Semi-supervised Taxonomy Aware Integration of Catalogs

JOTHI PRAKASH. V, NITHYA L.M.
Department of Information Technology
SNS College of Technology,
SNS Kalvi Nagar, Kurumbapalayam, Coimbatore, Tamil Nadu.
INDIA.
jothiprakashv@gmail.com

Abstract: - The main task of online commercial portals and business search engines is the integration of products coming from various providers to their product catalog. The commercial portal has its own master taxonomy while each data provider classifies the products into provider taxonomy. Classification of products from the data provider into the master catalog by using the data provider's taxonomy information is done by classifying the products based on their textual representations by using a simple text based classifier and then using the taxonomy information to adjust the results of the classifier to make sure that the products that are tied together in the provider catalog remain close in the master catalog. The taxonomy aware calibration takes place by tuning the values of three parameters k, θ and γ respectively. The major problem in classifying the products into the master taxonomy is the ability to identify candidate products for labeling. In this paper, we propose a Semi supervised learning methodology to overcome this problem by incrementally retraining the base classifier with parameters chosen during the taxonomy-aware calibration. Semi-supervised learning is a learning standard which deals with the study of how computers and natural systems such as human beings acquire knowledge in the presence of both labeled and unlabeled data. The proposed system finds each candidate parameter θi and then finds the optimal parameter γ such that the accuracy on the validation set is at the maximum. An experimental result shows that the Semi supervised learning algorithm is efficient and thus applicable to the large data sets on the web.

Key-Words: - semi-supervised learning, catalog integration, classification, web taxonomy, data mining

1 Introduction

An increasing number of web portals provide a user experience centered on online shopping. These web portals include several commercial sites such as Amazon and PriceGrabber and commerce search engines such as Bing Shopping. Hence data integration task is important for these commercial portals. The data integration task faced by these marketable portals is the integration of data coming from numerous data providers into a particular product catalog. This process is known as product categorization.

All web portals maintain their own master taxonomy for organizing products and it is used for both online shopping and searching purposes. When a new product arrives from dissimilar providers, it should automatically categorize the products into the master taxonomy according to their structure. But in web environment it is highly unlikely for the data providers to manually assign the products from the provider taxonomy to their corresponding categories in the master taxonomy. Automatic labeling techniques are needed for categorization products coming from data providers. Product catalog integration is the process of offering products from different vendor catalog for sale on a website. Another scenario that is conceptually similar concerns a company providing access to an office supply catalog on their internal website, allowing employees to order their own supplies for their office. This process needs to be considered from the perspective of the recipient of a vendor catalog as well as from the perspective of the vendor providing the catalog. Poor product categorization can frustrate shoppers and search engines. The product categorization should help shoppers find what they are looking for, but if the products are in the wrong place, they may be unseen by the shoppers which is bad for the economy of the shop.

Machine learning is a wide subfield of artificial intelligence. It involves creating algorithms and methods that allow computers to learn. This ability to learn from experience, analytical observation, and other means, results in a system that can endlessly improve itself and thereby offer

increased efficiency. During the time of training the learner has no knowledge about the test dataset. However, in transductive learning the learner is aware of the test dataset at the time of training and therefore only needs to shape a good classifier that generalizes to this known test dataset. Semi-supervised learning is the process of finding a better classifier from both labeled and unlabeled data. Semi-supervised learning methodology can deliver high performance of classification by utilizing unlabeled data. The methodology can be used to adapt to a variety of situations by identifying as opposed to specifying a relationship between labeled and unlabeled data from data. It can yield an improvement when unlabeled data can reconstruct the optimal classification boundary. Some popular semi-supervised learning models [11] include self-training [12], [14], mixture models [13], [16], co-training [15] and graph-based methods [17]. The success of semi-supervised learning depends completely on some underlying assumptions. So the emphasis is on the assumptions made by each model. The key stages of the work are as follows:

1.Formulate the taxonomy-aware catalog integration problem as a structured prediction problem by emphasizing the structure of the taxonomies in order to enhance the catalog integration.

2. Define taxonomy-aware classification as a two-step process with the first step being the base classification step where we classify the products based on their textual representations and in the second step called the taxonomy aware processing step we use the probability output by the base classifier to adjust according to the categories in the master taxonomy.

3. The label classification problem is overcome by using a scalable algorithm in the taxonomy aware processing step for the classification process.

