Recursive Product Catalog Pattern Matching and Learning for Categorization of Products in Commercial Portal

K.N.R.Manchusha
PGScholar: Department of CSE
Paavai Engineering College
Pachal,Namakkal
knrmanchusha@gmail.com

P.Renukadevi
AssistantProfessor: Department of CSE
Paavai Engineering College
Pachal,Namakkal
renuka@gmail.com

***Abstract*— In commercial portals and search engines complexity arise in data integration task for products coming from multiple providers to their product -of-word model leads to a generative classification model categorization of commercial portal based on the target product coming catalogs. Recursive Product Catalog Pattern Matching and Learning Scheme for product from other providers. Taxonomy is represented with a bagbased on mixture of multinomial. Recursive product catalog learning improves the identification of specific category. The desired result is achieved by using performance measure such as product catalog integration accuracy, execution time, number of products, separation cost, number of EM iteration**.

Keywords—Catalog integration; classification; data mining; taxonamy

I. Introduction

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data integration is the major important task for online commercial portals and commerce search engine based applications. The data integration task faced by online commercial portals and e-commerce search engines are the integration of products coming from multiple providers to their product catalogs.

This approach is based on a taxonomy-aware processing step that adjusts the results of a text-based classifier to ensure that products that are close together in the provider taxonomy remain close in the master taxonomy. An increasing number of web portals provide a user experience centered around online shopping. This includes e-commerce sites such as amazons and shopping.-com, and commerce search engines such as Google product search and bing shopping. A fundamental data integration task faced by these commercial portals is the integration of data coming from multiple data providers into a single product catalog. An important step in this process is product categorization.

An important observation in this scenario is that the data providers do have their own taxonomy and their products are already associated with a provider taxonomy category. The provider taxonomy maybe different from the master taxonomy, but in most cases, there is still a powerful signal coming from the provider classification. Intuitively, products that are in nearby categories in the provider taxonomy, should be classified into nearby categories in the master taxonomy. The work contains the following contributions:

Formulates the taxonomy-aware catalog integration problem as a structured prediction problem. To the best of knowledge, this is the first approach that leverages the structure of the taxonomies in order to enhance catalog integration.

Present techniques that have linear running time with respect to the input data and are applicable to large-scale catalogs. This is in contrast to other structured classification algorithms in the literature which face challenges scaling to large data sets due to quadratic complexity.

Performs an extensive empirical evaluation of these algorithms on real-world data. It shows that taxonomy-aware classification provides a significant improvement in accuracy over existing state-of-the-art classifiers.

II. Related Work

V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M. Kan, and K.R. McKeown [1] in their paper said that a statistical similarity measuring and clustering tool, SIMFINDER that organizes small pieces of text from one or multiple documents into tight clusters. By placing highly related text units in the same cluster, SIMFINDER SW23E21enables a subsequent content selection/generation component to reduce each cluster to a single sentence, either by extraction or by reformulation. It reports on improvements in the similarity and clustering components of SIMFINDER, including a quantitative evaluation, and establish the generality of the approach by interfacing SIMFINDER to two very different summarization systems. Summarization is an application that cuts across multiple natural language processing areas (search, text analysis, planning, generation) and for which disparate approaches have been used, including word counts, information retrieval based similarity measures , statistical models, positional information, and discourse structure. For multidocument summarization, where the source texts often contain the same information with variations in the presentation, an alternative approach is to explicitly seek similar pieces of the input text, on the assumption that recurring text units are probably the more central ones. Each set of similar text pieces can then produce one sentence in the summary, either by extraction or by reformulation.

