Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

Search Result Refinement via Machine Learning from Labeled-Unlabeled Data for Meta-search

I. Burak Ozyurt
Department of Psychiatry
UCSD LOCI MC 9151-B
9500 Gilman Drive
La Jolla, CA 92093
FAX: 858-642-1458
Email: iozyurt@ucsd.edu

Greg G. Brown
Department of Psychiatry
UCSD LOCI MC 9151-B
9500 Gilman Drive
La Jolla, CA 92093
FAX: 858-642-1458
Email: gbrown@ucsd.edu

Abstract— **For a user, retrieving relevant information from search engines involves encoding her intent, at best partially, in search keywords. A small amount of user feedback, can be beneficial in refining the results returned by the search engines and aiding exploratory search for scientific literature and data. In this paper, three new variants to EM method for semi-supervised document classification [1] is introduced for biomedical literature meta-search result refinement. Multi-mixture per class EM variant with Agglomerative Information Bottleneck clustering [2] using Davies-Bouldin cluster validity index [3], has shown retrieval performance rivaling the state of the art transductive support vector machines (TSVM) [4] with more than one order of magnitude improvement in execution time.**

I. Introduction

Retrieving relevant information from vast array of unstructured or semi-structured documents becomes more important with ever increasing amount of information produced every year. Since it is intractable for a single person to comb through the vast array of information produced by millions of people all around the world in search of his/her area of interest, automatic retrieval of information is the only tool we can rely on. There are numerous general purpose and domain specific search engines which usually suffer from low precision rate. This translates going through search results to look for relevant documents. Most search engines use term based queries and mostly differ in their document index model representation and document ranking function. Doing a search query, the user must translate the context of his/her intent into search terms. Usually this translation is at best partial, hence the search result reflect this partial context. Provided with search results, however, the user can indicate a small set of results as relevant providing feedback to the information retrieval system to re-rank the search results for higher precision rate. Since, most users don't bother looking more than a few search result pages, high precision is very important.

In this paper, a search result refinement system as a part of meta-search engine for Biomedical Informatics Research Network (BIRN)[1] Query Atlas [5] system is introduced. Query Atlas combines browsing/analysis of (functional) magnetic resonance imaging (fMRI) data with text/literature mining. The introduced search result refinement mechanism forms one of the key elements of text mining portion of the Query Atlas. Three new extensions to expectation maximization (EM) algorithm [1] for text classification from labeled and unlabeled data are introduced and compared against transductive Support Vector Machines (SVMs) [4] as candidates for the search result refinement mechanism for Query Atlas meta-search engine.

[1] http://www.nbirn.net

II. Transductive SVMs

Transductive SVM approach can be seen a form of learning from both labeled and unlabeled data similar to the improvement of Naive Bayes classification results via mixture parameter learning via EM from unlabeled data [1]. In relevance feedback, we cannot expect a user to provide enough labeled training data that would be necessary for inductive learning. Transductive approach [9] allows learning from small number of labeled and large set of unlabeled data. The unlabeled data, initially appears to be useless for learning a classification function. However, in natural language words occur in strong co-occurrence patterns [10]. Some words are more likely to occur in related documents than the unrelated documents. This phenomena is exploited by the unsupervised document categorization approaches. Transductive SVMs also exploit this phenomena to improve the test classification performance over limited number of labeled training data with readily available unlabeled data. The hypothesis space H for the learning task, which can have infinite dimensionality, can be divided into finite number of equivalence classes H' of functions that classify labeled and unlabeled training sample the same way [4]. Structural risk minimization [9] states that, for the smallest upper bound on the expected risk we should select a function from ordered set of equivalence classes H' of increasing VC-dimension that minimizes the confidence interval on the test error. This leads for transductive SVM learning for linearly separable case to the following convex

constrained optimization problem [4]

$$\begin{array}{ll} \min_{\vec{w},b,y_i^*} \frac{1}{2}||\vec{w}||^2 \\ \text{subject to:} & y_i(\vec{x_i}.\vec{w}+b) \geq 1 \; \forall i \\ & y_i^*(\vec{x_i^*}.\vec{w}+b) \geq 1 \; \forall i \end{array} \quad (1)$$

Here $\vec{x_i^*}$ corresponds to the ith unlabeled training instance and y_i^* corresponds to the *estimated* label for this unlabeled instance. In transductive SVM learning, besides the weights $\vec{w}$ and bias b , the unknown labels y_i^* are also estimated to find the maximum margin separating hyperplane for *both* labeled and unlabeled training data. For non-separable case, analogous to inductive discriminative SVMs, the constraints $y_i(\vec{x_i}.\vec{w}+b) \geq 1$ and $y_i^*(\vec{x_i^*}.\vec{w}+b) \geq 1$, are relaxed by introducing slack variables ξ_i for labeled examples and ξ_i^* for unlabeled ones.

$$\begin{array}{ll} \min_{\vec{w},b,y_i^*,\xi_i,\xi_i^*} \frac{1}{2}||\vec{w}||^2 + C\sum_i \xi_i + C^* \sum_i \xi_i^* \\ \text{subject to:} & y_i(\vec{x_i}.\vec{w}+b) \geq 1-\xi_i \; \forall i \\ & y_i^*(\vec{x_i^*}.\vec{w}+b) \geq 1-\xi_i^* \; \forall i \\ & \xi_i > 0 \; \forall i \\ & \xi_i^* > 0 \; \forall i \end{array} \quad (2)$$

Here C and C^* are user specified penalties for labeled and unlabeled misclassification error, respectively. In this paper, SVM^{light} [11] package is used.

