ISSN 2320 - 2602
Volume 2, No.8, August 2013
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2013/ijacst04282013.pdf

TAXONOMY AWARE CATALOG INTEGRATION WITH HIDDEN MARKOV MODEL (HMM)

P.Hema Priya[1], R. RangaRaj[2]
[1]Research Scholar, Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, India. priyapvc@gmail.com
[2] Head, Department of Computer Science, Hindusthan College of Arts and Science, Coimbatore, India. rraj75@rediffmail.com

ABSTRACT

Integration of data is the major important task for online ecommerce based web portals and commerce search engine based application. The integration of task faced by online commercial portals and e-commerce search engines are the integration of products coming from multiple providers to their creation of product catalogs. Cataloging of products from the data provider into the master taxonomy and while formulate use of the information provider taxonomy data become major problem. Conquer this difficulty classify the products based on textual based classifier and taxonomy-aware step with the purpose of adjust the outcome of a textual based classifier to make sure that products that are close as one in the provider taxonomy. In taxonomy aware calibration process base classifier that is text base classifier update the results based on parameters values specified at calibration step, but these values are manually given by user, at a halt it becomes major issues to identify candidate products for category the products, to conquer these problem Proposed a HMM based machine learning method to derive the parameter values automatically and continuously retrain the base classifier with fundamentals chosen during the taxonomy-aware calibration step. HMMs allow you to estimate probabilities of unobserved products in product categorization. In our approach estimate the probability between the categorized products and unobserved products .Unobserved categorization products are derived from the Transitional probability, observation probability products are used to find the observed or categorized products in the system .In this machine learning algorithm involves a large quantity of unlabeled product data with only a small number of labeled product data. It finds the each candidate parameter θ_i and after that get the optimal parameter γ from that the validation set is accuracy is maximized. An experimental result shows that the HMM based machine learning systems are efficient and thus appropriate to the huge data sets to be representative on the web.

Keywords: Catalog integration, Text based classifier taxonomies, HMM based machine learning system.

1. INTRODUCTION

In ever-increasing numeral of web portal give a user knowledge centered on online shopping. This contains e-commerce site such as Amazon and Shopping. com and commerce search engines such as Google Product Search and Bing Shopping. Generally this engine creates a product catalog to identify the present status of the products and number of user in their environment .For this main purpose data interaction step managed by these commercial portals is the incorporation of data upcoming from several data providers into a particular product catalog. It is named as product categorization. Each and every web portals keep up their own master taxonomy for organizing products and it is used for mutually browsing and searching purposes .In this process if a new products appear from the disparate providers, it routinely categorize the products in master taxonomy according to their users. But in website surroundings it becomes complicated to allocate the products from their catalog to the suitable category in the master taxonomy. Therefore we necessitate automated technique for categorizing products upcoming from the data providers into the master taxonomy.

A significant examination in this scenario is to the data providers do include their own provider taxonomy and their products are previously coupled with a provider taxonomy group. The provider taxonomy may be dissimilar beginning the master taxonomy, but in the majority of cases present is still an influential signal coming from the provider classification. In this scenario product taxonomy representation the products with the aim of nearby categories are supposed to be classified into nearby categories in the master taxonomy. To demonstrate this point considers the illustration in Fig. 1. The provider taxonomy is an extract from by using

Amazon and the master taxonomy is an extract by using Bing Shopping. Now specified a product tagged through a category from provider taxonomy, it also would like to categorize it in the master taxonomy. For example a specified the product "Boss Audio Systems CH6530" from the category of Electronics/Car Electronics/ Car Video/Car Speakers/Coaxial Speakers in the provider taxonomy. To categorize these products into the master taxonomy we need a text base classifier, because the existing work is undecided whether this product must be classified into Electronics/Car Electronics /Car Audio/Car Speakers or Electronics/Home Audio/Speakers.

