An Agent Based Catalog Integration System through Active Learning

G.Sindhu Priya [#1], P.Krubhala [*2], P.Niranjana [#3]
[#]*Assistant Professor & Department of Computer Science and Engineering & Anna University*
Nadar Saraswathi College of Engineering & Technology,Theni,India.

Abstract *Online Commercial data integration plays a vital role in categorizing the products from multiple providers all over the globe. An unique taxonomy is maintained by the Commercial portals and products of the providers are associated with their own taxonomy. In the existing work, an efficient and scalable approach to Catalog Integration is used which is based on the use of Source Category and Taxonomy structure Information. We formulate this intuition as a structured prediction optimization problem. Learning algorithms can actively query the user for labels. Active learning concept is used to identify candidate products for labeling and also used to obtain the desired outputs at new data points. It intends to develop the catalog integration process in automated fashion in an agent based environment in which agent can cooperate interact with the consumers to find the best classification based upon the consumer preferences.*

Keywords— *Active learning, Catalog Integration, classification, Master taxonomy, Provider taxonomy, Agent.*

I. INTRODUCTION

Integration of data is the important task for online ecommerce based web portals and commerce search engine based application. Catalog integration is the process of offering products from different vendor catalogs for sale on a Web site. Ecommerce based web portals include Amazon and Shopping.com. Search Engines such as Google and Bing Shopping. Two taxonomies are maintained as Master taxonomy and Provider taxonomy. Master taxonomy is used for organizing their products which is used for Browsing and Searching Purposes. The data providers do have their own taxonomy called Provider taxonomy. All web portals maintain their own master taxonomy for organizing products arrive from the different providers, it automatically categorize the products in master taxonomy according to their users. But in website environment it is difficult to assign the products from their catalog to the appropriate category in the master taxonomy .So automatic labeling techniques is used for categorizing the products.

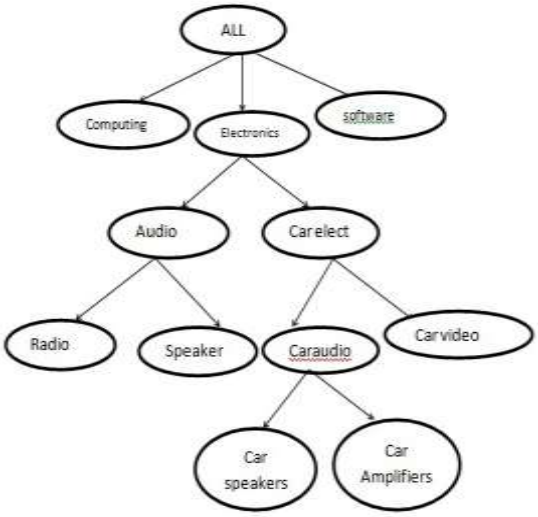

Fig .1.Master Taxonomy

For automatic categorization, the products are previously associated with the provider taxonomy, it will differ from the master taxonomy. For example , in Fig. 1 and 2 the product "Audio Systems CH2029" from the category Electronics/Car Electronics/Car Audio & Video/Car Speakers/Coaxial Speakers in the Provider taxonomy is mapped to Electronics/Car Electronics/Car Audio/Car Speakers or Electronics/Home Audio/Speakers. Text based classifier is used to adjust the results of taxonomy information. But we cannot get clear classification information and also we don't know where the product should be categorized at the leaf level. So Learning algorithms are used to actively query the user for automatic labeling. The major Contribution of the work as follows:

1. The taxonomy aware catalog integration problem is predicted as a Structured Prediction Problem. In this method, the approach leverages the taxonomy to enhance catalog integration.
2. Taxonomy aware Classification process with two steps: Base Classification Step and Taxonomy aware classification step.Products are classified using base classification step. In taxonomy aware classification step, the optimization problem can be overcomed by using TACI algorithm.

3. Active Learning concept is used to interactively query the user to enhance catalog prediction through structured analysis in an agent based environment.
4. Finally evaluate the experimental results and compare taxonomy aware classification and semi supervised active learning concept, it provides significant accuracy over existing algorithm.

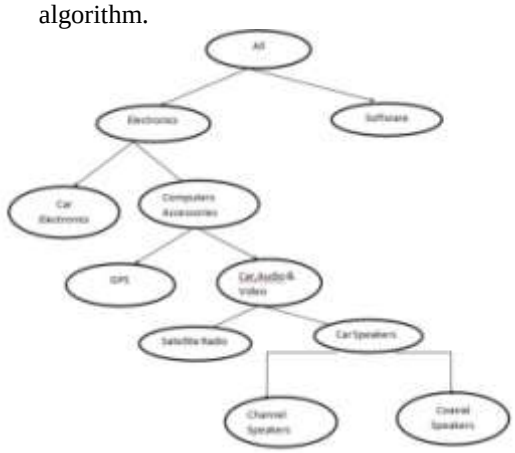

Fig. 2.Provider Taxonomy

II. RELATED WORK

In this section various methods are going to study to solve catalog integration problem such as metric labeling and structured prediction.

