w^j$  with probability  $\phi_j^l$ ,
     $D_t^i \leftarrow w^j$ 
  end
end
```

This process effectively re-defines the conditional probabilities of generating words

$$\tilde{\phi}_j^k = \sum_{l=1}^L \phi_j^l \alpha_l^k, \quad \text{where } \sum_{l=1}^L \alpha_l^k = 1, \quad k = 1, \dots, K \quad (9)$$

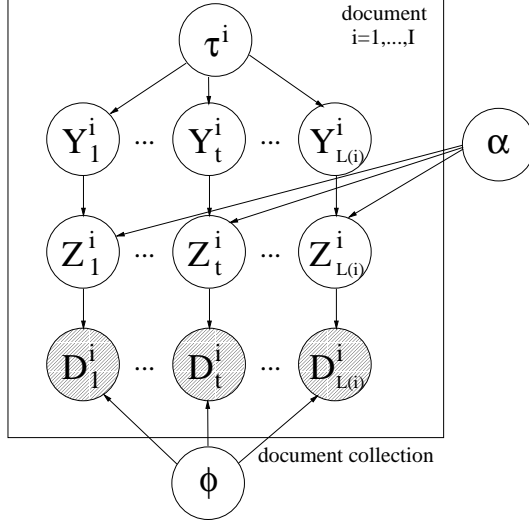


Fig. 3. Graphical model representation of the ProbMap.

and results in the overall model

$$\tilde{\pi}_j^i(\phi, \tau, \alpha) = \sum_{k=1}^K \tau_k^i \tilde{\phi}_j^k = \sum_{k=1}^K \tau_k^i \sum_{l=1}^K \phi_j^l \alpha_l^k \quad (10)$$

The graphical model representation of the ProbMap in Fig. 3 shows the additional layer of latent variables.

It is straightforward to verify that from a purely statistical point of view this does not offer any additional modeling power. Whatever the choice of  $\alpha_l^k$  and  $\phi_j^k$ , by re-defining  $\phi_j^k \equiv \tilde{\phi}_j^k$  one obtains the same conditional probabilities in the more parsimonious model of (4). Yet, we do *not* propose to fit the  $\alpha$  parameters from training data, but to set them to prespecified values derived from a *neighborhood function* in the latent variable space  $\mathcal{Z}$ . We will focus on means to enforce a topological organization of the topic representations  $\phi^k$  on a two-dimensional grid.

Let us therefore introduce the notation  $z(x, y)$ ,  $x = 1, \dots, R$ ,  $y = 1, \dots, R$  to identify latent states  $z(x, y) \equiv z^{k(x,y)} \in \mathcal{Z}$  with points  $(x, y)$  on a square grid.<sup>4</sup> By the Euclidean metric, this embedding induces a distance function on  $\mathcal{Z}$ , namely

$$d(z(x, y), z(x', y')) = d((x, y), (x', y')) = \sqrt{(x - x')^2 + (y - y')^2}. \quad (11)$$

With this distance function on  $\mathcal{Z}$  we propose to define the transition probabilities  $\alpha_l^k$  via a Gaussian function with standard deviation  $\sigma$

<sup>4</sup> For example, one may utilize the function  $k(x, y) = R(x - 1) + y$  to map points  $(x, y)$  on a two-dimensional  $R \times R$  grid to a set of indices  $k$  ( $1 \leq k \leq K = R^2$ ).