First use the text based classifier to adjust the results of taxonomy information. The text based classifier representation makes use of the taxonomy construction of the master and provider taxonomies in order to attain associations amongst the disparate categories in the taxonomy. Text based classifier the labeling problem occurs related to a diversity of optimization problems such as the metric labeling problem [1] or structured prediction problems [2]. These problems are known to be NP-hard when asking for a hard labeling of the data. For certain variant there are estimation algorithms [3], [4], [1], [5]. Optimization problem can be formulate as an Integer Linear Program (ILP) or a Quadratic Integer Program (QIP), still the numeral of variables is comparative to the numeral of products in the basis catalog, which is prohibitively large. To the best of our knowledge, not any of the solution that has been proposed is appropriate to web-scale categorization problems; anywhere the numeral of products to be classified is capable of order hundreds of thousands. In the text based classifier when new products arrives in the web portals or ecommerce applications categorization of the products at the base classifier becomes fails, because the parameters chosen at the base classifier are manually derives the value it is not match in all types of categorization products and labeling also becomes difficult because it does not provide any additional help out, while there is no noticeable mapping among categories at the leaf level. To conquer these problems we proposed a HMM based classifier to incrementally retrain the base classifier with elements chosen during the taxonomy-aware calibration step. In this step the classifier calculate the both observed categorization of products and unobserved categorization of products at taxonomy calibration step. Based on the probability values derived from the HMM learning only finds the parameters values and then finally classify the records or product information.
The major contributions of the effort as follows:

1. First derive the taxonomy-aware catalog integration complexity as a structured prediction problem. In this way the technique that leverages the construction of the taxonomies to categorize catalog integration.
2. Subsequent explain the taxonomy aware classification process with two ways:

- In first step product are classified under base classification step
- After that use taxonomies aware processing steps. During the taxonomy aware classification step the optimization problem or label classification problem have been overcome with TACI algorithm.

3. Incrementally retrain the base classifier with elements chosen during the taxonomy-aware calibration step we proposed a HMM based machine learning methods for best classification results .It select the best classifier b on the products and further it can be used for validation set process.

4. Finally evaluate the experimental results on real-world data and compare taxonomy-aware classification, proposed HMM based parameter calibration, it provides a considerable perfection in accuracy over existing Scalable algorithm.

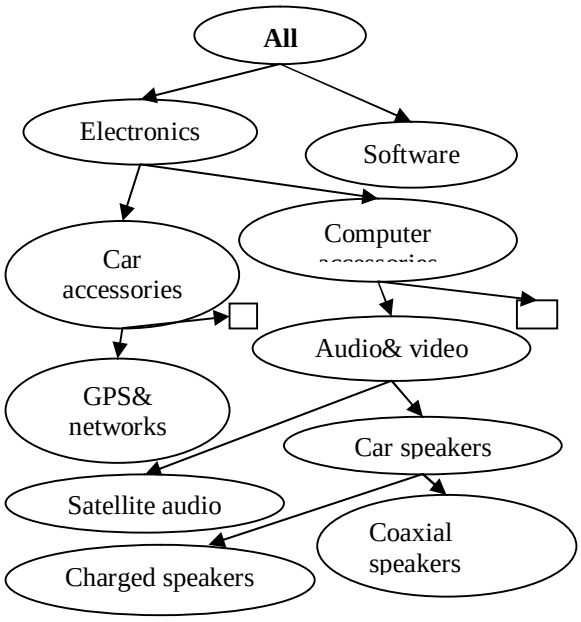

Figure 1: Provider taxonomy (Amazon)

