Baum-Welch Style EM Approach on Simple Bayesian Models for Web Data Annotation

Fatih Gelgi, Hasan Davulcu
Department of Computer Science and Engineering
Arizona State University
{fagelgi, hdavulcu}@asu.edu

Abstract

In this paper, our focus will be on weakly annotated data (WAD) which is typically generated by a (semi) automated information extraction system from the Web documents. The extracted information has a certain level of accuracy which can be surpassed by using statistical models that are capable of contextual reasoning such as Bayesian models. Our contribution is an EM algorithm that operates on simple Bayesian models to re-annotate WAD. EM estimates the parameters, i.e., the prior and conditional probabilities by iterating Bayesian model on the given Web data. In the expectation step, Bayesian classifier is trained from current annotations, and in the maximization step, the roles of all the labels are re-annotated to find the best fitting annotation with the current model then the probabilities are re-adjusted from the new annotations. Our experiments show that EM increases the Web data annotation accuracies up to 8%. We use Baum-Welch methodology in our EM approach.

Keywords: *Weakly annotated data, Baum-Welch, Expectation-Maximization, Bayesian Models.*

1. Introduction

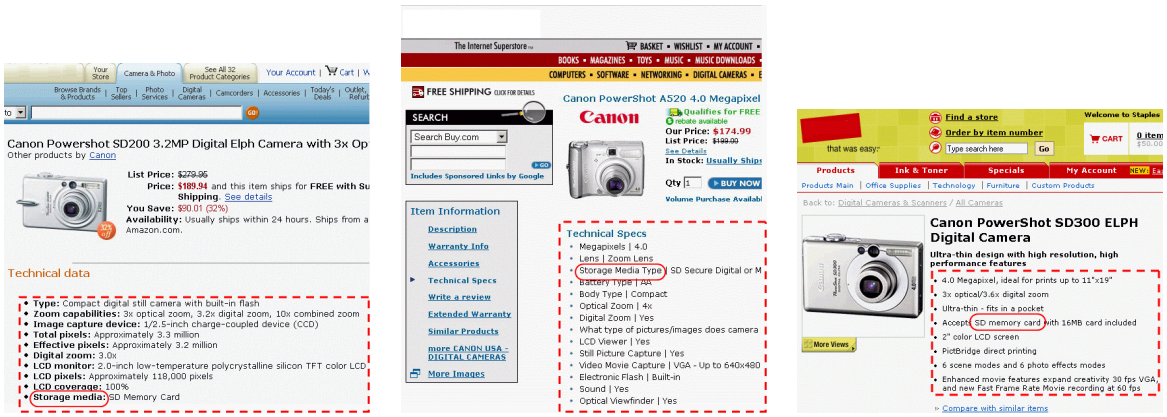
Information extraction (IE) systems have been quite successful in extracting and annotating information given in Web pages that are essentially unstructured. Some examples are (meta) data extraction from template driven Web sites using semi-automated techniques [9] and ontology-driven automated data extraction techniques [7] from text. And also, some work exists on fully automated techniques for extracting (meta) data from data rich segments of Web pages [4, 10, 16].

In this paper, our focus will be on *weakly annotated data* (WAD) which is typically generated by a (semi) automated information extraction (IE) system from the Web documents. In WAD, *annotations* correspond to ontological role assignments such as *Concept*, *Attribute*, *Value* or *Noise*. WAD has two major problems; (i) it might contain incorrect role assignments, and (ii) might have many missing attribute labels between its various entities.

We will use the Web pages in Figure 1 to illustrate WAD that might be extracted using an IE algorithm such as [4, 16]. Each of these pages presents a single instance of the 'Digital Camera' concept. In Figure 1(a), attributes such as 'storage media', and values such as 'sd memory card' have uniform and distinct presentation. However, for an automated system it would be extremely difficult to differentiate the 'storage media type' label as an attribute and 'sd secure digital' as its value due to their uniform presentation in Figure 1(b). On the other hand, in Figure 1(c) the attribute 'storage media' does not even exist, but only its value 'sd memory card' has been reported.

