

Conditional Information Bottleneck Clustering

David Gondek
Department of Computer Science
Brown University
dcg@cs.brown.edu

Thomas Hofmann
Department of Computer Science
Brown University
th@cs.brown.edu

Abstract

We present an extension of the well-known information bottleneck framework, called conditional information bottleneck, which takes negative relevance information into account by maximizing a conditional mutual information score. This general approach can be utilized in a data mining context to extract relevant information that is at the same time novel relative to known properties or structures of the data. We present possible applications of the conditional information bottleneck in information retrieval and text mining for recovering non-redundant clustering solutions, including experimental results on the WebKB data set which validate the approach.

1 Introduction

The information bottleneck method [11] has introduced a new principle for extracting relevant structure from data which is more general than, for instance, the standard supervised classification paradigm. The idea is to model structure extraction as data compression and to quantify the relevance of the extracted structure by how much information it preserves about a specified relevance variable. In contrast to example based methods, relevance can hence be defined in a task-specific manner on a conceptual level and can be built into the model, i.e. it needs not to be learned from data. The remaining computational task is to find the optimal, maximally information preserving data reduction. It is a natural extension of this framework to define the notion of *conditional relevance*. The latter refers to a situation where certain aspects of the data are assumed to be known *a priori* and the goal is to extract structure that provides not only relevant information, but also *novel* information, which is not implied by the known aspects.

This idea has been first formulated in [2] and [6] where the proposed solutions involve taking side information into account by treating it as *irrelevant information*. Formally [2] thus propose to maximize the information about the rel-

evance variable while at the same time *minimizing* the information between the extracted structure and the irrelevant noise. In this paper we follow a similar goal, but deviate from [2] in one important aspect: we propose to maximize the *conditional information* about the relevant structure, given the side information. As we will argue this formulation is conceptually preferable and alleviates from having to specify a weight to balance the trade-off between relevance and irrelevance.

We have also identified a new application for the conditional bottleneck method, which is to find clustering solutions that are “orthogonal” to a known categorization scheme. This is of great interest, for example in the context of document clustering, where many competing ways to cluster data may exist. One thus is often interested in discovering structure that is not already entailed by some known categorization scheme.

2 Information Bottleneck Reloaded

We will use the following notation: x refers to objects such as documents, y to features that are considered relevant, e.g. words, and c to clusters of objects. Upper case letters X, Y, C are used to denote the corresponding random variables. The information bottleneck criterion [11] proposes to find a probabilistic clustering, parameterized by probabilities $\{p(c|x)\}$ that minimizes

$$F = I(X, C) - \beta I(Y, C) \quad (1)$$

for some prespecified $\beta > 0$, which balances the trade-off between compression and preservation of information about the features of interest.

We first present a novel (and simple) way to derive the information bottleneck update equations in a variational setting. As a first step, we rewrite the mutual informations as (conditional) entropy differences

$$F = H(C) - H(C|X) - \beta H(Y) + \beta H(Y|C) \quad (2)$$

Notice that $H(Y)$ is a constant that can be safely omitted. We introduce cross entropies with auxiliary parameters $q(c)$

and $q(y|c)$, non-negative and normalized, i.e. $\sum_c q(c) = 1$, $\sum_y q(y|c) = 1$ for all c . Now we define

$$\tilde{H}(C) \equiv - \sum_c p(c) \log q(c) \quad (3a)$$

$$\tilde{H}(Y|C) \equiv - \sum_{c,y} p(y,c) \log q(y|c) \quad (3b)$$

and a new objective

$$\tilde{F} = \tilde{H}(C) - H(C|X) + \beta \tilde{H}(Y|C). \quad (4)$$

The advantage of \tilde{F} is that it is convex in $p(c|x)$ for given q , since $-H(C|X)$ is convex and the \tilde{H} terms are linear in $p(c|x)$. Moreover, \tilde{F} is also convex in the q parameters for given $p(c|x)$, since the negative logarithm function is convex. More precisely the solutions over q can be obtained as follows: First observe that the only term in \tilde{F} that depends on $q(c)$ is $\tilde{H}(C)$ and the only term that depends on $q(y|c)$ is $\tilde{H}(Y|C)$. Differentiating \tilde{F} with respect to the q parameters and setting to zero results in $q(c) = p(c)$ and $q(y|c) = p(y|c)$, as is straightforward to prove (cf. [11], Lemma 2). These correspond to minima of the function \tilde{F} for given variational parameters $p(c|x)$. Finally, in order to optimize \tilde{F} explicitly over the variational parameters one makes use of the relation

$$\frac{\partial p(y,c)}{\partial p(c|x)} = \frac{\partial \sum_x p(y|x)p(x)p(c|x)}{\partial p(c|x)} = p(y|x)p(x), \quad (5)$$

and arrives at

$$\frac{1}{p(x)} \frac{\partial \tilde{F}}{\partial p(c|x)} = [1 + \log q(c|x)] - \log p(c) - \beta \sum_y p(y|x) \log q(y|c). \quad (6)$$