III. Mixture of Unigrams Expectation Maximization from labeled and unlabeled data

A document d_i from a corpus can be seen as a list of ordered words $w_{t,d_{i,k}}$ $k = 1, \ldots, |d_i|$ and $t \in V$. Here $|d_i|$ the length(cardinality) of document i in terms of words and the set V is the vocabulary. The documents are assumed to be generated according a probability distribution with parameters θ. This probability distribution is assumed to be generated from a mixture of components/topics c_j $j = 1, \ldots, J$. Thus a document d_i given the parameter θ can be expressed as

$$p(d_i|\theta) = \sum_{j=1}^{J} p(c_j)p(d_i|c_j;\theta) \quad (3)$$

The probability of d_i given the mixture component c_j can be written as the probability of choosing document length $|d_i|$ and observing the sequence of words in the document as follows;

$$p(d_i|c_j;\theta) = p(|d_i|)\prod_{k=1}^{|d_i|} p(w_{t,d_{i,k}}|c_j; w_{t,d_{i,q}}, q<k;\theta) \quad (4)$$

Using the Naive Bayes assumption, i.e. the words of a document are generated independent of each other, (4) becomes

$$p(d_i|c_j;\theta) = p(|d_i|)\prod_{k=1}^{|d_i|} p(w_{t,d_{i,k}}|c_j;\theta) \quad (5)$$

Although Naive Bayes assumption is definitely unrealistic as a human language model, under zero-one loss function for classification error, learning methods with large modeling bias can work very well for classification, where the only requirement is a negative boundary bias [12]. The parameters of this mixture of components/topics model are $\theta = \theta_{w_t|c_j} : w_t \in V, c_j \in C; \theta_{c_j} \in C$.

Naive Bayes classifier uses the maximum a posteriori (MAP) estimate of the model parameters θ, i.e. $\text{argmax}_\theta \, p(\theta|D)$. Using Bayes rule and the conjugate prior of multinomial distribution, Dirichlet prior $p(\theta) \propto \prod_{c_j \in C} \theta_{c_j}^{\alpha-1} \prod_{w_t \in V} \theta_{w_t|c_j}^{\alpha-1}$;

$$\begin{array}{rcl} p(\theta|D) & \propto & p(\theta)p(D|\theta) \\ & \propto & \prod_{c_j \in C} \theta_{c_j}^{\alpha-1} \prod_{w_t \in V} \theta_{w_t|c_j}^{\alpha-1} \prod_{d_i} p(d_i|\theta) \end{array} \quad (6)$$

Using $\alpha = 2$ for Dirichlet prior and the constraint that the word probabilities in a class must sum to one, the parameters $\hat{\theta}$ maximizing $\log p(\theta|D)$ can be computed by Lagrange multiplier method. This specific value of α results in Laplace smoothing of the probability estimates. The word probability estimates $\hat{\theta}_{w_t|c_j} \equiv p(w_t|c_j;\hat{\theta})$, are expressed as;

$$\hat{\theta}_{w_t|c_j} = \frac{1 + \sum_{i=1}^{|D|} N(w_t, d_i)p(y_i = c_j|d_i)}{|V| + \sum_{s=1}^{|V|}\sum_{i=1}^{|D|} N(w_s, d_i)p(y_i = c_j|d_i)} \quad (7)$$

The class prior probabilities with Laplace smoothing are;

$$\hat{\theta}_{c_j} \equiv p(c_j|\hat{\theta}) = \frac{1 + \sum_{i=1}^{|D|} p(y_i = c_j|d_i)}{|C| + |D|} \quad (8)$$

The Naive Bayes classifier then uses the parameters calculated from training documents to estimate the most likely class y_i for a new document by using Bayes rule;

$$p(y_i = c_j|d_i;\hat{\theta}) = \frac{p(c_j|\hat{\theta})\prod_{k=1}^{|d_i|} p(w_{t,d_{i,k}}|c_j;\hat{\theta})}{\sum_{r=1}^{|C|} p(c_r|\hat{\theta})\prod_{k=1}^{|d_i|} p(w_{t,d_{i,k}}|c_r;\hat{\theta})} \quad (9)$$

Here one mixture component per class is used. Nigam et al. [1] also provide multiple mixture components/topics per class EM version, which is used as the basis for the cluster validity based enhancement introduced in this paper.