4. Calibration of the parameters k, θ and γ is important for the performance of the system. Semi supervised learning algorithm makes better use of the classification results. We apply the Semi supervised learning algorithm for selecting the optimal result output by the base classifier b on the products of the validation set.

5. Finally we evaluate the experimental results on real-world data and show that the proposed semi supervised learning algorithm for the parameter calibration step provides better accuracy than the simple taxonomy aware classification.

The overview of the various steps in the taxonomy aware catalog integration with semi supervised learning process is shown in figure 1.

2 Related Work

In this section we study the numerous methods followed to solve the catalog integration problem as well as the structured prediction problem.

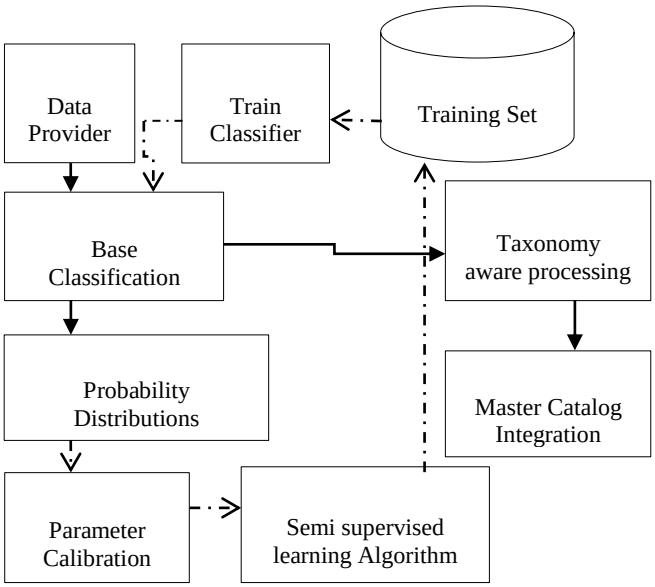

Fig.1 Overview of the taxonomy aware catalog integration with semi supervised learning

Papadimitriou and Tsaparas et. al [1] consider the problem of classifying the products from the provider taxonomy into the master taxonomy by making use of the taxonomy information of the provider. The approach is based solely on a taxonomy-aware processing step that adjust the results of the base classifier to make sure that the products that are tied together in the provider catalog remain close in the master catalog as shown in figure 2.

R. Agrawal and R. Srikant [2] consider the integration of documents from different sources into a master catalog is prevalent in web marketplaces and web portals. The current technology for automating this process consists of building a classifier that uses the categorization of documents in the master catalog to construct a model for predicting the category of unknown documents. But many of the data sources have their own categorization, and the accuracy of classification can be improved by factoring in the implicit information in these source categorizations. Classification is enhanced to incorporate the similarity information present in source catalogs. The experimental evaluation show substantial improvement in the accuracy of catalog integration.