$$\alpha_l^k = \frac{\exp \left[ -d(z^k, z^l)^2 / (2\sigma^2) \right]}{\sum_{m=1}^K \exp \left[ -d(z^k, z^m)^2 / (2\sigma^2) \right]}, \quad (12)$$

where  $\sigma$  is assumed to be fixed for now. To understand why this favors a topological organization of topics, consider a document  $d^i$  with its topic distribution parameterized by  $\tau^i$ . As an effect of the additional randomization, these probabilities are tilted to yield  $\tilde{\tau}_l^i = \sum_{k=1}^K \alpha_l^k \tau_k^i$ , because every time a state  $z^k$  is selected it has some probability  $\alpha_l^k$  to be switched to  $z^l$ . For simplicity assume that  $\tau_k^i = 1$  for a particular  $z^k \in \mathcal{Z}$ , then the transition probabilities will blend-in additional contributions, mainly from neighboring states  $z^l$  of  $z^k$  on the two-dimensional grid for which  $\alpha_l^k$  is non-negligible. If these neighboring states represent very different topics, the resulting word distribution  $\tilde{\pi}_j^i$  in (10) will significantly deviate from the distribution one would get from (4), which – assuming that  $\tau^i$  was chosen optimal – will result in a poor estimate. If on the other hand the neighbors of  $z^k$  represent closely related topics, this deviation will in general be much less severe. A meaningful topological arrangement of topics will thus pay off in terms of a higher value of the log-likelihood.

### 3.3 Parameter Estimation in the ProbMap

The next step is the derivation of the EM equations for the ProbMap. As in the case of PLSA, we have to compute (marginal) posterior probabilities for the latent variables:

$$P_Y(z^k | w^j; \tau^i, \phi) \equiv P(Y_t^i = z^k | D_t^i = w^j; \tau, \phi) = \frac{\tau_k^i \sum_{l=1}^K \alpha_l^k \phi_j^l}{\sum_{k'=1}^K \tau_{k'}^i \sum_{l=1}^K \alpha_l^{k'} \phi_j^l}, \quad (13)$$

$$P_Z(z^k | w^j; \tau^i, \phi) \equiv P(Z_t^i = z^k | D_t^i = w^j; \tau, \phi) = \frac{\phi_j^k \sum_{l=1}^K \alpha_l^k \tau_l^i}{\sum_{k'=1}^K \phi_j^{k'} \sum_{l=1}^K \alpha_l^{k'} \tau_l^i}. \quad (14)$$

In analogy to (7) we can compute the expected complete data log-likelihood  $\mathcal{Q}$  and obtain M-step re-estimation formulae

$$\phi_j^k = \frac{\sum_{i=1}^I n_j^i P_Z(z^k | w^j; \bar{\tau}^i, \bar{\phi})}{\sum_{j'=1}^J \sum_{i=1}^I n_{j'}^i P_Z(z^k | w^{j'}; \bar{\tau}^i, \bar{\phi})}, \quad \tau_k^i = \frac{\sum_{j=1}^J n_j^i P_Y(z^k | w^j; \bar{\tau}^i, \bar{\phi})}{\sum_{j=1}^J n_j^i}. \quad (15)$$

Again, re-iterating (13,14) and (15) results in a sequence of estimates that will converge towards a local maximum of the observed data log-likelihood (for fixed parameters  $\alpha$ ).

### 3.4 Topologies and Hierarchies

There are two ways in which hierarchies are of interest in the context of the ProbMap: (i) To accelerate the PLSA by a multi-resolution optimization over a sequence of refined grids. (ii) To improve the visualization by offering multiple levels of abstraction or resolution on which the data can be visualized.

A significant computational improvement can be achieved by performing PLSA on a coarse grid, say starting on a  $2 \times 2$  grid, and then recursively prolongating the found solution according to a quadtree-like scheme. This yields a successive map refinement procedure and involves copying the parameters  $\phi_j^k$  – with a small random disturbance – to the successors of  $z^k$  on the finer grid and distributing  $\tau_k^i$  from the coarse level among its four successor states on the finer grid. This procedure has the additional advantage that it often leads to better topological arrangements, since it is less sensitive to ‘topological defects’.<sup>5</sup> The multi-resolution optimization is coupled with a schedule for  $\sigma$ , which defines the length-scale for the transition probabilities in (12). In our experiments we have utilized a schedule  $\sigma_n = (1/\sqrt[m]{2})^n \sigma_0$ , where  $m$  corresponds to the number of iterations performed at a particular resolution level, *i.e.*, after  $m$  iterations we have  $\sigma_{n+m} = (1/2)\sigma_m$ . Prolongation to a finer grid is performed at iterations  $n = m, 2m, 3m, \dots$ . When the final resolution level is reached,  $\sigma$  is further reduced up to some small  $\sigma_\infty$  without prolongation.