For Nigam et al.'s [1] EM approach, the training data set consists of a small set of labeled documents and a much larger set of unlabeled documents $D = D^l \cup D^u$ which may convey further information in the form of co-occurrence of words both found in labeled and unlabeled documents. EM algorithm for mixture components without labeled data can be seen as the soft version of k-means clustering algorithm. In standard EM for mixture components, the unobserved (latent) mixing proportions ($\mathbf{Z}$) are estimated from the observed data (X) by finding the parameters $\hat{\theta}$ that maximizes the log likelihood of complete data $L(\mathbf{X}, \mathbf{Z}|\theta)$. EM algorithm (method) finds a local maximum by iterating an expectation step $\hat{\mathbf{Z}}^{(k+1)} = E[\mathbf{Z}|\mathbf{X};\hat{\theta}^{(k)}]$ followed by a maximization step

$$\hat{\theta}^{(k+1)} = \arg\max_{\theta} \; p(\theta|\mathbf{X};\mathbf{Z}^{(k+1)})$$

Basic EM method for one mixture component per class, starts by a 'priming' M-step, i.e. estimating parameters for Naive Bayes classifier from just the labeled set by (7) and (8). The EM iterations, then, begins by an E-step using the Naive Bayes classifier (9) to estimate the most likely class/mixture component for the unlabeled documents followed by the M-step where the new MAP estimates for parameters $\hat{\theta}$ are

calculated using the current estimates for $p(c_j|d_i;\hat{\theta})$ and by (7) and (8).

A. Multiple Mixture Component/Topic per class case

It is restrictive to assume that for each class documents will belong to a single topic. Nigam et al. [1] extended their basic EM method to multi topic/mixture component per class model. Using c_j to continue to denote the jth mixture component , we can define the ath class as G_a. Then, the class probability of a document can be expressed as the weighted sum of the mixture component probabilities;

$$p(G_a|d_i;\hat{\theta}) = \sum_{c_j \in G_a} p(G_a|c_j;\hat{\theta}) \times \frac{p(c_j|\hat{\theta}) \prod_{k=1}^{|d_i|} p(w_{t,d_{i,k}}|c_j;\hat{\theta})}{\sum_{r=1}^{|C|} p(c_r|\hat{\theta}) \prod_{k=1}^{|d_i|} p(w_{t,d_{i,k}}|c_r;\hat{\theta})} \quad (10)$$

IV. Multiple Mixture Per Class Component Number/Member selection by Hierarchical Clustering and Cluster Validity

Nigam et al. [1] report using cross-validation for mixture component/topic number selection and they uniformly distribute the labeled documents between the selected number of mixture components. However, in an on-line search refinement situation, what we can expect from the user is to indicate at most a few documents as relevant documents. The negative class consists of the documents in between the relevant ones and usually much larger than the positive (relevant) class. However, the most important constraint is the execution time. The user expects the filtered results within a few seconds, unlike offline document classification. The disadvantages of cross-validation in this situation is both the scarcity of the training data and most importantly time needed to do cross-validation which can be prohibiting. Due to the stochastic nature of mixture component initialization, in determining the number of mixture components multiple random initializations at each tested mixture component count will also be necessary; thus multi-folding the cross-validation time.

The effect of random initializations at different mixture component/cluster numbers ranging from 2 to 10 for the negative class for Nigam et al.'s multiple mixture component per class EM (MMEM) for some of the data sets with most variance is shown in Fig. 1. For the nine test cases used in this study, eight had enough data to do this analysis. For each of these eight data sets, for each cluster number from 2 to 10, 30 random initialization for MMEM is done. As the results show, MMEM is rather sensitive both to the number of mixture components and the initialization, which demonstrates the need for multiple initializations at multiple cluster numbers for cross-validation. For each data set 30 initializations for 2 to 10 clusters took roughly 10 minutes on a Pentium 4 3.0 GHz machine.

In the light of this findings a more time efficient approach for simultaneous mixture component initialization and "optimal" cluster number determination by using an agglomerative

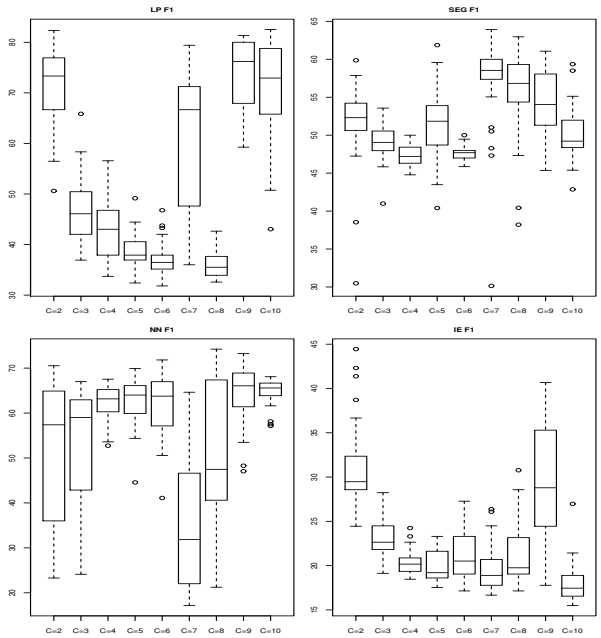

Fig. 1. Test F_1 percentage values for 30 random mixture component initializations for negative (non-relevant) class at cluster numbers 2 to 10.

