Probabilistic Topic Maps: Navigating through Large Text Collections

Thomas Hofmann

Computer Science Division, UC Berkeley &
International CS Institute, Berkeley, CA
hofmann@cs.berkeley.edu

Abstract. The visualization of large text databases and document collections is an important step towards more flexible and interactive types of information retrieval. This paper presents a probabilistic approach which combines a statistical, model–based analysis 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.

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'[5] – 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, since retrieving information from a glut of nuisance data can be tedious and extremely time consuming.

What is urgently needed are navigation aids, overlooks which offer uncomplicated and fast visual access to information, and maps that provide orientation, possibly on different level of resolution and abstraction. This paper deals with a statistical approach to provide such overlooks and maps for large collections of text documents. It aims at a concise visualization of conceptual and topical similarities between documents or aspects of documents in the form of *topic maps*. The proposed method has two building blocks:

i. A *latent semantic analysis* technique for text collections [3, 6] which models context–dependent word occurrences.
ii. A principle of *topology preservation* [11] which allows to visualize the extracted information, for example, in the form of a two–dimensional map.

D.J. Hand, J.N. Kok, M.R. Berthold (Eds.): IDA'99, LNCS 1642, pp. 161–172, 1999.

Herein, data analysis and visualization are not treated as separate procedural stages; as we will discuss in more detail later on, it is a benefit of our procedure that it unites both problems. This is formally achieved by optimizing a single objective function which combines a statistical criterion with topological constraints to ensure visualization. This coupling makes sense, whenever the final end is not the analysis per se, but the presentation and visualization of regularities and patterns extracted from data 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 and allow an effortless interpretation. 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.

Our approach is somewhat related in spirit to the WEBSOM learning architecture [10] which continues earlier work on semantic maps [15] and performs a topological clustering of words represented as context–vectors. However, the method presented here is based on a strictly probabilistic data model which is fitted by maximum likelihood estimation. The discrete nature of words is directly taken into account without deviation via a (randomized) vector space representation as in the WEBSOM. In addition, our model does not perform word *clustering*, but models topics via word *distributions*.

The rest of the paper is organized as follows: Section 2 briefly introduces a probabilistic method for latent semantic analysis [6], which is then extended to incorporate topological constraints in Section 3. Finally, Section 4 shows some exemplary results of multi-resolution maps extracted from document collections.

2 Probabilistic Latent Semantic Analysis

2.1 Data Representation

Probabilistic Latent Semantic Analysis (PLSA) [6, 7] is a general method for statistical factor analysis of two-mode and count data which we apply here to learning from document collections. Formally, text collections are represented as pairs over a set of documents $\mathcal{D} = \{d_1, \ldots, d_N\}$ and a set of words $\mathcal{W} = \{w_1, \ldots, w_M\}$, i.e, the elementary observations we consider are of the form (d, w), denoting the occurrence of a word w in a document d. Summarizing all observations by counts $n(d, w)$ of how often a word occurred in a document, one obtains a rectangular N by M matrix $\mathbf{N} = [n(d_i, w_j)]_{i,j}$ which is usually referred to as *term–document matrix*. The key assumption of this representation is the so–called 'bag-of-words' view which presupposes that conditioned on the identity of a particular document, word occurrences are statistically independent. This also the basis for the popular *vector-space model* of documents [16] 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, because just a small part of the vocabulary is typically used in a single document. 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*. 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, i.e, based exclusively on a document collection or corpus at hand. Given these expectations could be met, PLSA would offers great advantages in terms of flexibility as well as in terms of domain adaptivity.