Notice that the topological organization of topics in the ProbMap has the further advantage to support a simple coarsening procedure for visualization at different resolution levels. The fact that neighboring latent states represent similar topics suggests to merge states, *e.g.*, four at a time, to generate a coarser map with word distributions  $\phi_j^k$  obtained by averaging over the associated distributions on the finer grid with appropriate weights  $w_l \equiv \sum_i L(i)\tau_l^i / \sum_i L(i)$ . One can thus dynamically navigate in a three-dimensional information space: vertically between topic maps of different resolution and horizontally inside a particular two-dimensional map.

### 3.5 Related Work

**Latent Semantic Analysis** Latent Semantic Analysis (LSA) [3] is a well known technique in information retrieval that has been applied to various problems such as *ad hoc* retrieval (search) and information filtering. The main property which LSA and PLSA have in common is that both methods perform a low rank decomposition of the term-document matrix. However, LSA is

---

<sup>5</sup> There is a large body of literature dealing with the topology-preserving properties of SOMs. The reader is referred to [14] and the references therein.

based on the *Singular Value Decomposition* (SVD) which expresses  $\mathbf{N}$  as an expansion in terms of left/right eigenvectors

$$\mathbf{N} = \sum_{k=1}^{\min\{I,J\}} \lambda_k (u^k \otimes v^k) \approx \sum_{k=1}^K \lambda_k (u^k \otimes v^k) = \hat{\mathbf{N}}, \text{ where} \quad (16)$$

$$\langle u^k, u^l \rangle = \langle v^k, v^l \rangle = \delta_{kl} \text{ and } \lambda_k \geq \lambda_{k+1}. \quad (17)$$

By keeping only the  $K < \min\{I, J\}$  dominant contributions in (16) one obtains a rank  $K$  approximation  $\hat{\mathbf{N}}$  to  $\mathbf{N}$  that is optimal in the sense of the Frobenius norm.

In order to stress the similarities as well as the differences between LSA and PLSA, we switch to a symmetric parameterization of the model in (4) by defining new parameters  $\lambda \in [0; 1]^K$ ,  $\psi^k \in [0; 1]^I$ :

$$\lambda_k \equiv \frac{\sum_{i=1}^I L(i) \tau_k^i}{\sum_{i=1}^I L(i)}, \quad \psi_i^k \equiv \frac{L(i) \tau_k^i}{\sum_{i'=1}^I L(i') \tau_k^{i'}}. \quad (18)$$

Notice that  $\lambda_k$  denotes the probability that the latent class variable associated with a generic word occurrence in the corpus is in state  $z^k$ , while  $\psi_i^k$  denotes the probability that a word occurrence with associated latent state  $z^k$  will be found in document  $d^i$ . One thus gets

$$\mathbf{N} \approx \left( \sum_{i=1}^I L(i) \right) \sum_{k=1}^K \lambda_k (\phi^k \otimes \psi^k), \text{ where } \sum_{k=1}^K \lambda_k = \sum_{i=1}^I \psi_i^k = \sum_{j=1}^J \phi_j^k = 1. \quad (19)$$

Notice that (19) and (16) both perform a decomposition of  $\mathbf{N}$  into  $K$  rank 1 matrices. However, the constraints imposed on the vectors that generate these rank 1 matrices are different. In the case of LSA, the crucial conditions are pairwise orthogonality and  $L_2$  normalization, while PLSA imposes non-negativity and  $L_1$  normalization, which is consistent with a probabilistic interpretation. The conditional word probabilities  $\psi_j^k$  are an important plus of PLSA since they facilitate the interpretability of the extracted topic factors, which is crucial in the application presented here.