hierarchical clustering algorithm [13] with cluster validity based cluster number determination is introduced. Hierarchical clustering has the advantages of being applicable to cases where we are unable to supply a distance metric, but pairwise dissimilarity values for every pair of samples and generating a cluster hierarchy as the output. As the dissimilarity value between pairs of search result documents, which consists of a title and (if available) an abstract, symmetric Kullback-Leibler (KL) divergence is used as defined as;

$$sD_{KL}[p(x)||q(x)] = 0.5 \times (D_{KL}[p(x)||q(y)] + D_{KL}[q(y)||p(x)]) \quad (11)$$
$$D_{KL}[p(x)||q(x)] = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Symmetric KL Divergence is not a metric since it does not satisfy triangle inequality. A Farthest-Neighbor Agglomerative clustering based on symmetric KL divergence as dissimilarity measure between document word probability distributions is introduced (See Algorithm 1). Main disadvantages of the hierarchical clustering approaches are $O(cn^2d^2)$ time complexity and $O(n^2)$ space complexity and possible suboptimal clustering due to the greedy nature of the algorithm and non-convexity of most clustering problems. Here, n is the number of documents, d is the dimension of the solution space, which is one for this case and c is the number of clusters. Due to the nature of the on-line result refinement, the training data set (i.e. n) is rather small, hence both time and space complexity is manageable. Suboptimal clustering due to local minima is the common problem of all clustering algorithms, for which

Algorithm 1 Farthest-Neighbor Agglomerative Hierarchical Clustering

initialize $\hat{c} \Leftarrow n,\ C_i \Leftarrow \{d_i\}\ i = 1, \ldots, n$
while $\hat{c} > 1$ **do**
find nearest clusters by

$$\underset{i',j'}{\operatorname{argmin}} \max_{p(x)\in C_{i'}\ q(x)\in C_{i'}} sD_{KL}[p(x)||q(x)]$$

merge clusters C_i and C_j
$\hat{c} \Leftarrow \hat{c} - 1$
end while
return cluster hierarchy

there is no time efficient solution.

Many cluster validity indices are introduced over time to evaluate partitioning induced by clustering algorithms ([14],[3],[15]). Maulik and Bandyopadhyay [15], compares the most common cluster validity indices on center based clustering algorithms including well known K-means. In this study, the two best cluster validity indices for center based clustering, namely, the $\mathfrak{I}$ index [15] and Davies-Bouldin index [3] are adapted for pairwise clustering and tried to determine the "optimal" cluster number. Davies-Bouldin index [3] is found to be the better performing one. This index is a function of the ratio of the sum of within-cluster scatter to between cluster separation. The "optimal" cluster number occurs at the minimum of the DB index with increasing number of clusters. In practice, sometimes cluster validity indices keep decreasing with the increase of the number of clusters. In this case, the minimum slope of the DB index curve can be used as an indicator where the addition of another cluster will have only a marginal effect.

The within-cluster scatter for the ith cluster is calculated as $S_i = \frac{1}{|C_i|}\sum_{p_j(x)\in C_i} sD_{KL}[p_j(x)||q(x)]$, where $q(x)$ is word probability distribution for the centroid. The centroid of a cluster is defined as the document which has the least maximum dissimilarity from the other cluster documents, i.e.

$$\underset{j}{\operatorname{argmin}} \max_{k,k\neq j} sD_{KL}[p_j(x)||p_k(x)] \tag{12}$$

Cluster separation d_{ij} between cluster C_i and C_j is defined as $sD_{KL}[q_i(x)||q_j(x)]$, where $q_i(x)$ is the word probability distribution for the centroid of cluster i. The Davies-Bouldin (DB) index is defined as

$$DB = \frac{1}{K}\sum_{i=1}^{K} \max_{j,j\neq i}\left\{\frac{S_i + S_j}{d_{ij}}\right\} \tag{13}$$

The "optimal" number of clusters is determined by the minimum or the first zero crossing of the negative of the scaled numerical derivative of the $DB(t)$ function ($-\frac{DB_t - DB_{t+1}}{DB_t}$), where t, corresponding to K in (13), denotes the number of clusters for a particular partition.

A. Multiple Mixture Per Class Component Number/Member selection by Agglomerative Information Bottleneck and Cluster Validity

Agglomerative Information Bottleneck (AIB) [2] is a hierarchical, bottom-up, distributional hard clustering algorithm that maximizes the mutual information per cluster between the data and given categories. AIB, a variant of Information Bottleneck method, finds a compressed version of variable X, corresponding to the clusters Z, such that the mutual information between Z and a relevant variable Y, $I(Z,Y)$, which is conditionally dependent on X but independent on Z (i.e. forming a Markov chain $Z \rightarrow X \rightarrow Y$) is maximized under a constraint on the mutual information between X and Z. The solution to this constrained optimization problem results in self-consistent equations for the conditional distributions $p(y|z)$, $p(x|z)$ and $p(z)$. AIB uses Jensen-Shannon divergence as the distortion measure between conditional distributions of the relevant variable given clusters $p(y|z_i)$, $i \in \{1, \ldots, M\}$,

$$JS_\pi[p(y|z_1), \ldots, p(y|z_M)] = \sum_{i=1}^{M} \pi_i D_{KL}[p(y|z_i)||\sum_{i=1}^{M} \pi_i p(y|z_i)] \tag{14}$$