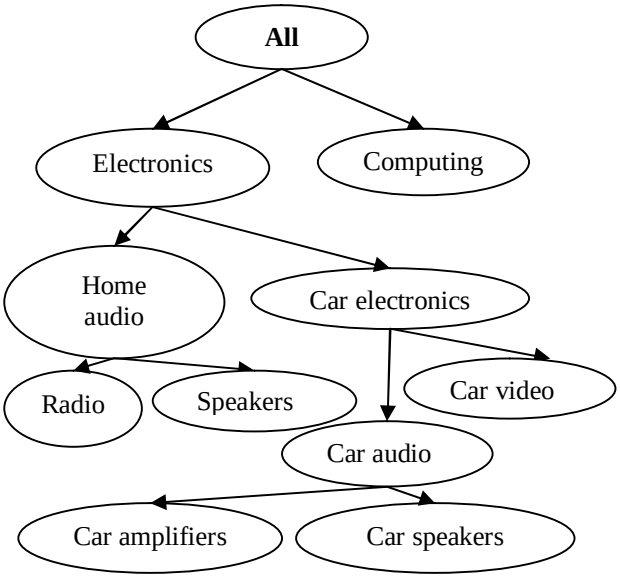

Figure 2: Master taxonomy (Bing)

2. RELATED WORK

Initially the major part of the work can be done with help of the background study papers and understand how the problem occurs and steps or way to solve the problem by using the datamining methods or classification ,categorization steps. So in this section study different catalog integration optimization problem and method follow to resolve the catalog integration problem as well as Metric labeling problem occurs at the classification step, finally study the structured prediction. In all of the previous work make use of source category information, except treat the source and target taxonomies as horizontal.

Pervasive web portal problem studied by R. Agrawal and R. Srikant et .al [6]. It repeatedly processes the product catalog to construct the base classifier for product integration of documents in the master catalog for predicts the category of unidentified documents. Our result on the way is that many of the data sources have their have control of categorization and classification accuracy can be enhanced by factoring in the implied information in the basis categorizations.

Semisupervised learning based categorization of documents from source to destination documents was introduced by Sarawagi et al [7] with cross training model. In this model if classify the documents with semi-supervised learning for document classification occurrence of multiple label sets. Document classification is an entrenched section of text mining. A text classifier is creative trained using documents with preassigned labels or classes selected from a set of labels it is called as taxonomy or catalog. Formerly the classifier is trained; it is obtainable test documents for which it necessities guess the best labels.

Learning based boosting algorithm was introduced by Zhang and Lee have also developed approach to catalog integration by using boosting [9] and transductive learning [8] [10]. While these approaches attain improved categorization accuracy, comparable to the cross-training move toward, they require training data that are labeled in equally the source and the target taxonomies. From this point these approaches are not applicable to our difficulty situation.

Nandi and Bernstein [20] propose an approach for matching taxonomies based on query term distributions. The approach is quite different to ours. First, it performs the mapping at the taxonomy level, mapping categories from the source to the target, while we perform the mapping at the instance level by categorizing individual product instances to the target taxonomy. Second, the approach is not based on classification but rather on exploiting distributions of terms associated with the categories

Matching the correct product into the sub product for product categorization ontology matching methods was proposed in previous work and alignment representation was also proposed in previous work. Glue [11] proposed a machine learning approach to learn the products and hoe to categorization of the product among the products with ontology matching and alignment of the product.

Iliad's [12] also introduces a machine learning and logical inference approach to production alignments. In universal the focal point in ontology alignment is to plan nodes of source taxonomy to nodes of target taxonomy. In difference metrics the similarity of the scheme not concerned in solving the alignment difficulty between taxonomies, but somewhat specified an instance the goal is to categorize it in the target taxonomy with aids from the taxonomy structure. The end purpose is evermore the classification of the product. This dissimilarity is very important in many realistic scenarios.

Catalog integration problem is defined as optimization problem it is motivated by the metric

labeling problem that was introduced by J. Kleinberg and E. Tardos [1]. It determines the optimal labeling from numeral of objects accordingly that they decrease an assignment and a separation cost. This type of problem is NP-hard problem and the dissimilar accessible predictable solution can be preparing it as an LP [1] or a QP [5]. The difficulty of all these method makes them inappropriate to large-scale data sets with further than an only some hundreds of products. The reason of our optimization problem is also equivalent to the objectives that arise in processor vision problems [3], [12], [13]. Nandi and Bernstein [14] recommend an approach for corresponding taxonomies based on query term distributions. Primary it perform the mapping at the taxonomy level, mapping category from the source to the target, while we achieve the mapping at the occurrence level by categorizing personality product instances to the target taxonomy. Following the approach is not based on classification but rather on exploiting distributions of terms associated with the categories.