Automated IE systems mostly make local reasoning, i.e., they use single Web pages, or a couple of instances of template-based Web pages. They are lack of using global statistics of the domain incorporating contextual information. Hence, the extracted information has a certain level of accuracy which can be surpassed by using statistical models that are capable of contextual reasoning such as Bayesian models [8]. For example, consider the label 'storage media' marked in Figure 1. A collection of Web pages such as those in Figure 1(a), would yield a strong association between 'canon sd200' as an object and 'storage media' as an attribute. Whereas, the incorrect annotation, extracted from Figure 1(b) would yield a weak association between 'canon a520' as an object and the 'storage media' as a value. Hence, a Bayesian classifier would be able to re-assign the attribute role to the 'storage media' label by using the domain statistics within its context. Similarly, the 'storage media' attribute of the 'sd card' value which is missing in Figure 1(c), could be inferred by utilizing its context and the model.

Bayesian models are capable of making contextual inference on Web data. Meanly, given the context of a label, one can use Bayesian models to infer the role of that label



(a) (b) (c)

Figure 1. The instance pages for Canon digital camera from three different Web sites. Digital camera specifications are marked by dashed boxes in each page. In page (a), attributes are explicitly given as bold whereas in (b) they are not obvious. On the other hand, the attributes are altogether missing in (c).

using the statistics of the entire data and the context of the label. The two problems of Bayesian models are structure and parameter estimation. In our problem, we work on simple Bayesian models, thus the structure is already known. However, parameters, i.e., prior and conditional probabilities can not be correctly estimated due to the partial observability and uncertainty of the Web data. Partial observability and uncertainty is due to the reasons; noisy and incorrect annotations of IE systems, and missing labels and relations between labels. The partial observability and uncertainty of the data lead inaccurate inference of Bayesian models due to the hidden distribution. That's why, the problem can be considered as a good match for an expectation-maximization (EM) algorithm.

In this paper, our contributions are; an EM algorithm that (i) operates on simple Bayesian models to re-annotate WAD, and (ii) follows Baum-Welch methodology. EM estimates the conditional probabilities by iterating Bayesian model on the given Web data. In the expectation step, prior and conditional probability distributions for all roles of labels are computed. In the maximization step, the roles of all the labels are re-annotated using the new Bayesian model to find the best fitting annotation with the current model.

In the rest of the paper, the related work is given in Section 2. Section 3 explains the EM model in details and Section 4 presents the experiments. Section 5 concludes the paper.

2. Related Work

EM models [6, 12, 5] became popular in recent years due to their high performance in many problems with partial observable data. In their recent work, Lowd and Domingos [11] argue that EM Naive Bayes is successful in general probabilistic learning and inference. Furthermore, it is comparable to Bayesian networks with context-specific independence.

Nigam et. al.'s work [13] is one of the main papers that fires the EM naive Bayes research on text and Web documents. Their EM model trains a naive Bayes classifier in E-step and uses it for labeling the unlabeled documents in M-step.

[15] extends the EM naive Bayes model and applies to classification of Web search results. Salvatierra states the incorrectness of the "bag of words" universe of the data assumed in previous work and uses pre-defined ontologies to develop a contextual model. Salvatierra similarly trains a classifier from labeled Web pages and labels the unlabeled ones using EM.

Our work differs at three points: (1) our data is fully annotated and unreliable Web data, (2) we use contextual models based on term associations without predefined ontologies, and (3) we train the simple Bayesian classifiers using the entire unreliable annotation which is similar to the Hidden Markov Model (HMM) training method in Baum-Welch algorithm.

3. Expectation-Maximization Approach

In this section, we first give a brief summary of Baum-Welch algorithm as the background information. Next, the statistical model of the domain which is called relational graph is explained, then the probabilistic framework is given in details.