Here π_i is the prior probability of z_i. For document classification, X corresponds to documents, Y corresponds to terms occurring in a document and Z corresponds to the document clusters we are after. The AIB algorithm, requires an empirical joint probability matrix $p(x,y)$ which is calculated from term frequency information. For a document set of N, first N clusters z_i are created with self consistent equations $p(z_i) = p(x_i), p(y|z_i) = p(y|x_i), p(z|x_i) = 1, p(z|x_{j\neq i}) = 0$. For each pair of clusters z_i and z_j, the distortion $d_{ij} = (p(z_i) + p(z_j))J_{\pi_2}[p(y|z_i), p(y|z_j)]$ is calculated. After initialization, AIB creates a cluster hierarchy in $N-1$ iterations. At each iteration, the two clusters with minimum distortion is merged, decreasing the the cluster number by one. The distributions $p(\bar{z}), p(y|\bar{z})$ and $p(\bar{z}|x)$ for the new merged cluster $\bar{z}$ and corresponding distortion values are updated.

Davies-Bouldin index [3] is adapted for AIB using Jensen-Shannon (JS) divergence instead of symmetric KL divergence in a similar fashion as defined before. The JS divergence is calculated between each pair of documents in each cluster for Davies-Bouldin index calculations. Also, the "optimal" number of clusters is determined by the minimum or the first zero crossing of the negative of the scaled numerical derivative of the $DB(t)$ function, as in the farthest-Neighbor hierarchical clustering algorithm.

V. Simulated Annealing EM for Labeled/Unlabeled Document Classification

Simulated annealing [16], is a well known stochastic search technique used to find global optimum in problems with multiple local extrema. Simulated annealing has a temperature parameter similar to its physical counterpart. Valleys and peaks of the solution space with objective function value differences

less than the temperature value become the search space for simulated annealing at that temperature. The annealing process starts at a high temperature, where the data points can belong to any mixture more or less with equal probability, the temperature is gradually lowered according to a cooling scheme, at zero temperature a data point will belong to a single mixture component with probability 1. For simulated annealing EM, E-step of EM as defined by 9 is replaced by

$$p(y_i = c_j|d_i;\hat{\theta}) = \frac{p(c_j|\hat{\theta})\exp\{\frac{\log p(d_i|c_j;\hat{\theta})}{T})\}}{\sum_{r=1}^{|C|} p(c_r|\hat{\theta})\exp\{\frac{\log p(d_i|c_r;\hat{\theta})}{T})\}} \quad (15)$$

where T is temperature parameter and $p(d_i|c_j;\hat{\theta})$ values are calculated by 5. The derivation for this method is given in Appendix.

VI. Experimental Design

To test the classification methods described above, the introduced meta-search engine is used to query the National Library of Medicine's search service PubMED[2]. A typical query result consists of the title, authors, journal information and abstract of a paper. The Query Atlas meta search engine[3] returns by default 200 top ranking search results from PubMED. Nine queries on topics the first author has some expertise about are used to generate test data sets. The queries as shown in Table I. From this query result set, the ones which seem relevant to the author is tagged as relevant. For example for "neural network" query the relevant documents are related to machine learning algorithms but not to the neural networks in the brain. The goal of an efficient search refinement algorithm is to detect complex implicit dependency patterns in relevant documents which cannot be expressed explicitly in search queries. Overall 1800 abstracts are tagged as either relevant or non-relevant. For training, first few relevant documents scanning from the top ranked document down are used as the positive examples and the non-relevant examples up to the last positive example make up the negative examples. The number of positive examples is selected such that the negative class has enough data to test mixture component clustering and EM algorithm improves upon Naive Bayes. The remaining documents in each 200 document data set constitute the test cases. The hand tagged document sets are summarized in Table I. Standard information retrieval performance measures *precision* P, *recall* R and harmonic mean of precision and recall F_1 are used. These measures are defined as

$$P = \frac{|\{\text{relevant docs}\} \cap \{\text{retrieved docs}\}|}{|\{\text{retrieved docs}\}|} \quad (16)$$

$$R = \frac{|\{\text{relevant docs}\} \cap \{\text{retrieved docs}\}|}{|\{\text{relevant docs}\}|} \quad (17)$$

$$F_1 = \frac{2 * P * R}{P + R} \quad (18)$$

Only the title and abstract of a search result is used for classification. The numeric values and special characters and stop-words in the documents are discarded. The common tf-idf term weighting scheme $N(w_t, d_i)\log\frac{|D|}{|d_i \supset w_t|}$ for word/term frequencies is used in all experiments instead of pure word frequencies. For KL and JS divergence calculations, Laplace smoothing is used to prevent zero probability problems. For simulated annealing version of EM, the starting temperature T_0 used is 5. Temperature is decreased linearly down to 1.0 in 10 iterations of EM. All EM variants are run for 10 iterations.