P. Ravikumar and J. Lafferty [5] are proposed a quadratic programming representation that represents a substitute to linear program relaxations and tree reweighted belief proliferation for the metric labeling or MAP assessment difficulty. An extra convex relaxation of the quadratic estimate is exposed to contain preservative approximation guarantee that relate even while the graph weights contain mixed sign or do not come from a metric. The approximations are comprehensive in a way that allows tight variational relaxations of the MAP difficulty, even though they normally engage non-convex optimization.

3. BASIC OF TERMS FOR TAXONOMY-AWARE CATAOG INTEGRATION AND HMM ALGORITHM FOR PARAMETER CALIBRATION

In this section first define the basic terms for product categorization and then from that create the taxonomy-aware catalog integration difficulty. The terms of the product is defined as p item that can be buying at a marketable portal. Every product has a textual demonstration that consists of a name of the product and possibly a set of attribute-value pair. The example was shown in Fig. 1 product name is “Boss Audio Systems CH6530”, and description attribute with value “Chaos Series 6.5-Inch 3-Way Speaker, 300W peak power.” Reminder that the name and the attributes of a product may vary across providers.

In this step the product taxonomy can be represented as a Graph G= {V, E} with a directed acyclic graph (DAG) whose nodes Cg characterize the set of possible category into which products are prearranged. Every graph in an edge (C_1, C_2) ∈ Eg represents a subsumption association between two categories C_1 and C_2 .A product catalog $\mathcal{K} = \{P, G, V\}$ is a taxonomy G populated with a set of products P as defined by the mapping function v : P → Cg that maps each product in P to a category in Cg. Since v is a function, we assume that each product is related with accurately one category, though our work is able to useful to suitcases where this supposition does not hold.

After that precedes the step toward the direction of to taxonomy aware categorization with two step process. Most important each product is classifying use a base classifier without aware of the taxonomies. Subsequently uses the formations of the source and target taxonomies in normalize to correct the output of the base classifier and create a concluding classification. It is named as the taxonomy-aware processing step.

3.1 Base classifier for taxonomy aware categorization

In the base foundation classification step categorize the products based on their textual manifestation. For this reason, prepare a text-based classifier with supervised machine learning approaches such as Naive Bayes (NB), and Logistic Regression (LR). We make use of a separation of the target list as the training set. These provide us among example of products labeled with category of the target taxonomy. The features of the classifier are extracting from the textual product representation. Note that at preparation time we don’t have any knowledge of the providers’ catalogs, and we make no use of the construction of the target taxonomy. Let b indicate the classification representation after training process competition. Given a product p ∈ P_s from the provider catalog we apply the classifier b on the textual representation of the product, as this appears in the provider’s catalog

3.2 Taxonomy-Aware Processing flow

After that classification process then the taxonomy-aware processing step is to categorize the target taxonomy results from the classification process by taking into description the associations of the products in the source and target taxonomies. The major problem occurs this categorization is labeling to handle all the products in efficient manner, to

conquer these talk the diverse parameters of the problem. It is defined from the given a source catalog $\mathcal{K}_s$, and a target catalog $\mathcal{K}_t$the objective is to find a labeling vector that minimize the subsequent cost function:

$COST(\mathcal{K}s, \mathcal{K}t, \ell) =$
$(1-Q)\sum_{p\epsilon P_s} A\ Cost(p, \ell_p) + Q\sum_{p,q\epsilon P_s} S\ Cost(p, q, \ell_p, \ell_q)$ ->(1)

The taxonomy-aware procedure f_T is the algorithm that finds the labeling ℓ that minimizes the cost function:

$$f_T(\kappa_S, \kappa_t) = \arg\min_{\ell} COST(\kappa_S, \kappa_t, \ell) \rightarrow (2)$$