3.1. Background

Traditional HMMs have three parameters $\lambda = (\pi, A, B)$ where π is the initial probability distribution of the states, A is the probability distribution of state transitions and B is the symbol emission probabilities of states [14]. Input data is the observation sequence $O = O_1 O_2 \dots O_T$ and each observation is considered to be produced from the corresponding state sequence $Q = q_1 q_2 \dots q_T$. Note that the observations are assumed to be *statistically independent*.

Our problem is related with 3. problem for HMMs given in [14]: "How do we adjust the model parameters $\lambda = (\pi, A, B)$ to maximize $P(O|\lambda)$?". $P(O|\lambda)$ can be obtained by summing the joint probability over all possible state sequences Q: $P(O|\lambda) = \sum_Q P(O, Q|\lambda)$.

Solving this problem means to optimize the parameters of HMM so that the model describes the generation of observation sequence best. This method is used when the observation sequences and their corresponding states are given as noisy or uncertain data in partially observable environments. Some of symbols in the sequence might be missing and the corresponding states might be incorrect. Hence the only difference of input data is being *contextual* instead of *sequential* in our case.

Baum-Welch algorithm [3] is a typical EM algorithm that uses the entire data, i.e., observation sequences and their corresponding states, as training data. It re-estimates the parameters and maximizes the number of correct individual states in the data.

We can summarize Baum-Welch algorithm as follows: probabilities are calculated for different annotations in E-step, and the observation sequence is re-annotated in the way that number of correct individual states is maximized and the parameters π, A and B are re-estimated in M-step. Initialization of parameters in the first E-step can be done by using the initial state sequences corresponding to the observations.

3.2. Relational Graph and Formalization

We assume that the IE system we used annotates the labels of a given individual Web page and transforms it into an XML/RDF-like hierarchical structure. The annotations correspond to four ontological roles:

- *Concept* (C): A concept defines a category or a class of similar items. E.g., 'books' and 'digital camera' are some of the concepts in the shopping domain.

- *Attribute* (A): An attribute is a property of an instance or a concept. E.g., 'storage media' is an attribute of the 'canon powershot sd200' instance and the 'digital camera' concept.

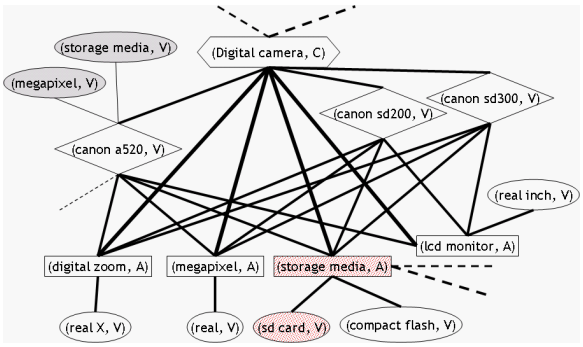


Figure 2. A fragment of the relational graph for the Shopping domain is shown. Each node is composed of a $\langle label, role \rangle$ pair. The thickness of an edge is proportional to the *association strength* between its nodes.

- *Value* (V): A value is a label that provides the value information for an attribute of a certain concept or an instance of a concept. E.g., 'storage media' attribute of the 'canon powershot sd200' instance has the value 'sd memory card'.

- *Noise* (N): Any label that does not have any of the above ontological roles are assigned to be noise. For example, some of the labels in headers, footers or navigation aids, such as 'back to' shown in Figure 1(c) could be annotated as noise.

From these hierarchical structures of the Web pages, we generate the *relational graph* which is a statistical domain model that shows how closely the terms in the domain are related with each other. It has been used to make contextual inference by Bayesian models as demonstrated in the example above. The relational graph is an undirected graph composed of $\langle label, role \rangle$ terms as nodes and their associations as edges. The weight of a node is the count of the corresponding term, and the weight of an edge is the number of parent-child relations of the corresponding nodes in the XML/RDF hierarchies in the entire collection. Formally, assuming w_{ij} as the weight between the terms i and j, and w_i as the weight of the node i, $w_{ij} = w_{ji} = |i \leftrightarrow j|$ and $w_i = |i|$ where $|i \leftrightarrow j|$ and $|i|$ denote the number of times the edge (i, j) and label i appeared in the entire domain respectively. Note that the edges are undirected since association strength between labels is a bi-directional measure. A fragment of the relational graph corresponding to *Shopping* domain is depicted in Figure 2.