The most fundamental difference between LSA and PLSA is the underlying optimization criterion. While PLSA maximizes a likelihood function, LSA is based on a least-squares approximation principle, reminiscent of a Gaussian noise assumption. The Gaussian noise model, however, is clearly not the adequate sampling model as can be seen, *e.g.*, from the fact that the rank  $K$  approximation in (16) may contain negative entries. PLSA on the other hand uses a multinomial sampling model and maintains a consistent probabilistic interpretation. The statistical approach offers important advantages since it

explicitly aims at minimizing word perplexity<sup>6</sup>. In addition, the probabilistic approach can take advantage of the well-established statistical theory for model selection and complexity control, *e.g.*, to determine the optimal number of latent space dimensions (cf. [8]). Last but not least, the statistical formulation can be systematically extended and generalized in various ways, one important example being the topological model presented in this section.

**The WebSOM Architecture** Our approach is related in spirit to the WebSOM learning architecture [12] which continues earlier work on semantic maps [18] and uses the Self-Organizing-Map [13] to perform a topological clustering of documents. Computationally, the SOM can be viewed as a stochastic gradient descent method of a clustering problem with a sum of squared distances as an objective function [19].

There are several differences between the ProbMap and the WebSOM approach. First of all, both approaches make different assumptions: The WebSOM is based on the idea of *document clustering* and hence aims at identifying cluster representatives. The topological arrangement of these cluster ‘centers’ then results in a document cluster map. On this map, every document can be assigned a unique coordinate given by the lattice point corresponding to the closest representative. The main idea behind the WebSOM is to find documents that are similar to a given query document by exploiting the spatial layout defined by the cluster map. The ProbMap on the other hand does not cluster documents, but performs a topical decomposition of the term-document matrix. Each document has a full distribution over topic factors and can thus ‘participate’ in several topics. This has two major advantages: (i) it offers more flexibility since single documents are modeled by topic *combinations* and (ii) topic factors typically are easier to interpret, since averaging over all words occurring in a group of documents is avoided. In fact, one might think of the WebSOM as a constrained version of the ProbMap, where the mixing weights  $\tau_k^i$  are restricted to Boolean values, in which case for a given document  $d^i$ ,  $\tau_k^i = 1$  for exactly one of the components. The latent states  $z^k$  encode document clusters in this case and one obtains a method known as distributional clustering (cf. [17,11]). The remaining difference between the WebSOM and this constrained version of the ProbMap is, of course, the objective function which is a likelihood in the case of the ProbMap in contrast to the squared error criterion utilized in the WebSOM. In addition, the ProbMap has the conceptual advantages of a statistical model which have already been stressed in the previous paragraph, while the WebSOM is by and large a heuristic procedure which lacks a precise probabilistic interpretation.

---

<sup>6</sup> Perplexity is a term from statistical language modeling which is utilized here to refer to the (log-averaged) inverse predictive probability,  $\text{Perplex} = \exp[-\sum_{i,j} \tilde{n}_j^i \log \pi_j^i / \sum_{i,j} \tilde{n}_j^i]$ , where  $\tilde{n}_j^i$  are counts on test data.