R. Agrawal and R. Srikant [1] addressed the problem of integrating documents from different providers into a master catalog. This is a pervasive problem in web portals and market places. It processes the product catalog to construct the base classifier for product integration of documents in the master catalog for predicting the category of unidentified documents.

Sarawagi et al [6] establish cross training model with semisupervised learning for document classification. A general semi-supervised learning framework called cross-training, a new technique for using sample documents from one taxonomy to improve classification tasks for another taxonomy.

Daume [9] proposed a method for integrating searching and learning that transforms complex problems into simple classification problems to which any binary classifier may be applied. SEARN, an algorithm for solving complex structured prediction problems with minimal assumptions on the structure of the output and loss function.

Zhang and Lee [3], [4], [5] have also proposed an approach to catalog integration by using boosting and transductive learning method. These approaches attain better categorization accuracy similar to the cross-training approach, but this approach needs training data that are labeled in both the provider and the master taxonomies. This method is not appropriate to our problem setting that is integrate the product from provider to master taxonomy.

III. TAXONOMY-AWARE CATALOG INTEGRATION AND SEMI SUPERVISED ACTIVE LEARNING CONCEPT IN AN AGENT BASED ENVIRONMENT

In this section we formulate the taxonomy aware catalog integration problem. Each product is identified by its name and attribute value.For example "Audio systems 2029".Here the product name is Audio systems and the attribute value description is chaos series 2.0 inch 2-way speaker,900W peak power." Taxonomy aware categorization as a two step process. First the products are classified based on their textual representation and then in second step the output is adjusted based on the structure of the master and provider taxonomy.

A.THE BASE CLASSIFICATION STEP

In this step, the products are classified based on their textual representation. Each product is classified by using a base classifier. the base classifier does not have any aware about the taxonomies. Machine learning techniques such as Naïve Bayes and Logistic Regression are used. The features of the product are extracted from the textual representation of the product.

B.THE TAXONOMY AWARE PROCESSING STEP

In the taxonomy aware processing step, the result of the base classification step can be adjusted by using the structure of the master and provider taxonomies. Here optimization problem occurs where the provider catalog K_p and the master catalog K_m ,the objective is to calculate labeling vector l that minimizes the cost function. The cost fuction formula is as follows.

$COST(K_p,K_m,l)=$
$(1-\gamma)\sum_{x\varepsilon PS}ACost(x,l_x)+\gamma\sum_{x,y\varepsilon PS} Cost(x,y,l_x,l_y)$
The taxonomy aware procedure f_T,
$f_T(K_p,K_m)= COST(K_p,K_m,l)$
Algorithm: TACI
Input: Source catalog K_p ,Target Taxonomy K_m, base classifier b and parameters θ ,k,γ
Output: Labeling vector l
1. $F_s \rightarrow \varphi$
2. For all $x \in P_s$ do
3. $\tau^* \leftarrow \arg\max_{\tau\varepsilon Ct,}\max_{\gamma\varepsilon Ct} P\gamma_b[\tau/x]$
4. if $P\gamma_b[\tau^*/x] \geq$ then
5. $lx \leftarrow \tau^*$
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 HT to empty

12. For all $(\sigma,\tau)\varepsilon\ H_{\theta,k}$ do
13. $HT(\sigma,\tau)=H(\sigma,\tau))$
14. For all $x \in Odo$
15. $l_x \leftarrow argmin_{\tau\varepsilon TOPk(x)}\{(1-\gamma)$
$ACOST_{x,\tau+\gamma HT(Sx,\tau)}\}$

The problem in using the Taxonomy aware processing step is that it involves large number of pair-wise relationships among the products categorized from the provider taxonomy to the master taxonomy. Computation over large data sets may occur. Assignment and separation costs for products are also increased. These problems are overcomed by using semi supervised active learning concept in an agent based environment.

C.SEMISUPERVISED ACTIVE LEARNING WITH WEKA TOOL IN AN AGENT BASED ENVIRONMENT

Active learning is a special case of semi-supervised machine learning in which a learning algorithm is able to interactively query the user to obtain the desired outputs at new data points.There are situations in which unlabeled data is abundant but manually labeling is expensive. In such a scenario, learning algorithms can actively query the user for labels. This type of iterative Supervised learning is called active learning.