V. Kolmogorov and R. Zabih, [2] Minimizing an energy function via graph cuts, however, remains a technically difficult problem. Each paper constructs its own graph specifically for its individual energy function, and in some of these cases the construction is fairly complex. The goal of this paper is to precisely characterize the class of energy functions that can be minimized via graph cuts, and to give a general-purpose graph construction that minimizes any energy function in this class. The results play a key role provide a significant generalization of the energy minimization methods used and show how to minimize an interesting new class of energy functions. This paper considers only energy functions involving binary-valued variables. At first glance this restriction seems severe, since most work with graph cuts considers energy functions that involve variables with more than two possible values. For example, the algorithms presented for stereo, motion and image restoration use graph cuts to address the standard pixel labeling problem that arises in early vision. In a pixel labeling problem the variables represent individual pixels, and the possible values for an individual variable represent, e.g., its possible displacements or intensities.

For accurate location, the user with few labelled data where a major difficulty arises from the need to label large quantities of user location data, which in turn requires knowledge about the locations of signal transmitters, or access points. To solve this problem, Jeffrey Junfeng Pan, Sinno Jialin Pan zand Jie Yin, have developed a novel machine-learning-based approach [3] that combines collaborative filtering with graph-based semi supervised learning to learn both mobile-users' locations and the locations of access points. The framework exploits both labeled and unlabelled data from mobile devices and access points. In the two-phase solution, they first build a manifold-based model from a batch of labelled and unlabelled data in an offline training phase and then use a weighted k-nearest-neighbor method to localize a mobile client in an online localization phase.

Linear Discriminant Analysis (LDA) is a traditional algorithm for supervised feature extraction. Recently, unlabeled data have been utilized to improve LDA. However, the intrinsic problems of LDA still exist and only the similarity among the unlabeled data is utilized. They propose a novel algorithm, called Semisupervised Semi-Riemannian Metric Map [4] (S3RMM), following the geometric framework of semi- Riemannian manifolds. S3RMM maximizes the discrepancy of the separability and similarity measures of scatters formulated by using semi-Riemannian metric tensors. The metric tensor of each sample is learned via semisupervised regression. The method can also be a general framework for proposing new semisupervised algorithms, utilizing the existing discrepancy-criterion-based algorithms.Semantic world model framework is for hierarchical distributed representation of knowledge [5] in autonomous underwater systems. This framework aims to provide a more capable and holistic system, involving semantic interoperability among all involved information sources. This will enhance interoperability, independence of operation, and situation awareness of the embedded service-oriented agents for autonomous platforms. The results obtained specifically affect the mission flexibility, robustness, and autonomy. The presented framework makes use of the idea that heterogeneous real-world data of very different type must be processed by (and run through) several different layers, to be finally available in a suited format and at the right place to be accessible by high-level decision-making agents. In this sense, the presented approach shows how to abstract away from the raw real-world data step by step by means of semantic technologies.

Temporal data clustering provides underpinning techniques for discovering the intrinsic structure and condensing information over temporal data. They present a temporal data clustering framework via a weighted clustering ensemble of multiple partitions produced by initial clustering analysis on different temporal data representations. In the approach, they propose a novel weighted consensus function [6] guided by clustering validation criteria to reconcile initial partitions to candidate consensus partitions from different perspectives, and then, introduce an agreement function to further reconcile those candidate consensus partitions to a final partition. As a result, the proposed weighted clustering ensemble algorithm provides an effective enabling technique for the joint use of different representations, which cuts the information loss in a single representation and exploits various information sources underlying temporal data. In addition, the approach tends to capture the intrinsic structure of a data set, e.g., the number of clusters. The approach has been evaluated with benchmark time series, motion trajectory, and time-series data stream clustering tasks.

Based on an effective clustering algorithm—Affinity Propagation (AP)—they present a novel semisupervised text clustering algorithm, called Seeds Affinity Propagation (SAP) [7] . There are two main contributions in the approach: 1) a new similarity metric that captures the structural information of texts, and 2) a novel seed construction method to improve the semisupervised clustering process. To study the performance of the new algorithm, they applied it to the benchmark data set Reuters-21578 and compared it to two state-of-the-art clustering algorithms, namely, k-means algorithm and the original AP algorithm.