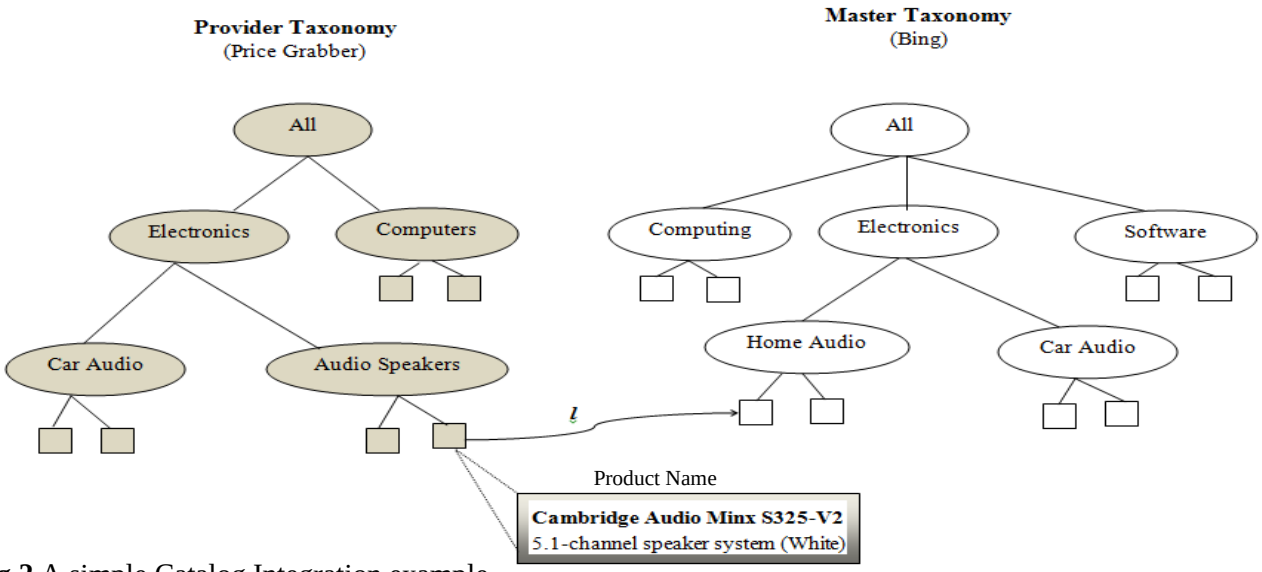

Fig.2 A simple Catalog Integration example

Y. Boykov and V. Kolmogorov [3] propose an energy minimization scheme provides two cost function models for creation of a labeling such that there is no swap move that decreases the energy. The two cost models are named as the Assignment cost and the Separation cost. The assignment cost is some positive cost associated with changing the intensity from its original value to a new value, the larger the change, the larger the cost.

The separation cost requires the smoothness term to be a metric. The separation cost is the cost of relabeling but that can be offset by the edge cost saved by being closer to its neighbors. The assignment cost will weigh the labeling in favor of the original values since most of the intensities are likely to be correct.

A. Fraser et. al and P. Ravikumar et. al provide the formulation of the catalog integration problem as an optimization problem is stimulated by the metric labeling problem. The metric labeling problem aims to discover the optimal labeling of a number of objects consequently that they reduce an assignment and a separation cost. The problem is NP-hard and the different obtainable estimated solutions formulate it as a Linear Programming problem (LP) [4] or a Quadratic Programming (QP) [5]. The purpose of our optimization problem is also comparable to the objective that arises in computer vision problems.

C. Chekuri et. al [6] consider the process of finding a label at minimum cost where the cost of a labeling is determined by the pairwise relations between the objects is considered. A distance function on labels; the distance function is assumed to be a metric is used. Each object also incurs an assignment cost that is label, and vertex dependent. The problem captures many classification problems that arise in computer vision and related fields. The solution to the problem is obtained from a general formulation. This formulation allows us to extend the ideas to obtain the first non-trivial approximation for the truncated quadratic distance function.

G. Ifrim et. al [7] show a Bayesian logistic regression approach that uses a Laplace prior to avoid over fitting and produces sparse predictive models for text data. This approach is applied to a range of document classification problems and show that it produces compact predictive models at least as effective as those produced by other classifiers. Lasso logistic regression provides state-of-the-art text categorization effectiveness while producing sparse and thus efficient models. The approach is also useful in other high dimensional data analysis problems

D. Zhang and W.S. Lee [8] discuss a straightforward approach to automating the process of catalog integration would be to learn a classifier that can classify objects from the source taxonomy into categories of the master taxonomy. The key vision is that the availability of the source taxonomy data could be helpful to build better classifiers for the master taxonomy if their categorizations have some semantic overlap.

Co-bootstrapping is used to enhance the classification by exploiting such implicit knowledge. It performs real-world web data show substantial improvements in the performance of taxonomy integration. Integrating objects from source taxonomy into a master taxonomy. This problem is not only currently prevalent on the web, but also very important to the upcoming semantic web. A direct approach to automating this process would be to train a classifier for all the categories

in the taxonomy of the master catalog, and then classify the objects from the source taxonomy into these categories. S. Sarawagi, et. al [9] provide a cross training model is established for document classification occurrence of multiple label sets. Document classification is a well-established region of text mining. A document classifier is original trained using documents with pre-assigned labels or classes picked from a set of labels it is named as taxonomy or catalog. Once the classifier is trained, it is offered test documents for which it must guess the best labels. General semi-supervised learning framework called cross-training which can exploit knowledge about label assignments in one taxonomy B to make better inferences about label assignments in taxonomy A. Cross-training generalizes several existing classification algorithms, while also comparing favorably with their accuracy on a host of related applications. Apart from increased classification accuracy, the benefits include a better understanding of probabilistic relationships between taxonomies, and more experience with encoding heterogeneous features for learning algorithms. It doesn't make different taxonomy models for each and every product mentioned in web search engine.