Semi supervised learning is a learning concept between supervised learning and unsupervised learning. Supervised learning contains labeled data whereas unsupervised learning contains unlabeled data. Semi supervised learning uses large number of labeled data with small number of unlabeled data.

The catalog integration process can be developed in automated fashion in an agent based environment through active learning. Agent can co-operate, interact with the consumers to find the best classification based upon the consumer preferences. Agent can be used to query the user for classification. Separate log is maintained for the products classified based on the active learning concept. The products from the provider taxonomy are categorized into the master taxonomy by using the log maintained. Agent can help the user to maintain the log for classification from the provider taxonomy to the master taxonomy.

Weka tool is used for classifying the products from the provider taxonomy to the master taxonomy. Weka is a collection of machine learning algorithms for data mining tasks. Here weka tool is used for classification process with machine learning techniques like active learning concept in an agent based environment. Computation problem may avoid by using weka tool for large data sets. The classifier accuracy can be improved by using weka with active learning machine concepts. The categorization of the product can be improved by using semi supervised learning algorithm.

Algorithm : Semi supervised learning for calibration step

Input: D = $\{\mathbf{a_i}, i = 1, \ldots, N \in \mathbf{a}i \in A\}$ be a collection of training examples, $\mathbf{y}l = (y1, \ldots, yn)\top$ are the labels randomly selected.S be the eigen vector.

1. Compute (ϕ_l,λ_l) i=1,….S the eigen functions and the eigen value for integral function is defined as
$L_N(f)(.)=1/N\sum^{N}_{i=1} k(a_{i,..}).f(a_i)$.
2. Compute the Prediction result g^(.),to retrain the base classifier prediction parameter Θ at the calibration step $g(X)=\sum_{j=1}^{s}\gamma_j*\phi_j(X)$ where $\gamma^*=\{\gamma_1*,\ldots,\gamma_s*)$.

IV.EXPERIMENTAL RESULTS

The classification accuracy for taxonomy aware classification step and the semi supervised active learning concept with weka can be measured. The accuracy can be measured by using three different providers such as Amazon, Pricegrabber, Etilize .Here we compare our semi-supervised active learning using weka with taxonomy aware catalog integration with Naïve Bayes and taxonomy aware catalog integration with Logistic Regression. We compare the accuracy of three algorithms in Table.1 and Fig.3

Providers	TACI NB	TACI LR	TACI SSAL with weka
Amazon	80.1	74.3	85.7
Pricegrabber	72.2	75.5	82.2
Etilize	81.2	85.3	89.5

3.

Table 1. Classification Accuracy Evaluation

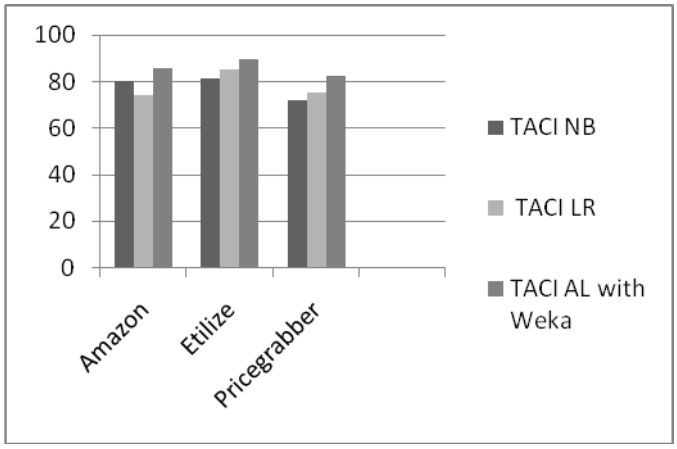

Fig. 3. Classification Accuracy Evaluation

V. CONCLUSION

In this research, catalog integration approach is used that is based on the use of source category and taxonomy structure information. The proposed semi supervised active learning algorithm with weka tool were used for retrain the base classifier and also used to increase the classification accuracy. The output of

the parameter result is used as a feature for item identical which matches the products in the master catalog to the products coming from the provider taxonomy. Experimental results are also shown which compares the new technique with the existing base classifier.

REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] H. Daume´ III, J. Langford, and D. Marcu, "Search-Based Structured Prediction," Machine Learning J., vol. 75, pp. 297-325, 2009.

[10] E. Rahm and P. Bernstein, "A Survey of Approaches to Automatic Schema Matching," The VLDB J., vol. 10, no. 4, pp. 334-350, 2001.