Setting to zero and accounting for the normalization of $p(c|x)$ one gets

$$p(c|x) \propto q(c) e^{\beta \sum_y p(y|x) \log q(y|c)}. \quad (7)$$

At a solution (corresponding to a minimum of saddle-point of \tilde{F}) we can identify q 's with p 's and Eq. (7) is equivalent to

$$p(c|x) \propto p(c) e^{-\beta D_{\text{KL}}(p(y|x)||p(y|c))} \quad (8)$$

which characterizes the true minimum of F , but potentially other solutions as well.

The asymptotic convergence of this scheme is a simple consequence of the fact that \tilde{F} is reduced in every update iteration and that is bounded from below. This may not correspond to a global minimum, but it will fulfill the optimality conditions over the p and q subspaces separately.

3 Conditional Information Bottleneck

Now we consider a domain with four discrete random variables X, Y, Z , and C . Our goal is again to find a stochastic mapping $p(c|x)$ of items or instantiations x to clusters $c \in \{1, \dots, K\}$. We assume that X, Y and Z have a joint distribution represented by a probability mass function $p(x, y, z) \equiv \Pr\{(X, Y, Z) = (x, y, z)\}$, which can be estimated to a sufficient degree of accuracy from data. As implied by the above notation, C only depends on X , i.e. the cluster membership is independent of Y and Z , given the value of X . We then suggest to solve the following optimization problem with respect to $p(c|x)$,

$$\min_{\{p(c|x)\}} F \equiv I(X, C) - \beta I(Y, C|Z) \quad (9a)$$

$$\text{s.t. } \sum_c p(c|x) = 1, \forall x \text{ and } p(c|x) \geq 0, \forall x, c, \quad (9b)$$

which apparently generalizes the standard information bottleneck. In this objective the weight $\beta > 0$ again controls the trade-off between compression (minimizing $I(X, C)$) and preservation (maximizing $I(Y, C|Z)$). Given a certain compression level or channel capacity, the goal is to preserve as much information about Y in C as possible, provided that Z is revealed to us as side information. Hence, we would like to encode properties of Y in C that cannot yet be reliably inferred based on Z .

For the following derivations it will be useful to again rewrite F in terms of entropies and conditional entropies. Dropping constants, we obtain the equivalent objective function

$$F = H(C) - H(C|X) + \beta H(Y|Z, C). \quad (10)$$

As in the presented information bottleneck derivation, we introduce auxiliary variables $q(c)$ and $q(y|z, c)$ and define

$$\tilde{H}(C) = - \sum_c p(c) \log q(c), \quad (11a)$$

$$\tilde{H}(Y|Z, C) = - \sum_{y,z,c} p(y, z, c) \log q(y|z, c), \quad (11b)$$

and an objective

$$\tilde{F} = -H(C|X) + \tilde{H}(C) + \beta \tilde{H}(Y|Z, C). \quad (12)$$

Using the Lemma in [11], the q parameters which minimize \tilde{F} for given $p(c|x)$ are

$$q(c) = p(c) = \sum_x p(x)p(c|x), \quad (13a)$$

$$q(y|z, c) = p(y|z, c) = \sum_x p(y|x, z)p(x|z, c) \quad (13b)$$

Using the relation

$$\frac{\partial p(y, z, c)}{\partial p(c|x)} = \frac{\partial \sum_x p(x, y, z, c)}{\partial p(c|x)} = p(x, y, z), \quad (14)$$

the optimal values of the variational parameters are given in terms of the auxiliary parameters as

$$p(c|x) \propto q(c) e^{\beta \sum_z p(z|x) \sum_y p(y|x, z) \log q(y|z, c)}. \quad (15)$$

At a stationary point this can again be rewritten more suggestively as a characterization of the resulting joint distribution $p(x, y, z, c)$

$$p(c|x) \propto p(c) e^{-\beta \sum_z p(z|x) D_{\text{KL}}(p(y|x, z) \| p(y|z, c))}. \quad (16)$$

A convergence argument identical to the information bottleneck derivation can be applied.