2.2 Probabilistic Latent Semantic Analysis

PLSA is based on a latent class model which associates an unobserved class variable $z \in \mathcal{Z} = \{z_1, \ldots, z_K\}$ with each observation (d, w). As will be explained in more detail, the intention pursued by introducing latent variables is to model *text topics* such that each possible state $z \in \mathcal{Z}$ would ideally represent one particular topic or subject. Formally, let us define the following multinomial distributions: $P(d)$ is used to denote the probability that a word is observed in a particular document.[1] $P(w|z)$ denotes a word distributions conditioned on the latent class variable z, which represent different *topic factors*. Finally, $P(z|d)$ is used to denote document-specific distributions over the latent variable space $\mathcal{Z}$. We may now define the following probabilistic model over $\mathcal{D} \times \mathcal{W}$

$$P(d, w) = P(d)P(w|d), \quad \text{where } P(w|d) = \sum_{z \in \mathcal{Z}} P(w|z)P(z|d) \ . \tag{1}$$

This model is based on a crucial conditional independence assumption, namely that d and w are independent conditioned on the state of the latent variable z associated with the observation (d, w). As a result, the conditional distributions $P(w|d)$ in (1) are represented as convex combinations of the K factors $P(w|z)$. Since in the typical case one has $K \ll N$, the latent variable z can be thought of as a bottleneck variable in predicting words conditioned on documents.

To demonstrate how this corresponds to a mixture decomposition of the term–document matrix, we switch to an alternative parameterization by applying Bayes' rule to $P(z|d)$ and arriving at

$$P(d, w) = \sum_{z \in \mathcal{Z}} P(z)P(d|z)P(w|z) \,, \tag{2}$$

which is perfectly symmetric in both entities documents and words. Based on (2) let us formulate the probability model (1) in matrix notation, by defining $\mathbf{U} = [P(d_i|z_k)]_{i,k}$, $\mathbf{V} = [P(w_j|z_k)]_{j,k}$, $\Sigma = \text{diag}[P(z_k)]_k$, so that $\mathbf{P} = [P(d_i, w_j)]_{i,j} =$

[1] This is intended to account for varying document lengths.

$\mathbf{U}\Sigma\mathbf{V}^t$. The algebraic form of this decomposition corresponds exactly to the decomposition of $\mathbf{N}$ obtained by *Singular Value Decomposition* (SVD) in standard *Latent Semantic Analysis* (LSA) [3]. However, the statistical model fitting principle used in conjunction with PLSA is the likelihood principle, while LSA is based on the Frobenius or L_2–norm of matrices. The statistical approach offers important advantages since it explicitly aims at minimizing word perplexity[2]. The mixture approximation $\mathbf{P}$ of the co-occurrence table is a well-defined probability distribution and factors have a clear probabilistic meaning in terms of mixture component distributions. In contrast, LSA does not define a properly normalized probability distribution and the obtained approximation may even contain negative entries. 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. [6]). Last but not least, the statistical formulation can be systematically extended and generalized in various ways, an example being the model presented in Section 3 of this paper.

2.3 EM Algorithm for PLSA

In order fit the model in (1) we follow the statistical standard procedure and perform maximum likelihood estimation with the EM algorithm [4, 17]. One has to maximize

$$\mathcal{L} = \sum_{d\in\mathcal{D}} \sum_{w\in\mathcal{W}} n(d,w) \log P(d,w) \tag{3}$$

with respect to all multinomial distributions which define $P(d,w)$. EM is guaranteed to find a local maximum of $\mathcal{L}$ by alternating two steps: (i) an expectation (E) step where posterior probabilities for the latent variables are computed based on the current estimates of the parameters, (ii) a maximization (M) step, where parameters are updated based on the posterior probabilities computed in the E–step. For the E–step one simply applies Bayes' formula, e.g., in the parameterization of (1), to obtain

$$P(z|d,w) = \frac{P(z|d)P(w|z)}{\sum_{z'\in\mathcal{Z}} P(z'|d)P(w|z')} \,. \tag{4}$$

It is straightforward to derive the M–step equations [9]

$$P(w|z) \propto \sum_{d\in\mathcal{D}} n(d,w)P(z|d,w), \quad P(z|d) \propto \sum_{w\in\mathcal{W}} n(d,w)P(z|d,w)\,. \tag{5}$$

The estimation of $P(d) \propto \sum_w n(d,w)$ can be carried out independently. Alternating (4) and (5) initialized from randomized starting conditions results in a procedure which converges to a local maximum of the log–likelihood in (3).