To classify the products from the base classifier b compute the probabilities of the base classifier to describe the task of cost function. A COST: $P_s * C_t \rightarrow \mathbb{R}^+$ .For a product x the cost of classifying product x to objective category ℓ_x is defined as follows:

$$A\ Cost(p, \ell_p) = 1 - Pr_b(\ell_p|\ell_q) \rightarrow (3)$$

Important similarity description is supposed to assure the perception those two categories that are close up in the taxonomy tree are more comparable than two categories that are far separately. For example, two categories that have a general parent are further similar than two categories that have dissimilar parents and an ordinary grandparent. The division cost as a function of the similarity $sim_S(s_p, s_q)$between categories and of p and q in the source taxonomy S and the similarity

$$S\ Cost(p, q, \ell_p, \ell_q) = \delta\left(sim_S(s_p, s_q), sim_T(s_p, s_q)\right) \rightarrow (5)$$

$\sum_{p,q\in P_s} S\ Cost(p, q, \ell_p, \ell_q)$=

$\sum_{\sigma,\bar{\sigma}\ \epsilon c_S}\sum_{\tau,\bar{\tau}\ \epsilon c_S} SCOST(\ \sigma, \bar{\sigma}, \tau, \bar{\tau})n(\sigma, \tau)\ n(\bar{\sigma}, \bar{\tau})$- >(6)

Optimization difficulty have be occur in all of the above mentioned steps, to overcome these problems , scalable algorithm for the taxonomy-aware categorization step to large data sets. Even though present our method with respect to our exact problem. It can be applied to other prearranged prediction problems in arrange to deal with the quadratic numeral of pairwise relationships. To perform this process using search pruning methods and then proceed calibration step to categorize the master and product taxonomy. Search Space Pruning presents a heuristic for proficiently performing arts the taxonomy-aware calibration step. The idea is to thoughtfully fix the group or category for a number of products in the foundation catalog in order to achieve a landscape of the mappings among the two taxonomies. From this define the subset of products that categorize the products .Let $\theta\epsilon$ [0,1] be a threshold value that define while the category probability approximation returned by the base classifier is great enough therefore that the predicted category is expected to be accurate. Let F_θ be the subset of products that pass the threshold is defined as,$F_\theta = \{p \in P_s | \max_{\gamma\epsilon C_t} Pr_b[\tau|p] \geq \theta\}$->(7)

$$\ell_p = \arg\max_{\gamma\epsilon C_t} Pr_b[\tau|p] \rightarrow (8)$$

Let $O_\theta = P_s / F_\theta$ denote the products whose classification remains open. Each open product $\in O_\theta$ autonomously and calculate a division cost for only with respect to the fixed products in F_θ .If s_p is the source category of p and t_p is a candidate target category, then the separation cost for this source-target pair is defined as follow:

$$h(s_p, t_p) = \sum_{\sigma\epsilon S, \tau\epsilon T} S\ COST(S\ Cost(s_p, \sigma, t_p, \tau)\bar{n}(s_p, t_p)\overline{n(\sigma, \tau)} \rightarrow (9)$$

Algorithm: TACI algorithm

Input: Source catalog s , Target Taxonomy T, base classifier b and parameters θ, k, γ

Output: Labeling vector ℓ

1. $F_s \leftarrow \theta$
2. For all $p \in P_s$ do
3. $\tau^* \leftarrow \arg max_{\tau\epsilon C_t}, \max_{\gamma\epsilon C_t} Pr_b[\tau|p]$
4. if $Pr_b[\tau^*|x] \geq \theta$ then
5. $\ell_p \leftarrow \tau^*$
6. $F_\theta \leftarrow F_\theta \cup \{p\}$
7. Else
8. $O_\theta \leftarrow O_\theta \cup \{p\}$
9. Compute $TOP_k(p)$
10. Compute candidate pairs $H_{\theta,k}$
11. Initialize hash table HT to empty
12. For all $(\sigma, \tau)\epsilon H_{\theta,k}$ do
13. HT(σ, τ)=H(σ, τ))
14. For all $p \in O_\theta$ do
15. $\ell_p \leftarrow$ $\arg\min_{\tau\epsilon \mathrm{TOPk}(p)}\{(1-\gamma)A\ COST_{\ p,\tau+\gamma HT(s_p,\tau)}\}$