Compared to [2] which uses the objective $I(X, C) - \beta(I(Y, C) - \gamma I(Z, C))$, the conditional information bottleneck has the advantage that it alleviates of the need to tune the additional trade-off parameter γ . Notice that the latter is problematic, since $I(Y, C)$ and $I(Z, C)$ may live on very different scales, e.g., $I(Y, C)$ may scale with the number of features, while $I(Z, C)$ may scale with the cardinality of the state space of Z . In the presented formulation, this is taken into account by conditioning on the side information Z in $I(Y, C|Z)$, which enforces non-redundancy without the need for an explicit term $I(Z, C)$ to penalize redundancy.

4 Non-Redundant Clustering

We propose to apply the above setting to the problem of data clustering, where a known classification of items is available. The goal then is to cluster the data in a way that is meaningful and at the same time as orthogonal as possible to the given classification.

More specifically, we will focus on the problem of document clustering. Hence \mathcal{X} , $|\mathcal{X}| = n$, will correspond to a document collection and x to the identity of an individual document. A document is represented as a vector of word counts $n(x)$ over a vocabulary of terms $\{w_1, \dots, w_m\}$, which is assumed to follow a multinomial distribution. The variable Y models the occurrence of a single word in a document, y_j denotes the occurrence of word w_j . Lastly, Z refers to an existing categorization scheme for documents, where $z(x)$ stands for the known class of document x .

This is a special case of the general conditional information bottleneck method. We propose to use the empirical feature counts or maximum likelihood estimates as a plug-in estimator for $p(y|x)$ and a uniform distribution over

documents,

$$p(z|x) = \delta(z, z(x)), \quad (17a)$$

$$p(y_j|x, z) = p(y_j|x) = \frac{n_j(x)}{\|n(x)\|_1}, \quad (17b)$$

$$p(x) = \frac{1}{n}. \quad (17c)$$

Hence Eq. (16) can be rewritten simply as

$$p(c|x) \propto p(c) e^{\text{EXP} \beta \sum_{j=1}^m \frac{n_j(x)}{\|n(x)\|_1} \log \frac{p(y_j, z(x), c)}{p(z(x), c)}} \quad (18)$$

Furthermore the self-consistency equations are given by

$$p(c) = \frac{1}{n} \sum_x p(c|x), \quad (19a)$$

$$p(z, c) = \frac{1}{n} \sum_{x:z(x)=z} p(c|x), \quad (19b)$$

$$p(y_j, z, c) = \frac{1}{\sum_x \|n(x)\|_1} \sum_{x:z(x)=z} n_j(x) p(c|x). \quad (19c)$$

Similarly, one can apply the conditional information bottleneck in the case of a Bernoulli sampling model for words. In this case, denote by $b(x) \in \{0, 1\}^m$ the binary representation of a document. The corresponding fixed point equations are given by

$$p(c|x) \propto p(c) e^{\beta \sum_{j=1}^m (b_j(x) \log \frac{p(y_j, z(x), c)}{p(z(x), c)})} \quad (20a)$$

$$\times e^{\beta \sum_{j=1}^m ((1-b_j(x)) \log (1 - \frac{p(y_j, z(x), c)}{p(z(x), c)}))}$$

and

$$p(y_j, z, c) = \frac{1}{n} \sum_{x:z(x)=z} b_j(x) p(c|x). \quad (20b)$$

5 Related Work

Techniques for introducing prior knowledge to perform constrained clustering have primarily focused on formulating the problem as one where the prior knowledge is expressed by *instance-level constraints*, such as [14, 9, 15]. As described in [14] these constraints typically take the form of relations such as *must-link* and *cannot-link* which are enforced between pairs of instances. This approach is extended by [9] which infers a proximity matrix for all instances from the given instance-level constraints and further by [15] where a formal distance metric is learned from the instance-level constraints information. In contrast to our setting, these approaches have been concerned primarily with settings in which the prior knowledge takes the form of positive information about the desired clustering.

Constraints may also be imposed at the feature level, as in [1] where domain knowledge is used to formulate rules which rank membership for clusters upon which an agglomerative technique is applied. This approach is designed for settings in which detailed expert knowledge is available from which to derive the rules.

Less work has been done on introducing prior knowledge as constraints at the model level. In [12], features are partitioned into relevant “useful” and irrelevant “noise” sets where the feature partition is incorporated as a parameter into the objective for model selection. Using this model, the enforcement of model-level constraints over the occurrence of noise and useful features is discussed in [13].