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

Fig. 1. The 3 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 most probable in the class-conditional distribution $P(w|z)$.

2.4 Example: Analysis of Word Usage with PLSA

Let us briefly discuss an elucidating example application of PLSA at this point. We have run PLSA with 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 chosen the word ‘segment’. Figure 1 shows the most probable words of 3 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” 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 Figure 2. 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 the Rwandan genocide. These example shows that the topic identified by PLSA might also correspond to something one might more appropriately refer to as *events*. De-

[2] Perplexity is a term from statistical language modeling which is utilized here to refer to the (log-averaged) inverse predictive probability $1/P(w|d)$.

“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. Three factors to most likely generate the word “UN” from a 128 factor decomposition of the TDT1 corpus.

pendent on the training collection and the specific domain the notion of topic has thus to be taken in a broader sense.

2.5 PLSA: What Is Missing?

From the example in Figure 1 one can see that the factors $P(w|z)$ extracted 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 the user to 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 nothing more than an ad hoc remedy to the problem.

What is really missing in PLSA as presented so far, is a relationship between the different factors. Suppose for concreteness one had identified a relevant topic represented by some $P(w|z)$; the identity of z does not provide any information about whether or not another topic $P(w|z')$ could be relevant as well. The generalization we present in the following section, extends the PLSA model in a way that enables it to captures 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

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) [11, 12]. While the formulation of the algorithm in [11] 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 relations to vector quantization for noisy communication channels (cf. [14, 2, 8]). 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.

3.1 Topologies from Confusion Probabilities

The key step in the proposed generalization is to introduce an additional latent variable $v \in \mathcal{Z}$ of the same cardinality as z to define the probability model

$$P(d,w) = P(d)P(w|d), \quad P(w|d) = \sum_{z \in \mathcal{Z}} P(w|z) \sum_{v \in \mathcal{Z}} P(z|v)P(v|d)\,. \qquad (6)$$

It is straightforward to verify that from a purely statistical point of view this does not offers any additional modeling power. Whatever the choice for $P(z|v)$ and $P(v|d)$ might be, one can simply define $P(z|d) = \sum_v P(z|v)P(v|d)$ to obtain exactly the same distribution over $\mathcal{D} \times \mathcal{W}$ in the more parsimonious model of (1). Yet, we do *not* propose to fit all model parameters in (6) from training data, but to fix the *confusion probabilities*[3] $P(z|v)$ 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 $P(w|z)$ on a two–dimensional grid with boundaries. Let us introduce the notation $z(x,y)$, $1 \leq x, y \leq L$, $x, y \in \mathbb{N}$ to identify latent states $z(x,y) \in \mathcal{Z}$ with points (x,y) on the grid. By the Euclidean metric, this embedding induces a distance function on $\mathcal{Z}$, namely

$$d(z(x,y), z(x',y')) = d((x,y),(x',y')) = \sqrt{(x-x')^2 + (y-y')^2}. \qquad (7)$$

Now we propose to define $P(z|v)$ via a Gaussian with standard deviation σ

$$P(z|v) = \frac{\exp\left[-d(z,v)^2/(2\sigma^2)\right]}{\sum_{z'} \exp\left[-d(z',v)^2/(2\sigma^2)\right]}\,, \qquad (8)$$

where σ is assumed to be fixed for now. To understand why this favors a topological organization of topics, consider a document d with its topic distribution $P(v|t)$. The confusion probabilities tilt this distribution to a distribution

[3] We use this terminology, because the relationship between z and v can be thought of in terms of a communication scenario: v represents the original message and z the message received after sending it via a noisy channel. $P(z|v)$ then correspond to the channel characteristic, i.e., how probable it is to receive z after sending v.