Supervised hyperspectral image classification is a difficult task due to the unbalance between the high dimensionality of the data and the limited availability of labeled training samples in real analysis scenarios. While the collection of labeled samples is generally difficult, expensive, and time-consuming, unlabeled samples can be generated in a much easier way. This observation has fostered the idea of adopting semisupervised learning techniques in hyperspectral image classification. The main assumption of such techniques is that the new (unlabeled) training samples can be obtained from a (limited) set of available labeled samples without significant effort/cost. In this paper, they develop a new approach for semisupervised learning which adapts available active learningmethods [8] (in which a trained expert actively selects unlabeled samples) to a self-learning framework in which the machine learning algorithm itself selects the most useful and informative unlabeled samples for classification purposes.

Continuous queries are used to monitor changes to time varying data and to provide results useful for online decision making. Typically a user desires to obtain the value of some aggregation function over distributed data items, for example, to know value of portfolio for a client; or the AVG of temperatures sensed by a set of sensors. In these queries a client specifies a coherency requirement as part of the query. They present a low-cost, scalable technique [9] to answer continuous aggregation queries using a network of aggregators of dynamic data items. In such a network of data aggregators, each data aggregator serves a set of data items at specific coherencies. Just as various fragments of a dynamic webpage are served by one or more nodes of a content distribution network, the technique involves decomposing a client query into subqueries and executing subqueries on judiciously chosen data aggregators with their individual subquery incoherency bounds. They provide a technique for getting the optimal set of subqueries with their incoherency bounds which satisfies client query’s coherency requirement with least number of refresh messages sent from aggregators to the client.

Effectively utilizing readily available auxiliary data to improve predictive performance on new modeling tasks is a key problem. In this research, the goal is to transfer knowledge between sources of data, particularly when ground-truth information for the new modeling task is scarce or is expensive to collect where leveraging any auxiliary sources of data becomes a necessity. Toward seamless knowledge transfer among tasks, effective representation of the data is a critical but yet not fully explored research area for the data engineer and data miner. Here, they present a technique [10] based on the idea of sparse coding, which essentially attempts to find an embedding for the data by assigning feature values based on subspace cluster membership. They modify the idea of sparse coding by focusing the identification of shared clusters between data when source and target data may have different distributions. In the paper, they point out cases where a direct application of sparse coding will lead to a failure of knowledge transfer.

The existing false data detection techniques consider false data injections during data forwarding only and do not allow any change on the data by data aggregation. However, they present a data aggregation and authentication protocol, called DAA [11] to integrate false data detection with data aggregation and confidentiality. To support data aggregation along with false data detection, the monitoring nodes of every data aggregator also conduct data aggregation and compute the corresponding small-size message authentication codes for data verification at their pairmates.

A Bayesian learning framework for adapting information extraction wrappers [12] with new attribute discovery, reducing human effort in extracting precise information from unseen Web sites. The approach aims at automatically adapting the information extraction knowledge previously learned from a source Web site to a new unseen site, at the same time, discovering previously unseen attributes. Two kinds of text-related clues from the source Web site are considered. The first kind of clue is obtained from the extraction pattern contained in the previously learned wrapper. The second kind of clue is derived from the previously extracted or collected items. A generative model for the generation of the site-independent content information and the sitedependent layout format of the text fragments related to attribute values contained in a Web page is designed to harness the uncertainty involved. Bayesian learning and expectation-maximization (EM) techniques are developed under the proposed generative model for identifying new training data for learning the new wrapper for new unseen sites. Previously unseen attributes together with their semantic labels can also be discovered via another EM-based Bayesian learning based on the generative model.