Ming Ji et. al [10] describe a Simple algorithm for semi-supervised learning that on one hand is easy to implement, and on the other hand is guaranteed to improve the generalization performance of supervised learning under appropriate assumptions. It is learned from the labeled examples the best prediction function that can be used for the parameter calibration and it can also be used to incrementally re-train the base classifier.

In the era of data divulgence, there has been broad interest in leveraging a massive amount of data available in open sources such as the Web to help solve long standing problems like object recognition, topic detection, and multimedia information retrieval. One promising direction gaining a lot of attention aims to develop the best ways of combining labeled data (often of limited amount) and a huge pool of unlabeled data in forming abundant training resources for optimizing machine learning models. This learning paradigm is referred to as semi supervised learning (SSL).

The main idea of the proposed algorithm is to estimate the top Eigen functions of the integral operator from the both labeled and unlabeled examples, and learn from the labeled examples the best prediction function in the subspace spanned by the estimated Eigen function. Unlike the previous studies of exploring Eigen functions for semi-supervised learning. To derive the generalization error bound, a different set of assumptions are made from previous studies.

3 Problem Definition

The taxonomy aware catalog integration problem is defined using some basic terminology. Let X be a product that can be bought at a commercial portal. Every product has a textual description that contains a name of the product and perhaps a set of attribute-value pairs. By considering this the taxonomy of the product can be represented as a Directed Acyclic Graph (DAG) $G = \{C_g, E_g\}$) whose nodes C_g represent the set of probable categories into which products are prearranged. Each edge of the graph $(C_1, C_2) \in Eg$ represents a subsumption association between two categories C_1 and C_2. Now we define the taxonomy aware catalog integration problem as, for a given source catalog K_s and a target catalog K_t, the problem is to learn a cross-catalog labeling function $l = f_T (K_s, K_t)$ by using a taxonomy-aware process $f_{T..}$

4 Taxonomy Aware Classification

The taxonomy aware classification is a two-step process. The first step is the base classification step where we classify the products based on their textual representations and the second step called the taxonomy aware processing step, we use the probability output by the base classifier to adjust according to the categories in the master taxonomy.

4.1 Base Classification Step

Base classification step classifies the products based solely on their textual representation. For this purpose, a text-based classifier is trained using standard supervised machine learning techniques. Naive Bayes (NB) and Logistic Regression (LR) are used. Then use a subset of the target catalog as the training set. This provides with examples of products labeled with categories of the target taxonomy. The attributes of the classifier are extracted from the textual product representations. Note that at training time knowledge of the providers' catalog is not known, and no use of the structure of the target taxonomy. During the base classification step any knowledge about the target or source taxonomy is not considered, either during training, or during the application of the classifier.
This refers to the structure of the taxonomy, as well as the category names. The classifier can possibly

try to match the names of the categories between the source and target taxonomies. However, this entails the danger of over fitting, and also as observed, category names often vary significantly between providers (e.g., Cameras versus Photography).

4.2 Taxonomy-Aware Processing Step

Taxonomy-aware processing step is that the target categories assigned by the base classification step can be adjusted by taking into account the relationships of the products in the source and target taxonomies. The objective of the taxonomy-aware processing is to allocate categories in the target taxonomy to the products coming from the catalog of the provider, such that the allocations respect the decisions of the base classifier and at the same time preserving the relative relationships of the products in the source taxonomy. The taxonomy-aware processing problem is defined [1] as an optimization problem with a given source catalog Ks and a target catalog Kt, the problem is to learn a cross-catalog labeling function l that minimizes the following cost function:

$$\mathrm{COST}(\kappa_S,\kappa_t,l)=(1-\Upsilon)\sum_{x\epsilon P_s}\mathrm{A\ Cost}(x,l_x)+\Upsilon\sum_{x,y\epsilon P_s}\mathrm{S\ Cost}(x,y,l_x,l_y) \quad (1)$$

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

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

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

$$\mathrm{A\ Cost}(x,\ell_x)=1-\Pr_b(\ell_x|\ell_y) \quad (3)$$

Important similarity description is supposed to assure the perception the two categories that are close together in the taxonomy tree are more comparable than two categories that are far away. For example, the two categories that have a common parent are more similar than two other categories that have dissimilar parents and a normal grandparent. The division cost called as the separation cost is defined as a function of the similarity sim_S (s_x, s_y) between categories and of x and y in the source taxonomy S and similarity sim_T (s_x,s_y) between categories and of x and y in the target taxonomy T is given by:

$$\mathrm{S\ Cost}(x,y,\ell_x,\ell_y)=\delta\left(\mathrm{sim}_S(s_x,s_y),\mathrm{sim}_T(s_x,s_y)\right) \quad (4)$$