[2] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
[3] Available at https://loci.ucsd.edu/qametasearch/

TABLE I
Data Sets

Document Set	# Relevant (Training)	# Irrelevant (Training)	# Relevant (Test)
Language processing (LP)	2	25	29
Segmentation (SEG)	4	12	50
Neural networks (NN)	4	21	82
Information Extraction (IE)	4	53	12
Text Categorization (TC)	4	4	41
Ontology (Ont)	9	17	83
Active Contour (AC)	9	22	31
Support Vector Machines (SVM)	5	13	35
Semantic (SEM)	4	30	32

VII. Experimental Results

The results for Naive Bayes (NB), EM with single mixture component per class (EM), MMEM with symmetric KL divergence based hierarchical clustering with Davies-Bouldin (DB) cluster validity index (MMEM(KL)), EM with Agglomerative Information Bottleneck clustering with DB cluster validity index (MMEM(AIB)), simulated annealing EM with single mixture component per class, K-nearest neighbor classifier with $k = 1$ (KNN) and Transductive SVM (TSVM) are summarized in Table II.

Based on the average F_1 values, MMEM(AIB) performed best followed by transductive SVM and SAEM. SAEM and MMEM(AIB) were also about 60 times and 20 times faster than TSVM on average, respectively. Single mixture component per class EM for the datasets SEG, NN and SEM showed the worst performance in all the EM methods leaving room for improvement. For SEG, SAEM was able to find a much better local maximum, but did only slightly better for NN and SEM datasets. Overall, all of the newly introduced EM variants achieved better average F_1 values than Nigam et al.'s EM. For data set SVM, single mixture component per class EM had better performance than MMEM variants. The determination of the degenerate case single "optimal" cluster is not possible with cluster validity indices, which are not even defined for a single cluster. Levine and Domany [17] proposed a cluster validity method based on re-sampling which can detect if the data has no cluster tendency. However, the data size requirements and time complexity of this method makes it infeasible for search results refinement. Also, for data set AC,

TABLE II
RESULTS SUMMARY FOR F_1 PERCENTAGES AND AVERAGE EXECUTION TIMES IN SECONDS

Data Set	NB	EM	MMEM(KL)	MMEM(AIB)	SAEM	KNN	TSVM
LP	45.0	70.8	85.2	74.6	72.0	61.2	61.9
SEG	10.5	22.6	24.6	46.5	60.0	40.0	54.2
NN	6.7	21.8	18.6	51.5	25.5	34.2	39.6
IE	28.6	64.5	66.7	45.8	68.8	34.5	54.5
TC	62.3	63.4	63.4	63.4	64.0	55.4	63.0
Ont	63.5	74.5	75.3	74.9	81.4	64.7	71.3
AC	51.2	48.6	48.1	48.6	43.3	41.3	57.5
SVM	40.0	62.5	56.1	62.3	63.8	47.4	55.8
SEM	10.5	14.3	18.6	69.8	18.2	39.2	61.5
Average	**35.4**	**49.2**	**50.7**	**59.7**	**55.2**	**46.4**	**57.7**
Time(SD)	0.7 (0.2)	1.1 (0.3)	3.5 (2.9)	3.6 (2.8)	1.2 (0.3)	1.0 (0.3)	72.4 (60.6)

EM and all its variants deteriorated Naive Bayes performance. Similar behavior is also reported in [1]. SAEM performance for data set AC was worse than EM. This was due to the fact that the cooling scheme is suboptimal. It is shown in [18], that in theory global optimum can be achieved if the cooling schedule obeys $T \propto \frac{1}{\log n}$ where n is the number of current iteration, which is unrealistic even for offline applications.

From the random initialization and different mixture component number experiments before, it seems that multi-mixture component EM is sensitive to initialization and number of mixture components. Both clustering based approaches introduced provide an one-pass, automatic way of selecting the number of mixture components and initialization of them while providing especially for MMEM(AIB), substantial improvement over one mixture component per class EM within user acceptable execution time. SAEM, on the other hand, assumes one mixture component per class, and tries to avoid getting stuck in a local maxima while still operating within user acceptable execution times. While the performance improvement over transductive SVM for MMEM(AIB) is not statistically significant using paired Wilcoxon signed rank test at $\alpha = 0.05$, the improvement in execution time, which is essential for the applicability of a method in search results refinement, is statistically significant.

The search result refinement mechanism integrated with Query Atlas meta-search engine uses KNN for degenerate cases with no negative examples and switches to MMEM(AIB) if there are more than three positive examples. This conservative threshold on positive examples is chosen to ensure that Naive Bayes, the underlying algorithm for MMEM(AIB), has always enough examples for best performance.

VIII. RELATED WORK

Relevance feedback, being one the most popular query reformulation strategies for IR, is based on query expansion and/or term reweighting techniques available for vector and probabilistic models of information retrieval [6]. The semi-supervised machine learning approach to relevance feedback taken here is akin to the Bayesian classification model of retrieval [7], where the relevant documents are used to model the relevant class for the query and the remaining corpus for the non-relevant class. The main difference is the incorporation of unlabeled data for better estimation of both relevant and non-relevant classes in the classification model.