$P(z|d) = \sum_z P(z|v)P(v|d)$. For simplicity assume that $P(v|d) = 1$ for a particular $v \in \mathcal{Z}$, then the confusion probabilities will blend-in additional contributions mainly from neighboring states z of v on the two–dimensional grid. If these neighboring states represent very different topics, the resulting word distribution $P(w|d)$ in (6) will significantly deviate from the distribution one would get from (1), which – assuming that $P(v|d)$ was chosen optimal – will result in a poor estimate. If on the other hand the neighbors of v 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 word perplexity.

3.2 EM Algorithm for Topological PLSA

The next step consists in deriving the EM equations for topological PLSA. Standard calculations yield the M–step re-estimation formulae

$$P(w|z) \propto \sum_d n(d,w)P(z|d,w), \text{ and } P(v|d) \propto \sum_w n(d,w)P(v|d,w)\,. \quad (9)$$

For the evaluation of (9) the marginal posterior probabilities are sufficient and it is not necessary to compute the joint posterior $P(v,z|d,w)$. The marginal posterior probabilities are given by

$$P(v|d,w) = \sum_z P(v,z|d,w) = \frac{P(v|d)P(w|v)}{\sum_{v'} P(v'|d)P(w|v')}, \quad \text{and} \quad (10)$$

$$P(z|d,w) = \sum_v P(v,z|d,w) = \frac{P(z|d)P(w|z)}{\sum_{z'} P(z'|d)P(w|z')}, \quad (11)$$

where $P(w|v) = \sum_z P(w|z)P(z|v)$ and $P(z|d) = \sum_v P(z|v)P(v|d)$. Notice also that the marginal posteriors are simply related by

$$P(v|d,w) = \sum_z P(v|z)P(z|d,w), \quad P(z|d,w) = \sum_v P(z|v)P(v|d,w)\,. \quad (12)$$

In summary, one observes that the EM algorithm for topological PLSA requires the computation of marginal posteriors and document/word conditionals for both variables v and z. Moreover, these quantities are related by a simple matrix multiplication with the confusion matrix $[P(z_k|v_l)]_{k,l}$ or its counterpart $[P(v_k|z_l)]_{k,l}$.

3.3 Topologies and Hierarchies

There are two ways in which hierarchies are of interest in the context of topological PLSA: (i) To accelerate the PLSA by a multi-resolution optimization over a sequence of coarsened 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 2×2 grid, and then recursively prolongating the found solution according to an quadtree–like scheme. This involves