3.3 Parameter calibration The fine-tuning of the parameter k, θ, and γ is significant for the performance of our algorithm. The validation set consists of products that are cross labeled in both the source and the target taxonomy. Base classifier training that involves tens of millions of features, while it is big enough to tune few parameters of the TACI algorithm. The first parameter set is parameter k, such that the accuracy of the classifier over the top-k categories is high. After that tune the parameters θ, and γ. For each candidate parameter we discover the "optimal" parameter γ such that the correctness of the TACI algorithm on the validation set is maximized. Let know all the parameters that are preferred such as to make the most of the accuracy in TACI algorithm on the validation set.

3.4 HMM based learning algorithm

In generally the knowledge methods can be separated into supervised and unsupervised learning methods. The supervised learning methods learner aims at evaluation of the input –output relationship by using objective function with training set data set for regression tasks and solves the classification problems .In unsupervised learning only the raw data x_i are available, not including the consequent labels y_i. It becomes difficult to handle the unlabeled data , to handle this situation where some labeled patterns are provided jointly with unlabeled ones arise frequently. It is called semi supervised learning. An HMM can be characterized by the following:

1. The number of states in the various categories of the products categorizes of products such as audio, video types in car N. The set of states is $S = \{S_1, S_2, \ldots S_N\}$, where S_i, $i = 1, 2, \ldots, N$ is an individual state.

2. The number of dissimilar observation symbols per state is M. The set of symbols is $V = \{V_1, V_2, \ldots V_3\}$, where V_i, $i = 1; 2; \ldots; M$ is an individual symbol.

3. The state transition probability matrix $A = [a_{ij}]$ where $a\_ij = P(q_t + 1 = S_j \mid q_t = S_i)$; $1 \le i \le N$; $1 \le j \le N$; $t = 1, 2, \ldots : N$ where $a_{ij} > 0$ for all i, j. Also, $\sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N$.

4. The observation product symbol probability matrix $B = [b_j(k)]$, where $b_j(k) = P(V_k \mid S_j), 1 \mid \le j \le N, 1 \le k \le M$ and $\sum_{k=1}^{M} b_j(k) = 1, 1 \le j \le N$

5. The initial state probability vector $\pi = [(\pi_i)]$, where $\pi_i = P(q_1 = S_i), 1 \le j \le N$, such that $\sum_{k=1}^{M} \pi_i = 1$

6. The observation sequence $O = O_1, O_2, O_3 \ldots O_R$, where each observation O_t is one of the symbols from V, and R is the number of observations from the products in the series.

It is clear that an absolute condition of an HMM requires the evaluation of two model parameters, N and M, and three probability distributions A, B, and π. We make use of the notation $\lambda = (A, B, \pi)$ to designate the complete set of parameters in the products categorization model, where A, B implicitly include N and M. An observation sequence O, as mentioned above, can be generated by many probable state sequences. Consider one such particular series $Q = q_1, q_2 \ldots \ldots q_R$ Where q_1 is the initial state.

The probability that O is generated from this state sequence is given by ,

$$P(O \mid Q, \lambda) = \prod_{t=1}^{R} P(O_t \mid , q_t \lambda)$$

where statistical independence of observations is assumed. Above equation can be described as

$$P(O \mid Q, \lambda) = b_{q_1}(O_1).\, b_{q_2}(O_2) \ldots \ldots b_{q_R}(O_R)$$

The probability of the state series Q is given as $P(Q \mid \lambda) = \pi_{q1} a_{q1q2} a_{q2q3} \ldots . a_{qR-1R}$

Thus, the probability of creation of the observation series O by the HMM specific by λ can be defined as an equation as follows:

$$P(Q \mid \lambda) = \sum_{\text{all } Q} P(O \mid Q, \lambda) P(Q \mid \lambda)$$