Similar to [8], multiple topic per class EM extensions introduced, use a language modeling perspective. Unlike [8], however, where the language model is applied to the user query, in these introduced extensions, the relatively more abundant non-relevant labeled documents are represented as a mixture of unknown but to be estimated number of topic language models.

IX. CONCLUSION

In this paper, three new variants to Nigam et al.'s [1] EM approach for document classification from small number of labeled documents and larger set of unlabeled ones are introduced in search for a time efficient search result refinement mechanism for BIRN Query Atlas meta-search engine. All of the methods introduced on average outperformed basic EM approach. MMEM(AIB) has shown better average F_1 performance, though not statisticaly significant, than the state of the art transductive SVMs with a more than one order of magnitude improvement in execution time. While the introduced EM variants are used for search result refining in Query Atlas meta-search engine, its application is not limited to meta-search. These approaches can also be used for other personalized information filtering tasks including email or news filtering.

X. FUTURE DIRECTIONS

One way for possible improvement of F_1 performance, which can be applied to all of the introduced approaches, is to use part-of-speech based morphological preprocessing of the unigrams to avoid term mismatches due to morphological term differences. Another approach, also equally applicable to all of the introduced methods, is to use a controlled vocabulary like Medical Subject Headings (MeSH) [19] to introduce synonyms in term/phrase matching and probability calculations. For SAEM, the effect of different cooling mechanisms on the performance can also be investigated.

APPENDIX
DERIVATION OF SIMULATED ANNEALING EM DOCUMENT CLASSIFICATION FROM LABELED AND UNLABELED DATA

The derivation follows directly Mak et al.'s [20] simulated annealing mixture model EM from entropy interpretation of the generic EM method. Let J stand for the number of mixture components and N stand for the number of observed data points $\mathbf{x_n} = (x_1, \ldots, x_k)$ $n = 1, \ldots, N$. In standard EM for mixture components, the unobserved (latent) mixing proportions ($\mathbf{Z}$) are estimated from the observed data (X) by finding the parameters $\hat{\theta}$ that maximizes the log likelihood of complete data $L(\mathbf{X}, \mathbf{Z}|\theta)$. The log likelihood for incomplete data can expressed as

$$L(\mathbf{X}|\theta) = \sum_{n=1}^{N} \log p(\mathbf{x_n}|\theta) \tag{19}$$

where

$$p(\mathbf{x_n}|\theta) = \sum_{j=1}^{J} p(\delta_{n,j} = 1|\theta) p(\mathbf{x_n}|\delta_{n,j} = 1, \phi_j) \tag{20}$$

Here $\delta_{n,j}$ is an indicator function, taking value of 1 if the nth data point belongs the jth mixture component, and zero otherwise. The ϕ_j represents the parameters for the jth mixture component. The probability $p(\delta_{n,j} = 1|\theta)$ corresponds to the prior probability for the jth mixture component.

$$\begin{aligned} p(\mathbf{x_n}|\theta) &= \sum_{n=1}^{N} \frac{p(\mathbf{x_n}|\theta)}{p(\mathbf{x_n}|\theta)} \log p(\mathbf{x_n}|\theta) \\ &= \sum_{n=1}^{N} \frac{\sum_{j=1}^{J} \pi_j p(\mathbf{x_n}|\delta_{n,j} = 1, \phi_j)}{p(\mathbf{x_n}|\theta)} \\ &\quad \times \log p(\mathbf{x_n}|\theta) \end{aligned} \tag{21}$$

Let define $h_{j,n}$ as the probability of a data point $\mathbf{x}_n$ belonging to the jth mixture component for the given parameter set ϕ_j. This can be expressed by using Bayes rule from the mixture prior π_j and likelihood $p(\mathbf{x}_n|\delta_{n,j} = 1, \phi_j)$ as

$$h_{j,n} \equiv p(\delta_{j,n} = 1|\mathbf{x}_n, \phi_j) = \frac{\pi_j p(\mathbf{x}_n|\delta_{j,n} = 1, \phi_j)}{p(\mathbf{x_n}|\theta)} \tag{22}$$

Using $h_{j,n}$, adding and subtracting $\log[\pi_j p(\mathbf{x}_n|\delta_{j,n} = 1, \phi_j)]$ to (21) and rearranging

$$\begin{aligned} L(\mathbf{X}|\theta) = &\sum_n \sum_j h_{j,n} \log \pi_j \\ &+ \sum_n \sum_j p(\mathbf{x}_n|\delta_{n,j} = 1, \phi_j) h_{j,n} \\ &- \sum_n \sum_j h_{j,n} \log h_{j,n} \end{aligned} \tag{23}$$

The last term in the above equation is entropy of the membership function h. In this equation, the first two terms constitute the complete data log likelihood. $L(\mathbf{X}|\theta)$ can be maximized also by maximizing the entropy term. From statistical physics/thermodynamics, it is known that if ensembles are distributed by a Gibbs/Boltzmann distribution, the entropy of the system is maximized.