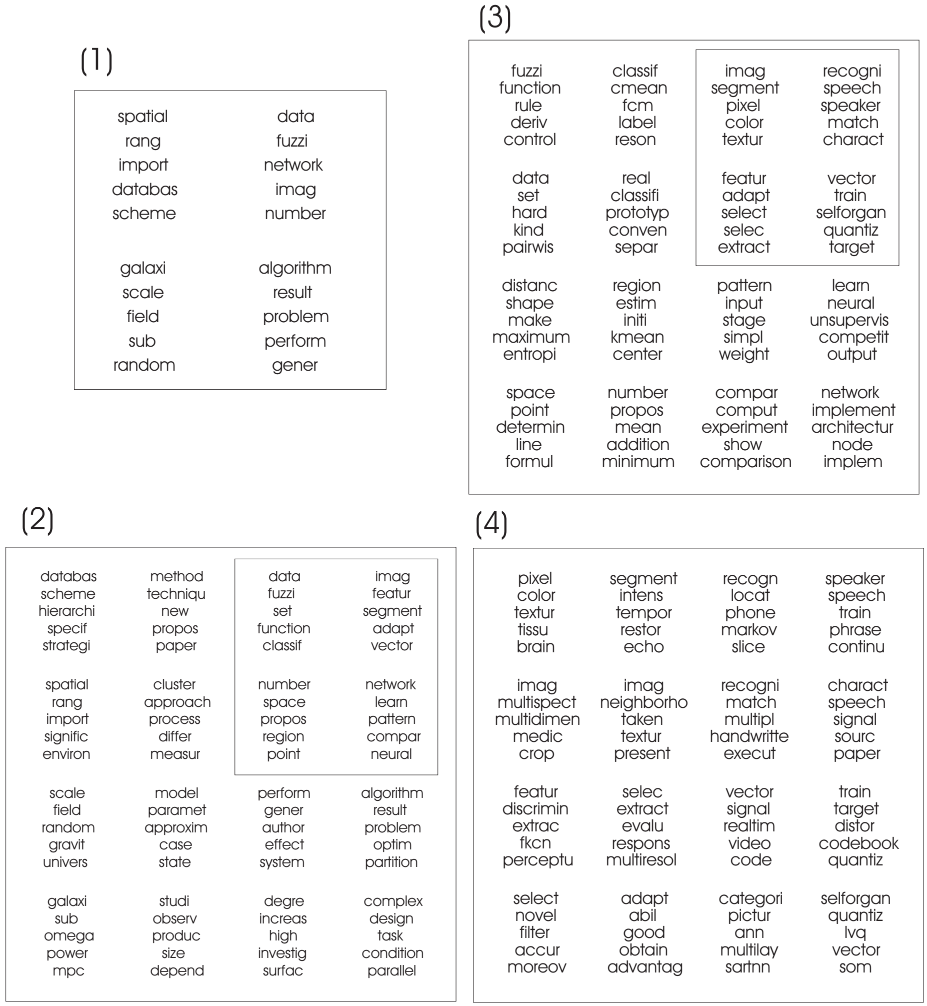

(1)
spatial rang import databas scheme
data fuzzi network imag number
galaxi scale field sub random
algorithm result problem perform gener
(2)
databas scheme hierarchi specif strategi
method techniqu new propos paper
data fuzzi set function classif
imag featur segment adapt vector
spatial rang import signific environ
cluster approach process differ measur
number space propos region point
network learn pattern compar neural
scale field random gravit univers
model paramet approxim case state
perform gener author effect system
algorithm result problem optim partition
galaxi sub omega power mpc
studi observ produc size depend
degre increas high investig surfac
complex design task condition parallel
(3)
fuzzi function rule deriv control
classif cmean fcm label reson
imag segment pixel color textur
recogni speech speaker match charact
data set hard kind pairwis
real classifi prototyp conven separ
featur adapt select selec extract
vector train selforgan quantiz target
distanc shape make maximum entropi
region estim initi kmean center
pattern input stage simpl weight
learn neural unsupervis competit output
space point determin line formul
number propos mean addition minimum
compar comput experiment show comparison
network implement architectur node implem
(4)
pixel color textur tissu brain
segment intens tempor restor echo
recogn locat phone markov slice
speaker speech train phrase continu
imag multispect multidimen medic crop
imag neighborho taken textur present
recogni match multipl handwritte execut
charact speech signal sourc paper
featur discrimin extrac fkcn perceptu
selec extract evalu respons multiresol
vector signal realtim video code
train target distor codebook quantiz
select novel filter accur moreov
adapt abil good obtain advantag
categori pictur ann multilay sartnn
selforgan quantiz lvq vector som

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

copying the distributions $P(w|z)$ – with a small random disturbance – to the successors of z on the finer grid and distributing $P(v|d)$ 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'.[4] The multi-resolution optimization is coupled with a schedule for σ, which defines the length-scale for the confusion probabilities in (8). 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 $\sigma_{n+m} = (1/2)\sigma_m$. Prolongations to a finer grid is performed at iterations $n = m, 2m, 3m, \ldots$.