Co-clustering, also referred to as block clustering or bi-clustering, is the problem of simultaneously clustering both instances and features of a data set. While co-clustering does not make use of prior knowledge, it can be seen as a form of constrained clustering. Co-clustering can be expressed in a probabilistic model as learning a joint distribution over the instances and features as in [7] and [8] or solved using spectral techniques as in [5]. A related information-theoretic approach to co-clustering is given in [4].

6 Experimental Results

6.1 Synthetic Data

6.1.1 Generation

We generate a m -dimensional binary-valued test set \mathcal{X} with two natural independent partitionings, P and Q , where P partitions the data into clusters P_1 and P_2 and Q into Q_1 and Q_2 . We associate m_P of the features with partition P and the remaining $m_Q = m - m_P$ with Q . Representative profiles $\hat{p}_1 \sim \{0, 1\}^{m_P}$ and $\hat{q}_1 \sim \{0, 1\}^{m_Q}$ are randomly chosen for P_1 and Q_1 . \hat{p}_2 and \hat{q}_2 are set to the complement of \hat{p}_1 and \hat{q}_1 . n instances are then generated where for each instance membership is randomly chosen from $P \times Q$ where each of the 4 configurations have equal probability. The instance is assigned the corresponding \hat{p} and \hat{q} profiles and noise is added by flipping each feature with probability $p_{noise} = 0.1$. The resulting set should contain natural partitionings P and Q where the relative strength of the two partitionings can be altered by varying m_P .

6.1.2 Results

Results for the synthetic sets are shown in Figure 1. We generate $d = 100$ such synthetic sets and compare the averaged performance of Conditional IB versus EM. Results are measured against true labeling L using two metrics, Overlap and Normalized Mutual Information, where we consider Overlap only for the case where $K = 2$:

- $\text{Overlap}(C, L) = 1 - \frac{\min(|C-L|_1, |C-\bar{L}|_1)}{n/2}$
- Normalized Mutual Information: $\text{norm } I(C, L) = \frac{I(C, L)}{H(L)}$

We hope to see results in which the solutions have large overlap with the desired partitioning, Q , and small overlap with the conditioned partitioning, P . A clustering which overlapped exactly with Q and had no overlap whatsoever with P would score $[0, 100]$ in both metrics. Consider the first case, where $m_P = m_Q = 4$, which means the features are evenly split between the Q and P partitions. Not surprisingly, the EM algorithm finds solutions close to the Q and P partitions with equal frequency, resulting in an average score which is equally similar to both Q and P . The Conditional IB algorithm, however, is successful at finding solutions which are, on average, considerably more similar to the desired partitioning Q .

As we increase the proportion of features which are correlated with P , the EM algorithm increasingly settles on solutions similar to P . Despite this bias towards P , the Conditional IB still finds solutions which are substantially unlike P and have higher overlap with Q , showing the success of the algorithm on these synthetic examples.

Note that in the extreme case, when $m_Q = 2$, the median Overlap for Conditional IB is $[6.00, 19.00]$, versus $[8.00, 84.00]$ for $m_Q = 3$. While the second term, the similarity to Q , has degraded noticeably, the first term has actually improved somewhat, reflecting that the solutions for $m_Q = 2$ are further from the conditioned P . This illustrates a difficulty of evaluating unsupervised learning with negative conditional information and the importance of considering both terms: while the objective may constrain the algorithm to selecting solutions which are unlike the conditioned information P , we cannot be sure the solution which is found will be similar to the known “desired” partitioning Q . In this case, where the influence from Q in the data is weak, the algorithm may indeed find other valid clusterings.

6.2 Real-world Datasets

In a second set of experiments, we have used the CMU 4 Universities WebKB data set as described in [3] which consists of webpages collected from computer science departments and has label information L1 for Pagetype: (‘course’, ‘faculty’, ‘project’, ‘staff’, ‘student’) as well as information on University: (‘Cornell’, ‘Texas’, ‘Washington’, ‘Wisconsin’) which we use for L2. Documents belonging to the ‘misc’ and ‘other’ categories, as well as the ‘department’ category which contained only 4 members, were removed, leaving 1087 pages remaining. Categories vary considerably in size, as can be seen in Figure 2. Stopwords were removed and the remaining terms were filtered with a frequency cutoff of 100, leaving 239 terms.