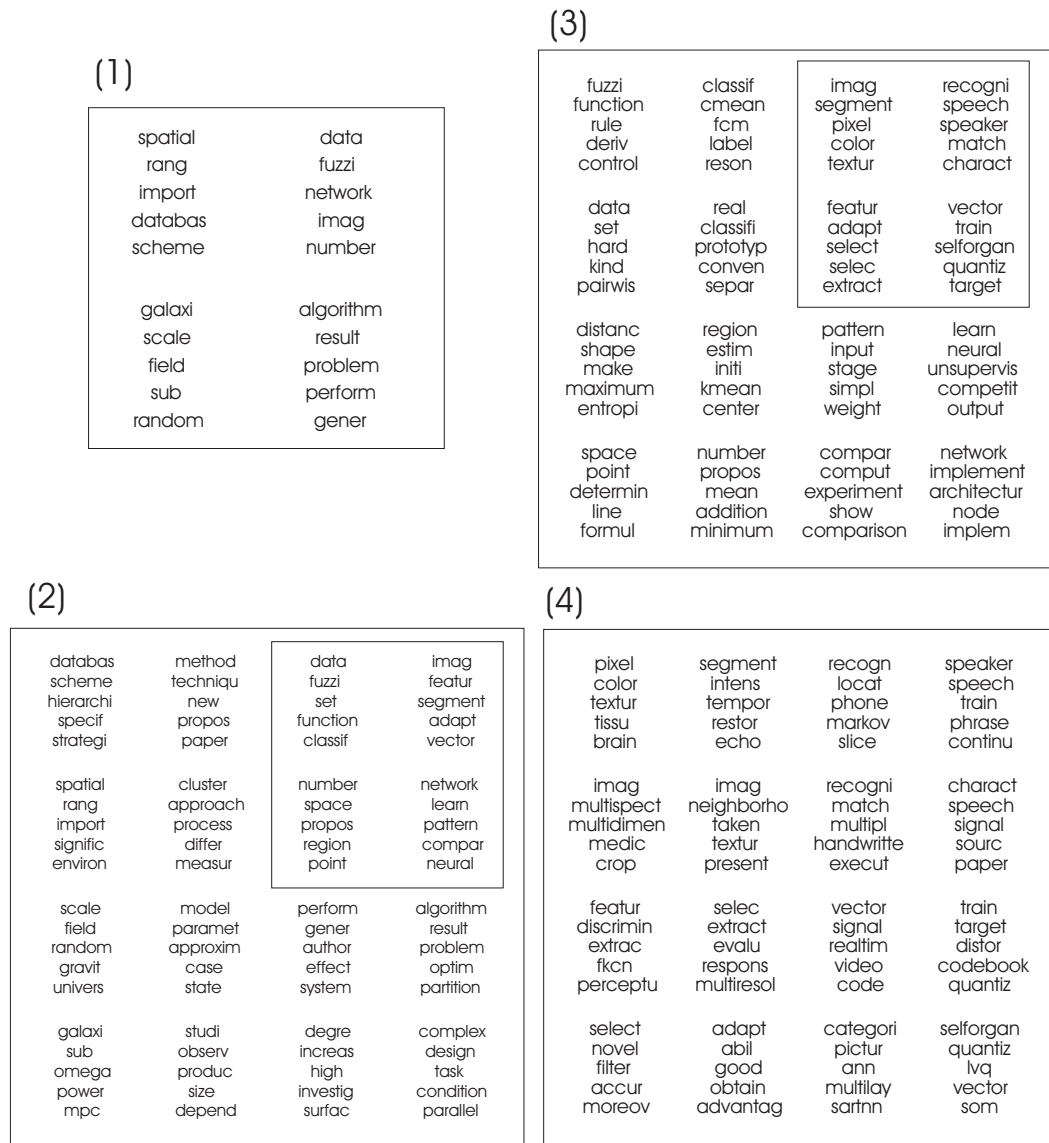


Fig. 4. Multi-resolution visualization of the CLUSTER collection with grid maps at  $2 \times 2$  (1),  $4 \times 4$  (2),  $8 \times 8$  (3), and  $16 \times 16$  (4). Subfigure (3) shows the  $4 \times 4$  subgrid obtained by zooming the marked  $2 \times 2$  window in subfigure (2). Similarly, subfigure (4) is a zoomed-in version of the marked window in subfigure (3).