The optimization problem occurs in the above steps and hence to overcome the problem of optimization we use the search space pruning for the parameter calibration. The aim is to fix the categories for some of the products wisely and obtain the landscape of mappings between the categories in either of the taxonomies. Then we can use this to find a related mapping to the products in the open category (non-fixed products) and find the separation cost using (4). Let $\theta\epsilon$ [0, 1] be a threshold value defined while the category probability distribution returned by the base classifier is great enough that the predicted category is expected to be accurate. Let $F\theta$ be the subset of products that pass the threshold is defined as,

$$F_\theta=\{x\in P_s|\max_{\gamma\epsilon C_t}\Pr_b[\tau|x\geq\theta\} \quad (5)$$

The products in F_θ are fixed with the output probability of the base classifier being large enough. That is, for all $x\epsilon F_\theta$,

$$\ell_x=\arg\max_{\gamma\epsilon C_t}\Pr_b[\tau|x] \quad (6)$$

Let $O_\theta=P_s/F_\theta$ denote the products with classification is still not fixed. Each open product $x\in O_\theta$ independently and calculate a division cost for only with respect to the products in the fixed category F_θ. If s_x is the source category of product x and t_x is a candidate target category, then the cost of separation for this source-target pair is defined as follows:

$$h(s_x,t_x)=\sum_{\sigma\epsilon S,\tau\epsilon T}\mathrm{S\ COST}(\mathrm{S\ Cost}(s_x,\sigma,t_x,\tau)\bar{n}(s_x,t_x)\overline{n(}\sigma,\tau) \quad (7)$$

Where, $\overline{n(}\sigma,\tau)$ is the number of products in the fixed category F_θ that belong to the category σ in S and are allocated to the category τ in T. We use $H_{\theta,k}$ to denote the set of the candidate source-target pairs for which the separation cost h has to be computed :

$$H_{\theta,k}=\{(s_x,\tau):x\in O_\theta,\tau\in TOP_k(x)\} \quad (8)$$

Algorithm 1 describes the modified Taxonomy Aware Catalog Integration (TACI) algorithm. The algorithm assumes the presence of a base classifier trained on data from the target catalog. The input to the algorithm consists of a source catalog S and target taxonomy T and also the parameters k, θ and γ. The output of the algorithm is a label l for the products in the source catalog.

Algorithm 1: modified TACI algorithm
Input: Source catalog κ_S, Target Taxonomy T, base classifier b and parameters θ, k and γ.

Output: Labeling vector *l*.

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

In the loop of Lines 2-9 is the base classification step where the algorithm applies the base classifier to each product in the provider catalog. Based on the base classifier output probability distributions, the algorithm either classifies the product to the top category (as in lines 4-6) given by the base classifier that is the fixed category F_θ, or it leaves its classification open (as in lines 7-9) and stores the top k categories, sorted by probability. Based on the set of open products O_θ, and their top-k candidate target categories the algorithm calculates (as in line 10) the set of candidate source-category pairs $H_{\theta,k}$. The algorithm then computes the separation costs for all of the candidate pairs $(\sigma, \tau) \in H_{\theta,k}$, and stores them in a hash table Ψ as in lines 12-13. It is to be noted that for each source-target pair value of h is computed only once, and the separation cost is never computed. In the loop of the lines 14-15, the algorithm classifies the open products in O_θ, the open category. A product $x \in O_\theta$ is assigned to the category l_x among the top-k categories in $TOP_k(x)$ that minimizes the objective function.

4.3 Parameter Calibration

The tuning of the parameters k, θ and γ is important for the performance of the algorithm. The chosen validation set consists of products that are cross labeled in both the source and the target taxonomy. The Base classifier is trained with a number of features and it is big enough to tune few parameters of the TACI algorithm. The first parameter set is the parameter k, such that the accuracy of the classifier over the top-k categories is high. Then, we tune the parameters θ which determines the anchor set F_θ by choosing N equally spaced probability values. For each candidate parameter we find the optimal parameter γ such that the accuracy of the TACI algorithm on the validation set is maximized. We notify all the parameters that are selected such as to maximize the accuracy of the TACI algorithm on the validation set. A detailed explanation on tuning the parameters k, θ and γ can be found in [1].