For the rest of the paper, we use the following notations:

- A **label** is a word or a phrase in a Web page.

- $\mathcal{R}$ denotes the **set of roles** used for annotating the labels in Web pages.
- An **annotation** is to assign a role to a label.
- A **term** is a pair $\langle O, q\rangle$ composed of a label O and a role $q \in \mathcal{R}$. In other words, terms are annotated labels in the Web pages.
- In this setting, we consider all the labels in each Web page are tagged with roles, hence we define an **annotated Web page** to be a list of its terms. Formally, assuming m labels in a Web page $\mathcal{W} = [\langle O_1, q_1\rangle, \langle O_2, q_2\rangle, \ldots, \langle O_m, q_m\rangle]$.
- The **context** of a label O_j in an annotated Web page $\mathcal{W}_i$ is the set of all the terms in that Web page excluding the terms that contain O_j. That is, $C_j^{(i)} = \bigcup_{T\in\mathcal{W}^{(i)}}\{T\}\setminus\bigcup_{q\in\mathcal{R}}\{\langle O_j^{(i)}, q\rangle\}$.

3.3. Probabilistic Framework

We propose a Baum-Welch style EM model on simple Bayesian models. To demonstrate the methodology, we build EM model on two Bayesian models explained in [8]: the *simple Bayesian classifier* (SBC) and the *naive Bayes classifier* (NBC). These Bayesian classifiers have been used to re-annotate the labels obtained from an IE system given in [16]. Simple Bayesian model structure is already known[1]; EM is used to estimate the prior and conditional probabilities of the Bayesian models in partially observable data.

NBC-EM described here can be considered as a modification of [13]. In this one, NBC classifies the labels in the Web pages instead on Web pages themselves. Additionally, the entire data is used for training the classifier whereas in [13] only the labeled data is used.

For the convenience of the reader, we use a similar notation to [14] for EM to easily relate our methodology to Baum-Welch. In our problem, input data $\mathcal{D}$ is the set of Web pages with annotations. The parameter λ is the role probability distribution of the Bayesian model. The outputs are the corrected annotations of the input data.

By default in EM data, the observations are assumed to be independent from each other. For the Web data, we have two more assumptions as discussed in [8].

Assumption 1 *The role of a label O_j is unique in a context.*

Assumption 2 *The prior probabilities of all the roles of a label O_j are uniform.*

[1]Recall that in Baum-Welch algorithm, the Markovian structure is also assumed to be known which is an HMM and only the parameters are estimated.

The first assumption states that in a specific context, a label can not have more than one roles. The second one is for SBC to reduce the effect of the dominant terms in the collection.

3.3.1 Parameters

The probability of the current data $P(\mathcal{D}|\lambda)$ which is composed of K Web pages is:

$$P(\mathcal{D}|\lambda) = \prod_{i=1}^{K} P(O^{(i)}|\lambda). \quad (1)$$

By the independence assumption, the probabilities of the labels in each Web page can be calculated independently. Given the Web page $\mathcal{W}^{(i)}$,

$$P(O^{(i)}|\lambda) = \sum_{Q} P(O^{(i)}, Q|\lambda) \quad (2)$$

Now, we will explain how to calculate the probability $P(O^{(i)}, Q|\lambda)$ for any given annotation Q.

$$P(O^{(i)}, Q|\lambda) = P([O_1^{(i)}, \ldots, O_m^{(i)}], [q_1, \ldots, q_m]|\lambda) \quad (3)$$