Deriving the value of $P(Q \mid \lambda)$ using the direct description of above equation is divisionally exhaustive. Consequent to the HMM parameters are exposed; we take the symbols from a product training data and form an initial state series of symbols. Let $O_1, O_2, O_3 \ldots O_R$ be one such series of length R. This observation series results is formed from the product categorization .They produce this input sequence to the HMM and compute the probability of acceptance in training stage calculated by HMM. Let the probability θ can be formulated as follows:

$$\theta = P(O_1, O_2, O_3 \ldots O_R \mid \lambda)$$

Providers	NB	LR	TACI-NB	TACI-LR	TACI-HMM
Amazon	77.2	70.1	82.2	76.3	85.78
Etilize	75.1	80.1	82.7	91.74	93.2
pricegrabber	41.5	48.5	73.2	75.5	85

Table 1: Classification Accuracy

This probability computation was performed for all θ and maintained these results. If $\theta > 0$ the new sequence is verified and accepted by the HMM with low probability, and it could be a categorization by the way of if the percentage change in the probability is above a threshold, that is,

$$\theta \geq \text{Threshold}$$

4. EXPERIMENTAL RESULTS

Before concluding the results of the existing and proposed HMM system results the major part is to compare the accuracy of the system in terms of the classification performance evaluation and time taken to complete the product categorization in web portal environment, commercial search engine applications.

4.1 Classification Accuracy

Finally in this section measure the classification accuracy of the taxonomy classification step for TACI algorithm and TACI with HMM learning at calibration step. The results show the benefits of our taxonomy-aware calibration step and compare the taxonomy-aware algorithm. Three different providers such as Amazon, Etilize, and Pricegrabber are used to measure the classification accuracy of both master taxonomy and provider taxonomy .The use as master catalog of Bing Shopping, which cumulative data feeds from retailers, distributors, resellers, and other profitable portals.

In all the experiments, consider a target taxonomy that consists of all the categories in Bing Shopping taxonomy that is related to consumer electronics. Measure the classification accuracy Naive bayes(NB),Linear Regression(LR), Taxonomy-Aware Catalog Integration with Naive bayes(TACI-NB) , Taxonomy-Aware Catalog Integration with Linear regression(TACI-LR) and Taxonomy-Aware Catalog Integration with HMM(Hidden Markov Model). The outcome for all algorithms more than all data sets are shown in Table 1 and the resultant figure are also shown in Figure 2.

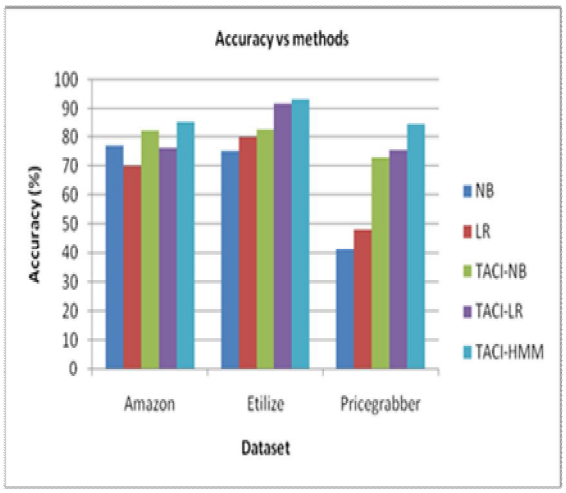

Figure 3: Classification Accuracy Evaluation

4.2 Time Comparison

In this section, we compare the time comparison of the different approaches for catalog integration. The results for all algorithms over all data sets are in Table 2 and the corresponding figure are shown in Figure 3.