III. ARCHITECTURE DIAGRAM

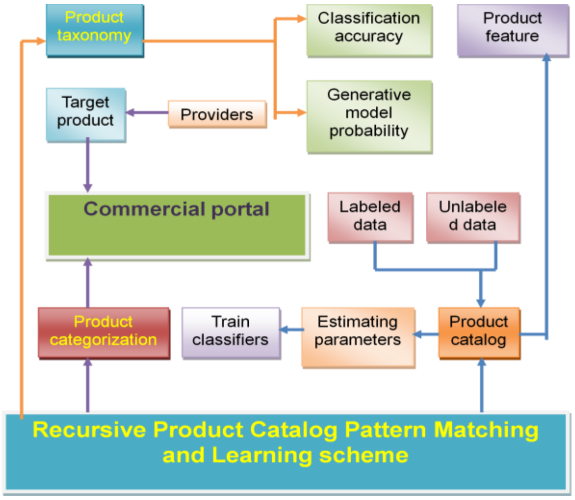

Fig. 1. The architecture of recursive product catalog pattern matching and learning for categorization of products in commercial port.

The phases involved in the proposed scheme are:

A. Generative Semi-supervised Learning

B. Recursive Product Catalog Pattern Learning

C. Recursive Pattern Matching for Catalog Integration

A. Generative semi supervised learning

The semi-supervised setting with labeled and unlabeled data. To find maximum a posterior (MAP) parameter estimates. No labels found for unlabeled data**.** The used Expectation-Maximization (EM) technique find locally MAP parameter estimates for generative model. EM technique applied to case of labeled and unlabeled data with naive Baye's yields better product catalog appropriation. The Naive Bayes classifier is built from limited amount of labeled training data. To perform classification of unlabeled data with naive Bayes model of probabilities associated with each class. To rebuild a new naive Baye's classifier using all the data labeled and unlabeled using the estimated class probabilities as true class labels. The unlabeled product catalog taxonomy are treated as several sub classes according to these estimated class probabilities.

B. Recursive Product Catalog Pattern Learning

The Iterative EM did not refer to all mixture components**.** The Multiple mixture component generative model account for many-to-one correspondence between mixture components and classes. The Address two missing values for each unlabeled product catalog taxonomy class and sub-topic. Even for the labeled data called missing values even class is known its sub-topic is not used. EM to estimate local MAP generative parameters. The pattern learning is done through recursive algorithm to alternates between estimating values of missing class and sub-topic labels. To calculate pattern learning parameters using estimated labels. The Recursive learning converges to high probability parameter estimates. The generative model used for product catalog taxonomy integration. It specifies a separation between mixture components and classes represents pre-determined and deterministic many-to-one mapping between mixture components and classes.

C. Recursive Pattern Matching for Catalog Integration

The Pattern Matching for catalog integration is done through recursive learned outcomes begins by maximizing on a very smooth, convex surface that is only remotely related to true probability surface of target product catalog. Initially find global maximum of simple pattern match surface**.** To change the surface of master taxonomy to become both more dense and close to true probability surface of target product catalog. If follow the original maximum as the surface gets more complex then when original surface is given still have a highly probable maximum. It avoids many of local maxima from recursive learning**.** To alter the class-to-cluster correspondence based on classification of each labeled example after deterministic annealing is complete. The Recursive pattern matching improves accuracy.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In the section the performance of recursive product catalog pattern matching and learning for categorization of products in commercial port through Java is evaluated. To confirm the analytical results, Recursive Pattern Matching for Catalog Integration evaluated the performance of technique is implemented. The performance of the system is evaluated by the following metrics.i.Execution time ii.Separation cost iii.No of EM iteration

TABLE I. EXECUTION TIME

Number of Products	**Execution Time in Existing System(S)**	**Execution Time in Proposed System(S)**
5	45	40
10	49	45
15	53	50
20	58	53
25	61	59