In this formula, q_j is the corresponding annotation for $O_j^{(i)}$. Thus, we write the same formula with the following representation for our convenience:

$$P(\langle O_1^{(i)}, q_1\rangle, \ldots, \langle O_m^{(i)}, q_m\rangle|\lambda) \quad (4)$$

Due to the independence assumption, it is equal to the product of the probabilities of individual terms,

$$\prod_{j=1}^{m} P(\langle O_j^{(i)}, q_j\rangle|\lambda). \quad (5)$$

In this formula, the probability $P(\langle O_j^{(i)}, q_j\rangle|\lambda)$ is calculated by a simple Bayesian model with context dependent reasoning that corresponds to E-step. Each term $\langle O_j^{(i)}, q_j\rangle$ is conditioned on the context C:

$$P(\langle O_j^{(i)}, q_j\rangle|\lambda) = \sum_{C} P(\langle O_j^{(i)}, q_j\rangle|C, \lambda). \quad (6)$$

Since the context of a label $O_j^{(i)}$ is only $C_j^{(i)}$ and fixed in our model, $P(\langle O_j^{(i)}, q_j\rangle|C \neq C_j^{(i)}, \lambda) = 0$. That makes,

$$\sum_{C} P(\langle O_j^{(i)}, q_j\rangle|C, \lambda) = P(\langle O_j^{(i)}, q_j\rangle|C_j^{(i)}, \lambda). \quad (7)$$

Again, by the independence assumption, SBC and NBC given in [8] calculate this probability respectively as follows:

$$\frac{\prod_{T\in C_j^{(i)}} P(\langle O_j^{(i)}, q_j\rangle|T)}{Z} \quad (8)$$

$$\frac{\left(\prod_{T\in C_j^{(i)}} P(T|\langle O_j^{(i)}, q_j\rangle)\right) P(\langle O_j^{(i)}, q_j\rangle)}{Z'} \quad (9)$$

where Z and Z' are normalization factors. From Equation 8, the parameters for SBC to be estimated are in the form:

$$\lambda_{ij}(q) = P(\langle O_j^{(i)}, q\rangle|T) \quad (10)$$

where $q \in \mathcal{R}$ is an arbitrary annotation for the label $O_j^{(i)}$. Similarly, from Equation 9, the parameters for NBC are:

$$\lambda_{ij}(q) = [P(\langle T|O_j^{(i)}, q\rangle), P(\langle O_j^{(i)}, q\rangle)] \quad (11)$$

3.3.2 E-step

In the expectation step, the role probability distribution for each label in a given context is calculated. Formally, $P(\langle O_j^{(i)}, q\rangle|C_j^{(i)}, \lambda)$ in Equation 7 is calculated for any $q \in \mathcal{R}$. This probability computation directly follows from Equations 8 and 9 for SBC and NBC respectively.

3.3.3 M-step

The maximization step reduces to maximizing the number of correct individual annotations of the terms which is similar to maximizing the number of correct individual states as in the Baum-Welch algorithm. Likelihood of the data can be maximized by annotating the labels with their maximum likely roles. Then, the new label-role annotations are used to maximize the parameter λ. For each Web page $\mathcal{W}^{(i)}$,

$$\max_Q P(O^{(i)}, Q|\lambda) = \prod_{j=1} \max_q P(\langle O_j^{(i)}, q\rangle|C_j^{(i)}, \lambda) \quad (12)$$

from the independence of the observation. From Equations 8 and 9, the probability $\max_q P(\langle O_j^{(i)}, q\rangle|C_j^{(i)}, \lambda)$ is:

$$\frac{\max_q \prod_{T\in C_j^{(i)}} P(\langle O_j^{(i)}, q\rangle|T)}{Z} \quad (13)$$

$$\frac{\max_q \left(\prod_{T\in C_j^{(i)}} P(T|\langle O_j^{(i)}, q\rangle)\right) P(\langle O_j^{(i)}, q\rangle)}{Z'} \quad (14)$$

for SBC and NBC respectively. Since maximization step, chooses the role $\bar{q}_j$ for a label $O_j^{(i)}$, the annotation that maximizes the likelihood is:

$$\bar{q}_j = \operatorname*{argmax}_q \prod_{T\in C_j^{(i)}} P(\langle O_j^{(i)}, q\rangle|T). \quad (15)$$