Synthetic Data sets: $d = 100, n = 100, Z = P, K = 2$									
m_P, m_Q	Algorithm	Overlap				norm I(C,Z)			
		mean		median		mean		median	
		P	Q	P	Q	P	Q	P	Q
4, 4	EM	51.90	52.30	57.00	57.00	40.12	40.98	34.69	33.51
	Cond IB	7.26	51.62	6.00	87.00	0.65	42.72	0.27	66.39
5, 3	EM	81.62	19.16	98.00	8.00	76.72	9.78	86.67	0.57
	Cond IB	7.72	52.64	8.00	84.00	0.64	41.07	0.42	61.89
6, 2	EM	97.92	8.00	98.00	0.80	91.74	0.75	92.44	0.37
	Cond IB	5.78	40.62	6.00	19.00	0.41	24.58	0.22	2.63

Figure 1. Results on Synthetic Sets

	course	faculty	project	staff	student	TOTAL
Cornell	44	34	20	21	128	247
Texas	38	46	20	3	148	255
Washington	77	31	21	10	126	265
Wisconsin	85	42	2	12	156	320
TOTAL	244	153	86	46	558	1087

Figure 2. Document counts for WebKB labelings

In our experimental framework, we assume that one of the categorizations is known and we condition on that known categorization as Z . We also consider the setting where there is some labeled data available for the desired categorization L . More specifically we assume that in each category there are some l documents labeled according to the desired categorization, where l is typically considerably smaller than the total number of documents. We examine the effect of varying the proportion of labeled data on the achieved accuracy. We compare performance against basic EM augmented with a Naive Bayes model for labeled data, as described in [10]. In particular, the Naive Bayes model is initialized with the l documents labeled according to the desired categorization L while the Z information is ignored. For evaluation purposes, K is set to the labelset cardinality of the desired classification and the average of 5 initializations is taken.

The results in Figure 3 compare the performance between EM and Conditional IB with conditioned information. In the $Z = L2$ case, where we hope to find the $L1$ categorization, Conditional IB slightly outperforms EM in finding solutions similar to $L1$ and significantly outperforms EM in finding solutions which are unlike $L2$. For $Z = L1$, EM converges over all the values tested to solutions which are on average more similar to $L1$ than the desired $L2$. Even at the maximum proportion tested, EM still converges to solutions which look more like $L1$ than $L2$. This suggests that in this data set, the $L1$ (Pagetype) structure is stronger than $L2$ (University). Even so, the Con-

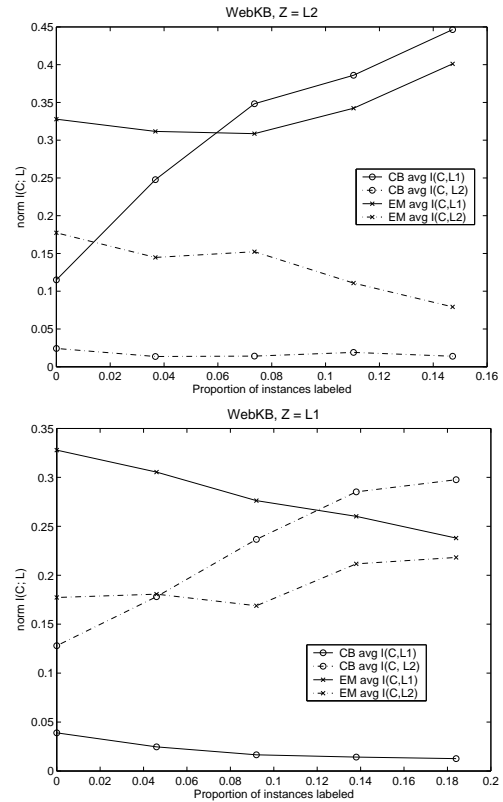


Figure 3. Results on WebKB with labeled instances

c = 1: cornell(0.049), lecture(0.024), upson(0.023), hours(0.016), midterm(0.013), ithaca(0.012), _time_(0.012), ny(0.012), homework(0.012), assignments(0.0099)
c = 2: austin(0.17), utexas(0.16), texas(0.16), tx(0.12), contact(0.074), ut(0.069), tay(0.057), sciences(0.052), address(0.027), fax(0.027)
c = 3: engineering(0.09), seattle(0.081), science(0.077), wa(0.074), project(0.055), student(0.045), work(0.045), year(0.042), system(0.042), working(0.041)
c = 4: madison(0.19), wisconsin(0.17), wi(0.17), dayton(0.16), sciences(0.1), street(0.096), _phonenum_(0.064), st(0.063), department(0.043), interests(0.042)