4.4 Semi Supervised Learning for Parameter Calibration

In general the learning methods can be divided into supervised and unsupervised learning. In the supervised learning methods learner aims at estimation of the input –output relationship by using objective function with training set data set $\{x_i, y_i\}$, i = 1, . . . , N where the inputs x are n-dimensional vectors and the labels y are continuous values for regression tasks and discrete for classification problems; In unsupervised learning only the raw data x_i are available, not including the consequent labels y_i. This type of the algorithm belonging to the group are clustering and independent component analysis routines .It becomes difficult to handle the unlabeled data, to handle this situation where some labeled patterns are provided jointly with unlabeled ones arise frequently. This type of learning is named as the semi supervised learning. The proposed algorithm for semi-supervised learning during calibration step that on one hand is easy to execute and on the other hand is guaranteed to improve the categorization of the product result performance. The main idea of the proposed algorithm is to estimate the top eigen functions of the integral operator from the both labeled and unlabeled examples, and learn from the labeled examples the best prediction function in the subspace spanned by the estimated eigen functions.

Let X be a compact domain or a manifold in the Euclidean space R_d. Let $D = \{x_i, i = 1, \ldots, N \in x_i \in X\}$ be a collection of training examples. Randomly select the n examples from D for labeling. Without loss of generality, we assume that the first n examples are labeled by $y_l = (y_1, \ldots, y_n)^T \in R_n$. We denote by $y = (y_1, \ldots, y_N)^T \in RN$ the true labels values for parameters such as for all the examples in D. In this study, we assume y = f(x) is decided by an unknown deterministic function f(x). Our goal is to learn an accurate prediction parameter θ to incrementally retrain the base classifier at calibration step.

Algorithm 2: Semi supervised learning for calibration step

Input: D = $\{x_i, i = 1, \ldots, N \in x_i \in X\}$ be a collection of training examples , $y_l = (y_1, \ldots, y_n)^T$ labels for the n examples selected randomly ,s be the eigen vectors selected
Output: Parameter θ.

1: Compute $(\widehat{\emptyset_I}, \widehat{\lambda_1})$,i=1,……,s the first eigen functions and eigen values for the integral operator is defined as

$$\widehat{L_N}(f)(.) = \frac{1}{N}\sum_{i=1}^{N} k(x_I, .).f(x_i)$$

2: Compute the prediction result $\hat{g}(.)$ which is to be considered as the prediction parameter θ to incrementally retrain the base classifier at calibration step

$$\hat{g}(X) = \sum_{j=1}^{s} \gamma_j^* \widehat{\emptyset_j}(X)$$

Where,$\gamma^* = \{\gamma_1^*, \ldots \ldots, \gamma_s^*\}$ is given by solving the following equation,

$$\gamma^* = \arg\min_{\gamma \epsilon R^s} \sum_{i=1}^{n}\sum_{j=1}^{s} (\gamma_j \widehat{\emptyset_j}(X_i) - y_i)^2$$

5 Experimental Evaluation

In this section we measure the accuracy of classification by the TACI algorithm and TACI with semi supervised learning. The results show that the semi supervised TACI algorithm outperforms the unsupervised TACI algorithm. We measure the accuracy for TACI with Naïve Bayes (TACI-NB), TACI with Logistic Regression (TACI-LR) and TACI with semi supervised learning (TACI-Semi Supervised) methods by considering the PriceGrabber dataset with 21 channels and 193 categories is chosen as the data provider. BingShopping dataset with 30 channels and 376 categories is used as the master catalog, which combined data feeds from vendors, suppliers, resellers, and other profitable portals. We consider a target taxonomy that consists of all the categories in Bing Shopping taxonomy that is related to consumer electronics and computing. It is seen that in all the experiments, Taxonomy-Aware Catalog Integration with Semi supervised learning (TACI-Semi Supervised) shows a better accuracy than the Taxonomy-Aware Catalog Integration with Naive bayes (TACI-NB), Taxonomy-Aware Catalog Integration with Linear regression (TACI-LR).

5.1 Classification Accuracy Evaluation

The evaluation metric chosen here is the ratio of the number of source products for which the correct target category is predicted to the total number of source products. Table 1 shows the accuracy of all the algorithms and the following figure 3 shows the comparison of accuracy with all the methods.