For NBC, it is:

$$\bar{q}_j = \operatorname*{argmax}_q \left(\prod_{T\in C_j^{(i)}} P(T|\langle O_j^{(i)}, q\rangle)\right) P(\langle O_j^{(i)}, q\rangle). \quad (16)$$

As emphasized before, these equations maximize the correct number of individual roles in the data. After the role inferences, all node and edge weights in the relational graph are re-adjusted according to the re-annotations. Next, parameter λ is maximized using the new statistics of the data as follows in SBC:

$$\bar{\lambda}_{ij} = P(\langle O_j^{(i)}, \bar{q}_j\rangle|T) \quad (17)$$

Similarly, conditional probabilities of NBC to be estimated are:

$$\bar{\lambda}_{ij} = [P(T|\langle O_j, \bar{q}_j\rangle), P(\langle O_j, \bar{q}_j\rangle)] \quad (18)$$

Following the methodology in the association rules [1, 2], conditional probabilities are calculated from the current statistics of the entire data:

$$P(T'|T) = \frac{P(T', T)}{P(T)} = \frac{w_{T'T}}{w_T}, P(T) = \frac{w_T}{Z} \quad (19)$$

$$P(T|T') = \frac{P(T, T')}{P(T')} = \frac{w_{TT'}}{w_{T'}}, P(T') = \frac{w_{T'}}{Z'} \quad (20)$$

where $T' = \langle O_j^{(i)}, q_j\rangle$, $w_{T'T}$ and w_T are the weight of the association between terms T' and T, and the weight of the term T respectively. Z and Z' are normalization factors. These weights are retrieved from the relational graph of the current annotation.

Given the label $O_j^{(i)}$, the stochastic constraint of the model which is the summation of the probabilities of all role annotations in a context has to be 1:

$$\sum_{q\in\mathcal{R}} = P(\langle O_j^{(i)}, q\rangle|C_j^{(i)}, \lambda) = 1. \quad (21)$$

The initialization of parameter λ comes from the initial annotations as in the maximization step using Equations 19 and 20. Recall that Baum-Welch similarly uses entire initial state sequences to initialize the parameters in the beginning.

4. Experiments

We used the same two sources and experimental setup in [8] for our data sets to evaluate the efficacy of our algorithms: TAP and CIPS. The first one is a synthetic source whereas the second one is natural Web data.

- **TAP Dataset**: The data set is Stanford TAP Knowledge Base 2 [9] containing the categories AirportCodes, CIA, FasMilitary, GreatBuildings, IMDB, MissileThreat, RCDB, TowerRecords and WHO. These categories alone comprise RDF files of 9,068 individual Web pages.

- **CIPS Dataset**: This data set is composed of faculty, course home pages, shopping and sports Web pages consisting of 225 Web sites and more than 20,000 individual pages. The computer science department Web sites are meta-data-driven, i.e., they present similar meta-data information across different departments. Shopping and sports are some popular attribute rich domains.

In the rest of this section, experimental setup will be explained in details and the results will be discussed in terms of performance of EM, its convergence and the effect of initialization method.

4.1. Experimental Setup

TAP is a synthetic and a template-driven data set that is used to measure accuracy of our algorithm. In the data set, RDF files have been converted into $(Concept, Attribute, Value)$ triples first, then distorted to obtain the inputs. Distortion are applied by deleting the labels and changing the roles of the labels with certain rates in the triples. Two sets of data is prepared from TAP with $(15\%, 15\%)$ and $(35\%, 35\%)$ distortions where the percentages are the random deletion and role change ratios respectively. Each set consists of a mixture of categories. The first set has initially 80% accuracy and the second has 40% accuracy.