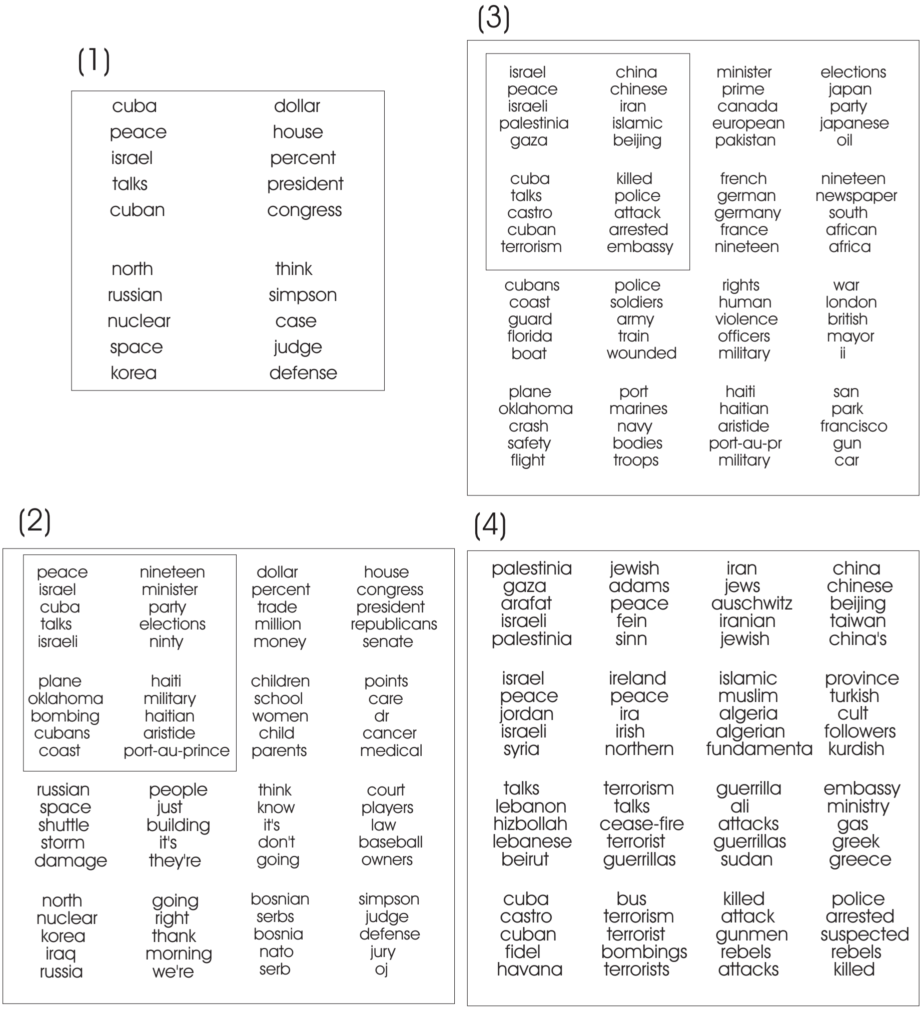
Notice that the topological organization of topics has the further advantage to support a simple coarsening procedure for visualization at different resolution levels. The fact that neighboring latent states represent similar topics suggests to merge states, e.g., four at a time, to generate a coarser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

4 Experimental Results

We have utilized two document collections for our experiments: (i) the TDT1 collection (Topic Detection and Tracking, distributed by the *Linguistic Data Consortium* [13]) with 49,225 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 3 occurrences have also been eliminated. For the TDT1 collection word frequencies have been weighted with an entropic term weight [16]. The 5 most probable words in factors $P(w|z)$ have been utilized for visualization and are displayed at the position corresponding to the topic on the two–dimensional grid to produce topic maps. In an interactive setting one would of course vary the number of displayed terms according to the user's preferences.

A pyramidal visualization of the CLUSTER collection based on a 256 factor, 16×16 topological PLSA is depicted in Figure 3. One can see that a meaningful coarsened maps can be obtained from the 16×16 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 recogniton, is well–preserved by the topic map. A similar map hierarchy for the TDT1 collection is depicted in Figure 4. Different topics and events can effortlessly be identified from the word distributions. Again, subtopics like the ones dealing with different events of international politics are mapped to neighboring positions on the lattice.