## 4 Experimental Results

There are essentially two ways to evaluate a visualization method like the ProbMap: (i) by a quantitative evaluation of the quality of the underlying statistical model in terms of log-likelihood, perplexity or task-specific measures like precision-recall, and (ii) by assessing the generated visual map representations. Careful experimental evaluations of the first type that have consistently demonstrated the advantages of PLSA can be found in [8,9,6]. Here, we restrict

points game round lead final	seven thirty second san hit	coast florida guard miles sea	area plane damage emergency rescue	police building injured car dead	north south korea shot korean	killed attack sources bomb violence	israel israeli palestinian west peace
team players series strike owners	fourty went lost took left	water board base cubans center	air airport small safety spokeswoman	city spokesman local authorities officials	sunday radio saturday arrested released	town refugees people prisoners men	army soldiers peace border east
television best film smith love	years home york long young	fifty british service died old	thousand number department total available	said officials reported report news	statement group office german reports	government capital return thousands civil	troops military forces fighting operation
year-old prison woman wife county	children death life family hospital	times drug school known university	million eighty food company industry	nineteen ninety newspaper march april	official saying quoted interview central	end ministry embassy western south	war russian russia region yeltsin
court case trial judge charges	justice blood black legal dr	women use medical cases study	dollar business cost paid environmental	year private commission bank ban	rights human japan japanese china	minister foreign prime visit relations	nuclear talks europe soviet cooperation
los simpson angeles defense jury	asked information investigation letter question	law help public federal need	percent money pay services increase	billion economic economy aid development	trade world conference deal countries	agreement meeting agreed meet negotiations	french france paris croatia contact
live today reporting yesterday jim	right know people believe point	work americans american health care	plan program tax cut cuts	state national support speech action	policy administration free american step	united states nations international washington	un bosnian serbs serb bosnia
p joining space john gary	good come course little lot	house clinton president white gingrich	senate congress republican vote committee	party political elections campaign election	haiti general power haitian president	council security force sanctions canada	iraq iraqi arms northern kuwait

Fig. 5. Coarsened  $8 \times 8$  map of the TDT1 collection derived from a  $32 \times 32$  ProbMap with cyclic boundary conditions.

our attention to the visualization aspect.

We have utilized two document collections in our experiments: (i) the TDT1 collection (Topic Detection and Tracking, distributed by the *Linguistic Data Consortium* [15]) with 15,863 transcribed broadcast news stories, (ii) a collection of 1,568 abstract of research papers on ‘clustering’ (CLUSTER). All texts have been preprocessed with a stop word list, in addition very infrequent words with less than three occurrences have been eliminated. For the CLUSTER collection words have been reduced to a root form with a standard stemmer, for the TDT1 collection word frequencies have been weighted with an entropic term weight [20].

In all our experiments, we have visualized topics/events by the 5 most probable words of the corresponding factor (*i.e.*, words  $w^j$  for which  $\phi_j^k$  is maximal). These words have been utilized to produce topic maps by displaying them at positions  $(x, y)$  corresponding to the index  $k(x, y)$  on the two-dimensional grid. In an interactive setting one could imagine varying the number of displayed terms according to the user’s preferences.

A pyramidal visualization of the CLUSTER collection based on a 256 factor,

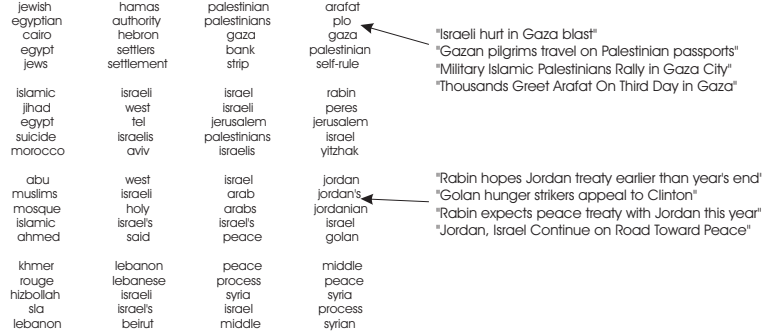


Fig. 6. Portion of the TDT1 ProbMap at full ( $32 \times 32$ ) resolution. The displayed part is an enlargement of the upper right corner in Fig. 5. Attached with arrows are titles of documents  $d^i$  with maximal weights  $\tau_k^i$  for these topics.