Introducing a simulating annealing like temperature parameter for the entropy and prior terms and defining $s(\mathbf{x}_n, \phi_j)$ as $\log p(\mathbf{x}_n|\delta_{n,j} = 1, \phi_j)$, $L(\mathbf{X}|\theta)$ becomes

$$\begin{aligned} L(\mathbf{X}|\theta) &= \sum_n L_n \\ &= \sum_n [-T \sum_j h_{j,n} \log h_{j,n} \\ &\quad + T \sum_j h_{j,n} \log \pi_j + \sum_j h_{j,n} s(\mathbf{x}_n, \phi_j)] \end{aligned} \tag{24}$$

The maximum likelihood wrt to membership function $h_{j,n}$ subject to the constraint $\sum_j h_{j,n} = 1$, can be solved by the method of Lagrange multipliers. Since each L_n be maximized independently, the Lagrangian becomes

$$\mathfrak{L} = L_n + \lambda \left(\sum_j h_{j,n} - 1) \right) \tag{25}$$

Taking partial derivatives wrt $h_{j,n}$ and λ and equating to 0, we get

$$\begin{aligned} \frac{\partial \mathfrak{L}}{\partial h_{j,n}} &= -T [\log h_{j,n} - \log \pi_j + 1] + s(\mathbf{x}_n, \phi_j) + \lambda = 0 \\ \frac{\partial \mathfrak{L}}{\partial \lambda} &= \sum_j h_{j,n} - 1 = 0 \end{aligned} \tag{26}$$

Solving the first partial derivative for $h_{j,n}$ and eliminating the λ using the second, we get the membership function that maximizes entropy as expected having Gibbs/Boltzmann distribution

$$h^*_{j,n} = \frac{\pi_j e^{s(\mathbf{x}_n, \phi_j)/T}}{\sum_j \pi_j e^{s(\mathbf{x}_n, \phi_j)/T}} \tag{27}$$

Since Nigam et al.'s EM method [1] is a specific application of the general EM for mixture components, and the membership function $h_{j,n}$ for Nigam et al.'s EM method is defined by (9), for simulated annealing EM for document classification, the equivalent of $h^*_{j,n}$ becomes

$$p(y_i = c_j|d_i; \hat{\theta}) = \frac{p(c_j|\hat{\theta}) \exp\{\frac{\log p(d_i|c_j;\hat{\theta})}{T})\}}{\sum_{r=1}^{|C|} p(c_r|\hat{\theta}) \exp\{\frac{\log p(d_i|c_r;\hat{\theta})}{T})\}} \tag{28}$$

ACKNOWLEDGMENT

This research was supported by 1 U24 RR021992 to the Function Biomedical Informatics Research Network (BIRN, http://www.nbirn.net), that is funded by the National Center for Research Resources (NCRR) at the National Institutes of Health (NIH).

REFERENCES

[1] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents using EM," *Machine Learning*, vol. 39, pp. 103–134, 2000.

[2] N. Slonim and N. Tishby, "Agglomerative information bottleneck," *Proceedings of NIPS-12*, 1999.

[3] D. Davies and D. Bouldin, "A Cluster Separation measure," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 1, pp. 224–227, 1979.

[4] T. Joachims, "Transductive inference for text classification using support vector machines," *Proceedings of the International Conference on Machine Learning (ICML)*, 1999.

[5] G. G. Brown, S. Pieper, M. Martone, N. Aucoin, A. Joyner, A. Bischoff-Grethe, and V. Torvik, "The Query Atlas: A Brain Referenced Knowledge Discovery Tool," *Annual Neuroscience meeting*, 2004.

[6] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press, New York, 1999.

[7] C. J. van Rijsbergen, *Information Retrieval*. Butterworths, 1979.

[8] W. B. Croft, S. Cronen-Townsend, and V. Lavrenko, "Relevance feedback and personalization: A language modeling perspective," in *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.

[9] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998.

[10] C. van Rijsbergen, "A theoretical basis for the use of co-occurence data in information retrieval," *Journal of Documentation*, vol. 33, no. 2, pp. 106–119, 1977.

[11] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Proceedings of the 10th European Conference on Machine Learning*, pp. 137–142, 1998.

[12] J. H. Friedman, "On bias, variance, 0/1-loss, and the curse-of-dimensionality," *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 55–77, 1997.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[14] J. C. Dunn, "Well Separated Clusters and Optimal Fuzzy Partitions," *J. Cybernetics*, vol. 4, pp. 95–104, 1974.

[15] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[16] C. G. S. Kirkpatrik and M. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, pp. 671–680, 1983.

[17] E. Levine and E. Domany, "Resampling Method for Unsupervised Estimatition of Cluster Validity," *Neural Computation*, vol. 13, pp. 2573–2593, 2001.

[18] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and Bayesian restoration of images," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[19] S. J. Nelson, W. D. Johnston, and B. L. Humphreys, *Relationships in the organization of knowledge*. Kluwer Academic Publishers, 2001, ch. Relationships in Medical Subject Headings (MeSH), pp. 171–184.

[20] S. K. M. W. Mak and S. H. Lin, *Biometric authentication: A machine learning approach*. Prentice Hall, 2004, ch. Expectation-Maximization Theory, pp. 50–84.