In the CIPS data set, we used the semantic partitioner [16] to obtain initial annotations of the labels from Web pages. The semantic partitioner is an IE system that transforms a given Web page into an XML-like structure by separating and extracting its meta-data and associated data. For the evaluations, we created a smaller data set containing randomly selected 160 Web pages from each domain. We divided samples into 4 groups with 40 pages from each category, and each group is evaluated by a non-project member computer science graduate student.

In our experiments, data has been preprocessed using simple regular expressions to identify the common data types of values such as percentage, dates, numbers etc. which is a traditional method in Web mining. The reported accuracies are measured according to the formula $Accuracy = \frac{\#\ of\ correct\ annotations}{\#\ of\ total\ annotations}$.

In the rest of this section, experimental setup will be explained in details and the results will be discussed in terms of performance of EM, its convergence and the effect of initialization method.

4.2. Results and Discussions

Our EM model is evaluated with three experiments. The first one demonstrates the quality performance of EM. The final accuracies of EM models have been compared with simple Bayesian models without EM. The results are given in 3. NBC-EM and SBC-EM are the EM models of NBC and SBC. The effect of EM on SBC is in the interval of $[2\%, 7\%]$ whereas in NBC it is $[3\%, 8\%]$. EM has been well performed especially in *Sports* and *TAP(35%,35%)* data sets since their initial annotations have lower accuracy which means more instability compared to the other data sets.

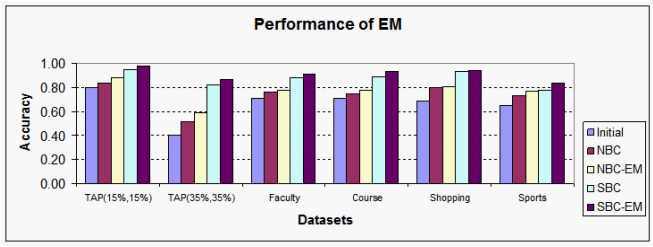


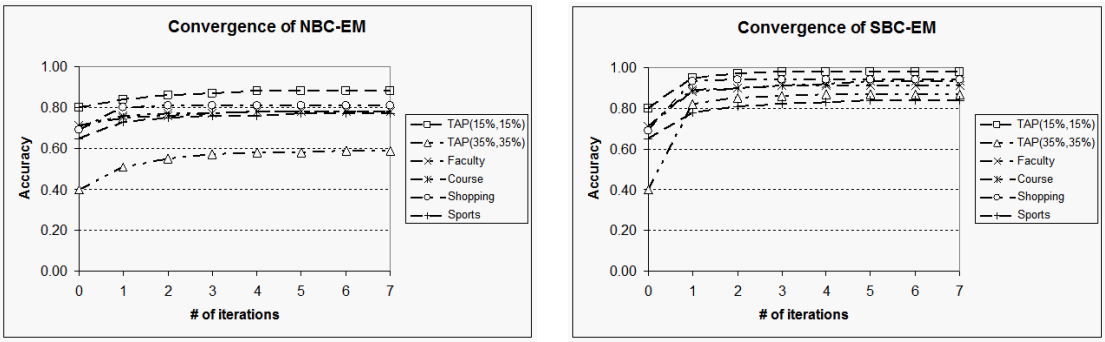
Figure 3. Performance of EM on different Bayesian Models.

Convergence of EM is tested in the second experiment. Figure 4 shows the convergence performance of EM on the same data sets. For both EM models, 7 iterations have been sufficient for convergence. As expected, the big leap has been in the first iteration and models have stably converged following logarithmic curves which is the usual characteristics of EM models.