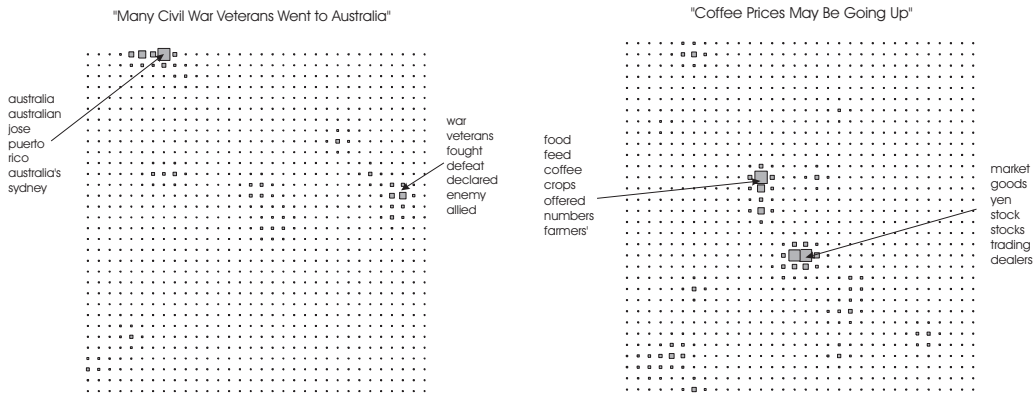


Fig. 7. Visualization of the topic weights  $\tau_k^i$  for two documents. The title of the two documents is displayed at the top; the map displays at each grid location  $(x, y)$  a square with area proportional to  $\tau_{k(x,y)}^i$  and has keywords attached to the two most significant local maxima.

$16 \times 16$  ProbMap is depicted in Fig. 4. One can see that meaningful coarsened maps have been obtained from the  $16 \times 16$  map: different areas like astronomy, physics, databases, and pattern recognition can be easily identified. In particular on the finer levels, the topological organization is very helpful where the relation of different subtopics in signal processing, including image processing and speech recognition, is well-preserved by the topic map.

A  $8 \times 8$  ProbMap of the TDT1 collection is depicted in Fig. 5. The map provides a good overview of the news stories contained in this collection. In Fig. 6 the upper right corner of Fig. 5 has been enlarged. The represented factors obviously deal with various events/topics about Israel and the Arabic world. The topological arrangement has successfully positioned these factors close to one another on the grid. The attached exemplary document titles provide additional information and confirm the characterization by keywords.

The information contained in a ProbMap can be used in various other ways. In Fig. 7 the parameters  $\tau_k^i$  have been displayed for two example documents.

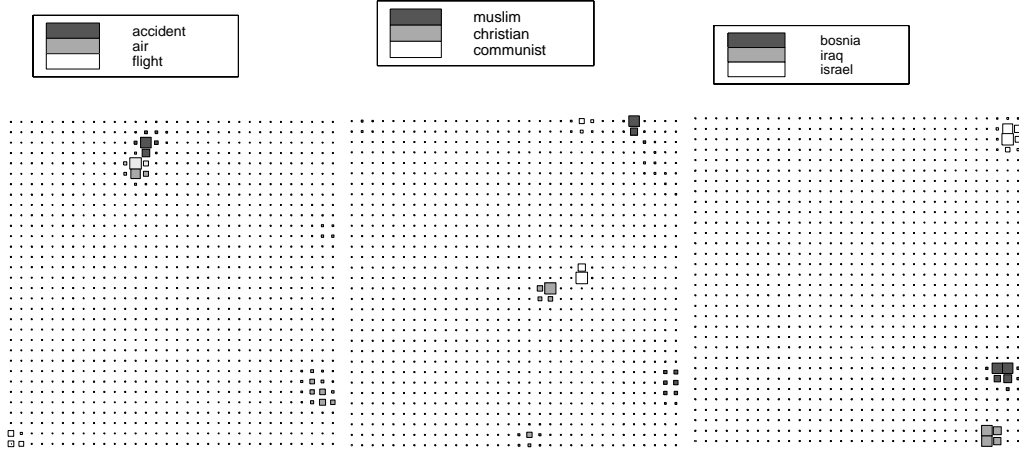


Fig. 8. Visualization of three probability maps for keywords. The legend at the top explains which words have been visualized; the map for  $w^j$  displays at each grid location  $(x, y)$  a square with area proportional to the normalized weight  $\phi_j^{k(x,y)}$ .