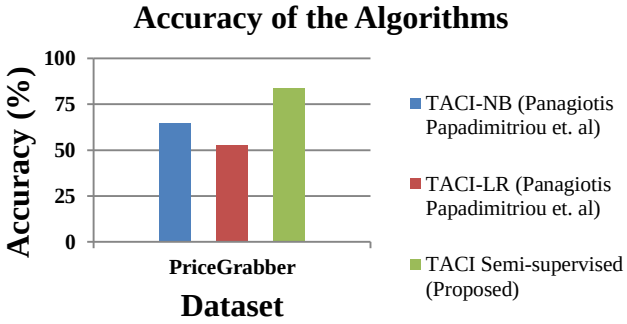

Fig.3 Comparison of Accuracy of the TACI-NB, TACI-LR and TACI-Semi supervised

Dataset	Algorithm	Accuracy
PriceGrabber	TACI NB	64.50
	TACI LR	52.35
	TACI Semi-supervised	83.52

Table 1 Classification Accuracy Evaluation

5 Conclusion and Future Work

The presented efficient and scalable approach to catalog integration is based on the use of source category and taxonomy structure information. In this research, a well-organized approach to catalog integration that is based on the use of source category and taxonomy structure information is presented. The proposed semi supervised learning algorithm is used for retraining the base classifier during the parameter calibration step; they can also be used for other problems. Several algorithms were used for classification; they can also be used as a feature for item matching, when the elements classified under the master taxonomy (e.g., the products in the master catalog) has to be matched to incoming offers from the providers. This move

toward can lead to considerable gains in correctness with respect than the existing calibration step based classifier. It also showed that this approach leads to substantial gains in accuracy with respect to existing classifiers.

In future, the base classification step using machine learning algorithms and the proposed technique can be used in an active learning environment in order to identify candidate products for labeling. Finally parameter selection can be performed with optimization methods that can select and can retrain the base classifier with attributes chosen during the taxonomy-aware calibration step.

References:

[1] Panagiotis Papadimitriou, Panayiotis Tsaparas, Ariel Fuxman, Lise Getoor, "TACI: Taxonomy-Aware Catalog Integration",*IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, July 2013.

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

[3] Y. Boykov and V. Kolmogorov, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124-1137, Sept. 2004.

[4] A. Fraser and D. Marcu, "Getting the Structure Right for Word Alignment: Leaf," *Proc. Joint Conf. Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL), 2007.

[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] 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.

[7] Georgiana Ifrim, Gökhan Bakir, Gerhard Weikum, "Fast logistic regression for text categorization with variable-length n-grams" KDD '08 *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* Pages 354-362.

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

[10] Ming Ji, Tianbao Yang, Binbin Lin, Jiawei Han, "A Simple Algorithm for Semi-supervised Learning with Improved Generalization Error Bound" *Proc. 29th Int'l Conf. on Machine Learning*, Edinburgh, Scotland, UK, 2012.

[11] Xiaojin Zhu, "Semi-Supervised Learning Literature Survey", Computer Sciences TR 1530 University of Wisconsin, Madison. Last modified on July 19, 2008.

[12] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-Supervised Self-Training of Object Detection Models," *Proc. Seventh Workshop Applications of Computer Vision*, vol. 1, pp. 29-36, Jan.2005.

[13] Fujino, A., Ueda, N., & Saito, K. "A hybrid generative/discriminative approach to semi-supervised classifier design". AAAI-05, *the Twentieth National Conference on Artificial Intelligence*.2005.

[14] Riloff, E., Wiebe, J., &Wilson, T. "Learning subjective nouns using extraction pattern bootstrapping." *Proceedings of the Seventh Conference on Natural Language Learning* CoNLL-2003.

[15] Blum, A., Mitchell, T. "Combining labeled and unlabeled data with co-training" COLT: *Proceedings of the Workshop on Computational Learning Theory*, Morgan Kaufmann, 1998, p. 92-100.

[16] Xiaojin Zhu ,John Lafferty "Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning",*Proc. of the 22nd Int'l Conference on Machine Learning*, Bonn, Germany, 2005.

[17] Zhou, D., Huang, J., & Sch¨olkopf, B. "Learning from labeled and unlabeled data on a directed graph." ICML05, *22nd International Conference on Machine Learning*. Bonn, Germany.