The third experiment evaluates the effect of initialization of EM models. In [14], it is indicated that good initial estimates of parameters effects the performance. Figure 5 compares random initialization to the initialization with an IE system. In Figure 5(a), the accuracy of EM with random initialization is the average of 20 runs. The variation in the initial and final accuracies is limited in the interval of 7%. Although, there is not marginal difference in the accuracy increase, the final accuracies are substantially different. Since the likelihood function has many local maximums, EM with random initialization has been stuck in one of the local maximum which has very low accuracy; less than 50% for all data sets.

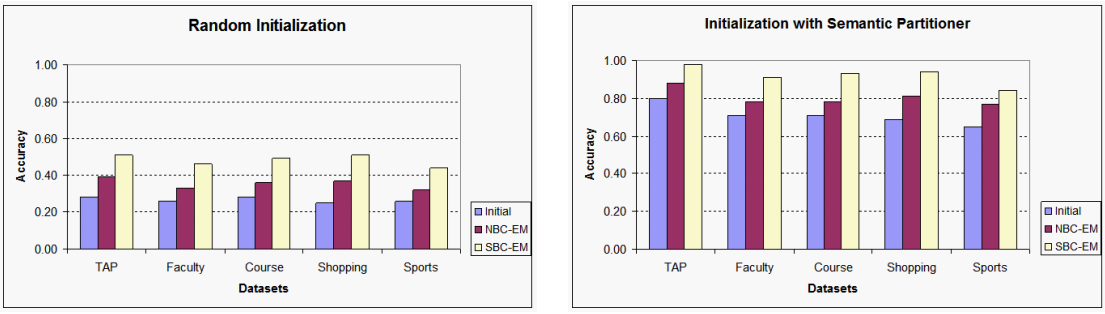
5. Conclusion and Future Work

In this work, we present a EM methodology on simple Bayesian models and exemplify on a prior work [8]. Our work is inspired from the Baum-Welch algorithm which is essentially an EM model on HMMs to estimate its parameters. As we demonstrate in the experiments, EM increases the accuracy of the data from 2% to 8%. As the future work, we consider generalizing the approach for Bayesian networks with arbitrary structures.



(a) (b)

Figure 4. Convergence of EM.



(a) (b)

Figure 5. Effect of the initialization method: (a) random initialization, (b) initialization with Semantic Partitioner.

References

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD*, pages 207–216, Washington, D.C., 1993.

[2] E. Alpaydin. *Introduction to Machine Learning*, chapter 3, pages 39–59. MIT Press, 2004.

[3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in statistical analysis of probabilistic functions in markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[4] V. Crescenzi and G. Mecca. Automatic information extraction from large web sites. *Journal of ACM*, 51(5):731–779, 2004.

[5] F. Dellaert. The expectation maximization algorithm. Technical Report GIT-GVU-02-20, Georgia Institute of Technology, 2002.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39.

[7] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, K. S. McCurley, S. Rajagopalan, A. Tomkins, J. A. Tomlin, and J. Y. Zien. A case for automated large-scale semantic annotation. *Journal of Web Semantics*, 1(1):115–132, 2003.

[8] F. Gelgi, S. Vadrevu, and H. Davulcu. Fixing weakly annotated web data using relational models. In *ICWE*, 2007.

[9] R. Guha and R. McCool. TAP: A semantic web toolkit. *Semantic Web Journal*, 2003.

[10] K. Lerman, L. Getoor, S. Minton, and C. Knoblock. Using the structure of web sites for automatic segmentation of tables. In *ACM SIGMOD*, pages 119–130, Paris, France, 2004.

[11] D. Lowd and P. Domingos. Naive bayes models for probability estimation. In *ICML*, pages 529–536, 2005.

[12] T. Minka. Expectation-maximization as lower bound maximization, 1998.

[13] 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.

[14] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[15] S. M. Salvatierra. Using unlabeled data to improve classification in the naive bayes approach: Application to web searches. *Journal of Computational Methods in Science and Engineering*, 4(1-2):65–74, 2004.

[16] S. Vadrevu, F. Gelgi, and H. Davulcu. Semantic partitioning of web pages. In *WISE*, 2005.