[4] There is a large body of literature dealing with the topology–preserving properties of SOMs. The reader is referred to [12] and the references therein.



(1)

cuba peace israel talks cuban	dollar house percent president congress
north russian nuclear space korea	think simpson case judge defense

(2)

peace israel cuba talks israeli	nineteen minister party elections ninty	dollar percent trade million money	house congress president republicans senate
plane oklahoma bombing cubans coast	haiti military haitian aristide port-au-prince	children school women child parents	points care dr cancer medical
russian space shuttle storm damage	people just building it's they're	think know it's don't going	court players law baseball owners
north nuclear korea iraq russia	going right thank morning we're	bosnian serbs bosnia nato serb	simpson judge defense jury oj

(3)

israel peace israeli palestinia gaza	china chinese iran islamic beijing	minister prime canada european pakistan	elections japan party japanese oil
cuba talks castro cuban terrorism	killed police attack arrested embassy	french german germany france nineteen	nineteen newspaper south african africa
cubans coast guard florida boat	police soldiers army train wounded	rights human violence officers military	war london british mayor ii
plane oklahoma crash safety flight	port marines navy bodies troops	haiti haitian aristide port-au-pr military	san park francisco gun car

(4)

palestinia gaza arafat israeli palestinia	jewish adams peace fein sinn	iran jews auschwitz iranian jewish	china chinese beijing taiwan china's
israel peace jordan israeli syria	ireland peace ira irish northern	islamic muslim algeria algerian fundamenta	province turkish cult followers kurdish
talks lebanon hizbollah lebanese beirut	terrorism talks cease-fire terrorist guerrillas	guerrilla ali attacks guerrillas sudan	embassy ministry gas greek greece
cuba castro cuban fidel havana	bus terrorism terrorist bombings terrorists	killed attack gunmen rebels attacks	police arrested suspected rebels killed

Fig. 4. Multi-resolution visualization of the TDT1 collection with grid maps at 2×2 $4\times$, 8×8 (upper left corner), and 16 (upper left corner). Subfigure (3) shows the 4×4 subgrid obtained by zooming the marked 2×2 window in subfigure (2). Similarly, subfigure (4) is a zoomed-in version of the marked window in subfigure (3).

5 Conclusion

We have presented a novel probabilistic technique for visualizing text databases by *topic maps*. The main advantages are (i) a 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 briefly with simple two–dimensional maps, however, since arbitrary topologies can be extracted, one might expect even more benefits in combination with more elaborate interfaces.

Acknowledgment

This work has been supported by a DAAD postdoctoral fellowship.

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, G. W. Dumais, S. T. amd Furnas, Landauer. T. K., and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 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] J. G. Herder. *Sprachphilosophische Schriften.* Felix Meiner Verlag, Hamburg, 1960.

[6] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the 15th Conference on Uncertainty in AI*, 1999.

[7] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd ACM-SIGIR International Conference on Research and Development in Information Retrieval, Berkeley, California*, 1999.

[8] T. Hofmann and J. M. Buhmann. Competitive learning algorithms for robust vector quantization. *IEEE Transaction on Signal Processing*, 46(6):1665–1675, 1998.

[9] T. Hofmann and J. Puzicha. Statistical models for co–occurrence data. Technical report, AI Memo 1625, M.I.T., 1998.

[10] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. WEBSOM–self-organizing maps of document collections. *Neurocomputing*, 21:101–117, 1998.

[11] T. Kohonen. *Self–organization and Associative Memory.* Springer, 1984.

[12] T. Kohonen. *Self-Organizing Maps.* Springer, 1995.

[13] Linguistic Data Consortium. TDT pilot study corpus. Catalog no. LDC98T25, 1998.

[14] S.P. Luttrell. Hierarchical vector quantization. *IEE Proceedings*, 136:405–413, 1989.

[15] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cyberbetics*, 61:241–254, 1989.

[16] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw–Hill, 1983.

[17] 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*, 1997.