Both cases illustrate the advantages of a topic decomposition technique like PLSA: each document can be represented as a combination (distribution) of topics and does not have to be assigned to a single location on the map as in the WebSOM. For example, the document with the title ‘Coffee Prices Maybe Going Up’ deals with agricultural goods (‘food’, ‘crops’, ‘coffee’), but is also related to trading (‘market’, ‘trading’).

In yet another visualization mode one can display the distribution of words over the various topics, *i.e.*, by applying Bayes’ rule to parameters  $\phi_j^k$  for a fixed word  $w^j$ . This may support information access where the user specifies keywords and wants to see where on the map corresponding topics can be found. Fig. 8 displays maps of this type for three different examples with three keywords each. The displayed maps also illustrate how words with multiple meanings or different types of usage are represented. ‘flight’ is used in the context of planes, but also for space flight; ‘air’ occurs in the context of civil aviation, but also in the context of air raids and bombing campaigns, etc. The examples show that these word distributions are typically sparse, but often multimodal.

## 5 Conclusion

We have presented a novel probabilistic technique for visualizing text databases by *topic maps*, called the *ProbMap*. The main advantages are (i) the sound statistical foundation on a latent class model with EM as a fitting procedure, (ii) the principled combination of probabilistic modeling and topology-preservation, and (iii) the natural definition of resolution hierarchies. The benefits of this approach to support interactive retrieval have been demonstrated

with simple two-dimensional maps, however, since arbitrary topologies can be extracted, one might expect additional benefits in combination with more elaborate graphical interfaces.

### *Acknowledgments*

I would like to thank the special issue guest editors Michael Berthold, David Hand, and Doug Fisher, as well as the anonymous referees for their helpful comments.

### **References**

- [1] J. M. Buhmann. Stochastic algorithms for data clustering and visualization. In M.I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Publishers, 1998.
- [2] J. M. Buhmann and H. Kühnel. Complexity optimized data clustering by competitive neural networks. *Neural Computation*, 5:75–88, 1993.
- [3] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B*, 39:1–38, 1977.
- [5] D. Geiger, D. Heckerman, H. King, and C. Meek. Stratified exponential families: Graphical models and model selection. Technical Report MSR-TR-98-31, Microsoft Research, July 1998.
- [6] D. Gildea and T. Hofmann. Topic based language models using EM. In *Proceedings of 6th European Conference On Speech Communication and Technology (Eurospeech)*, volume 5, pages 2167–2170, 1999.
- [7] J. G. Herder. *Sprachphilosophische Schriften*. Felix Meiner Verlag, Hamburg, 1960.
- [8] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in AI*, pages 289–296, 1999.
- [9] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval, Berkeley, California*, pages 50–57, 1999.
- [10] T. Hofmann and J. M. Buhmann. Competitive learning algorithms for robust vector quantization. *IEEE Transaction on Signal Processing*, 46(6):1665–1675, 1998.

- [11] T. Hofmann, J. Puzicha, and M. I. Jordan. Unsupervised learning from dyadic data. In *Advances in Neural Information Processing Systems*, volume 11, pages 466–472. MIT Press, 1999.
- [12] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM–self-organizing maps of document collections. *Neurocomputing*, 21:101–117, 1998.
- [13] T. Kohonen. *Self-Organization and Associative Memory*. Springer, 1984.
- [14] T. Kohonen. *Self-Organizing Maps*. Springer, 1995.
- [15] Linguistic Data Consortium. TDT pilot study corpus. Catalog no. LDC98T25, 1998.
- [16] S. P. Luttrell. Hierarchical vector quantization. *IEEE Proceedings*, 136:405–413, 1989.
- [17] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proceedings of the ACL*, pages 183–190, 1993.
- [18] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989.
- [19] H. Ritter and K. Schulten. Kohonen’s self-organizing maps: exploring their computational capabilities. In *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, pages 109–116, 1988.
- [20] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw–Hill, 1983.
- [21] L. Saul and F. Pereira. Aggregate and mixed–order Markov models for statistical language processing. In *Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing*, pages 81–89, 1997.