Figure 4. Top 10 per-cluster informative terms for $Z = \text{Pagetype}$, ranked by $I(W = w, C = c)$

c = 1: hours(0.18), assignments(0.18), instructor(0.16), _time_(0.16), class(0.15), lecture(0.14), syllabus(0.13), homework(0.12), final(0.11), assignment(0.1)
c = 2: conference(0.11), publications(0.1), professor(0.1), proceedings(0.1), acm(0.096), research(0.091), recent(0.089), ieee(0.086), workshop(0.085), journal(0.082)
c = 3: high(0.011), papers(0.01), support(0.0098), set(0.0096), students(0.0078), level(0.0074), people(0.0069), projects(0.0064), based(0.0064), performance(0.005)
c = 4: university(0.072), department(0.067), science(0.066), _fivedigit_(0.064), _phonenum_(0.052) computer(0.051), ithaca(0.05), student(0.048), ny(0.043), hall(0.043)
c = 5: _fivedigit_(0.039), wi(0.031), madison(0.03), dayton(0.03), austin(0.028), tx(0.027), wisconsin(0.026), street(0.025), student(0.021), sciences(0.021)

Figure 5. Top 10 per-cluster informative terms for $Z = \text{University}$, ranked by $I(W = w, C = c)$

ditional IB algorithm is able to consistently obtain solutions in which the similarity to $L2$ is close to zero, and where for fractions of labeled data > 0.046 , the similarity to $L1$ becomes significantly higher than the solutions found by EM.

We can also evaluate the clustering solutions by examining the influential terms in each clustering. In particular, we take a representative clustering where proportion of labeled instances $\approx 14\%$. In order to consider only those terms which are discriminative between clusters, we rank the top 10 terms w per cluster c by $I(W = w, C = c) = p(w, c) \log \frac{p(w|c)}{p(w)}$. Non-discriminative terms, which would be equiprobable in all clusters, will have $p(w|c) = p(w)$ so $I(W = w, C = c) = 0$. As can be seen in Figure 4 where conditioning is on Pagetype, the resultant clustering is consistent with the University classification, with informative terms such as “cornell”, “austin”, “seattle”, and “madison”. The clusters resulting from conditioning on University are less sharp. As suggested by Figure 5, examining the labelings show that cluster 1 contains class webpages, cluster 2 is primarily faculty pages, and cluster 3 contains mostly project webpages. Clusters 4 and 5 contain the majority of the student webpages.

7 Conclusion

We have presented an extension of the information bottleneck framework, called conditional information bottleneck, which allows the incorporation of known information in a natural way. We have investigated an application of this approach to non-redundant data clustering where a known categorization is available and the goal is to find an alternative way of clustering the data. Our experiments on synthetic data and the WebKB document clustering task have validated our approach. In both data sets, the proposed approach is able to recover good approximations of one classification, given the other classification as side information. We have also shown that negative feedback can greatly increase the classification accuracy in the case of learning with partially labeled data.

References

- [1] J. Béjar, U. Cortés, and R. Sanguesa. Experiments with domain knowledge in knowledge discovery. In *Proceedings of the 1st Int. Conference on the Practical Application of Knowledge Discovery and Data Mining (PADD '97)*, pages 187–197, London, UK, 1997.
- [2] G. Chechik and N. Tishby. Extracting relevant structures with side information. In *Advances in Neural Information Processing Systems 15 (NIPS '02)*, 2002.

- [3] M. Craven, D. DiPasquo, D. Freitag, A. K. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the World Wide Web. In *Proceedings of the 15th Conference of the American Association for Artificial Intelligence, (AAAI '98)*, pages 509–516, Madison, US, 1998. AAAI Press, Menlo Park, US.
- [4] I. Dhillon, S. Mallela, and D. S. Modha. Information-theoretic co-clustering. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*, 2003.
- [5] I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*, pages 269–274, 2001.
- [6] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI '01)*, 2001.
- [7] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval (SIGIR '99)*, 1999.
- [8] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI '99)*, 1999.
- [9] D. Klein, S. D. Kamvar, and C. D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML '02)*, 2002.
- [10] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
- [11] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.
- [12] S. Vaithyanathan and B. Dom. Model selection and document clustering. In *Neural Information Processing Systems (NIPS '99)*, 1999.
- [13] S. Vaithyanathan and D. Gondek. Clustering with informative priors. Technical report, IBM Almaden Research Center, 2002.
- [14] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00)*, pages 1103–1110, 2000.
- [15] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems, 15 (NIPS '03)*, 2003.