Table 2: Time comparison Accuracy

Providers	NB	LR	TACI-NB	TACI-LR	TACI-HMM
Amazon	852	829	745	675	589
Etilize	92	86	71	55	65
pricegrabber	452	420	370	320	249

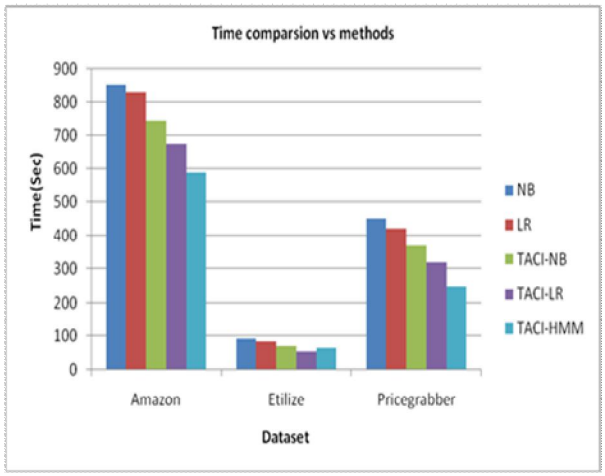

Figure 4: Time comparison Accuracy Evaluation

5. CONCLUSION

In this research, it has an well-ordered learning method to catalog integration with the aim to use of basis category and taxonomy organization information. The proposed HMM based learning algorithm were used for retrain the base classifier during the product calibration step, they can also be used for solve the other problems such as product categorization to which category in the catalog. The yield of the parameter outcome as choosen might be used as an attribute for item equal, whereas would like to match elements classified under the master taxonomy to received offer on or after the providers. Experimental results also showed that was leads to considerable accuracy with value than the presented calibration step based classifier.

REFERENCES

1.J. Kleinberg and E. Tardos, "Approximation Algorithms for Classification Problems with Pairwise Relationships: Metric Labeling and Markov Random Fields," J. ACM, vol. 49, no. 5, pp. 616-639, 2002.

2. G. Bakir, T. Hofmann, B. Schlkopf, A. Smola, B. Taskar, and S. Vishwanathan, Predicting Structured Data. MIT Press, 2007.

3. Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 23, no. 11, pp. 1222-1239, Nov. 2001.

4. C. Chekuri, S. Khanna, J.S. Naor, and L. Zosin, "Approximation Algorithms for the Metric Labeling Problem via a New Linear Programming Formulation," Proc. 12th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA), pp. 109-118, 2001.

5. P. Ravikumar and J. Lafferty, "Quadratic Programming Relaxations for Metric Labeling and Markov Random Field Map Estimation," Proc. 23rd Int'l Conf. Machine Learning (ICML), pp. 737-744, 2006.

6. R. Agrawal and R. Srikant, "On Integrating Catalogs," Proc. 10th Int'l Conf. World Wide Web (WWW), pp. 603-612, 2001.

7. S. Sarawagi, S. Chakrabarti, and S. Godbole, "Cross-Training: Learning Probabilistic Mappings between Topics," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Datamining (KDD), 2003.

8.D. Zhang and W.S. Lee, "Web Taxonomy Integration through Co- Bootstrapping," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 410-417, 2004.

9.D. Zhang and W.S. Lee, "Web Taxonomy Integration Using Support Vector Machines," Proc. 13th Int'l Conf. World Wide Web (WWW), pp. 472-481, 2004.

10 .D. Zhang, X. Wang, and Y. Dong, "Web Taxonomy Integration Using Spectral Graph Transducer," Proc. ER Workshop, pp. 300- 312, 2004.

11. A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy, "Learning to Match Ontologies on the Semantic Web," The VLDB J., vol. 12, no. 4, pp. 303-319, 2003.

12. Y. Boykov, O. Veksler, and R. Zabih, "Fast Approximate Energy Minimization via Graph Cuts," IEEE Trans. Pattern Analysis Machine Intelligence, vol. 23, no. 11, pp. 1222-1239, Nov. 2001.

13. V. Kolmogorov and R. Zabih, "What Energy Functions can be minimized via Graph Cuts?" IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 2, pp. 147-159, Feb. 2004.

14. A. Nandi and P.A. Bernstein, "Hamster: Using Search Clicklogs for Schema and Taxonomy Matching," Proc. VLDB Endowment, vol. 2, no. 1, pp. 181-192, 2009