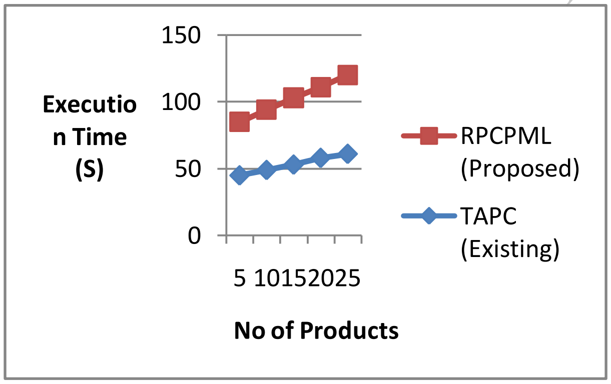

Fig. 2. Execution Time

Fig.2. demonstrates the Execution time. X axis represents the number of Products whereas Y axis denotes the Execution time using both the proposed Recursive product catalog pattern matching and learning. When the number of products increased, Execution time rate gets decreases accordingly. The rate of Execution Time is illustrated using the existing Taxonomy aware product categorization and proposed. Recursive product catalog pattern matching and learning Fig. 2. Shows better performance of proposed Recursive product catalog pattern matching and learning in terms of No of products density than existing and proposed Recursive product catalog pattern matching and learning. RPCPML achieves 15 to 25% less Execution Time variation when compared with existing system.

TABLE II. SEPERATION COST

Number of Products	**Separation Cost in ExistingSystem**	**Separation Cost in Proposed System**
5	23	18
10	32	29
15	45	33
20	58	47
25	62	55

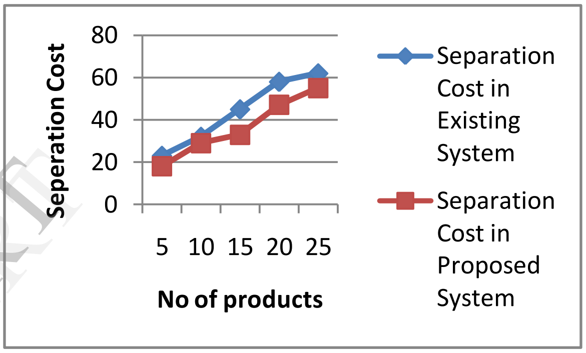

Fig. 3. Separation Cost

Fig. 3. demonstrates the Separation Cost. X axis represents the number of Products whereas Y axis denotes Separation Cost the using both the TAPC and the proposed RPCPML. When the number of Products increased, Separation also gets increases accordingly. The Separation Cost is illustrated using the existing TAPC and the proposed RPCPML. Fig. 3. shows better performance of Proposed RPCPML in terms of Products than existing TAPC and the proposed RPCPML.RPCPML achieves 20 to 35% less Separation Cost variation when compared with existing system.

TABLE III. NO .OF EM ITERATION

Number of Products	**No of EM iteration in Existing System**	**No of EM iteration in Proposed System**
5	8	5
10	10	8
15	12	9
20	14	10
25	16	12

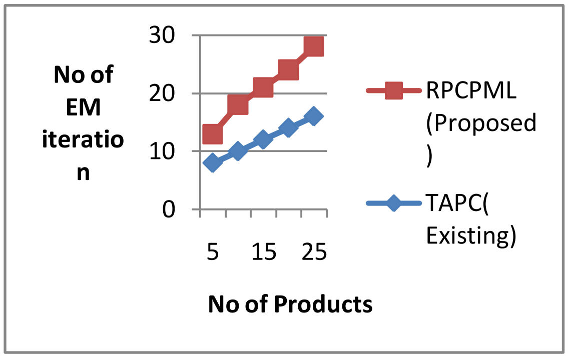

Fig. 4. No .of EM iteration

Fig. 4. demonstrates the No of EM iteration. X axis represents number of Products whereas Y axis denotes the No of EM iteration using both the TAPC and proposed RPCPML Technique. When the number of Products increases the no of EM iteration also gets decreased. Fig. 4. shows the effectiveness of No of EM iteration over different number of products than existing TAPC and the proposed RPCPML. RPCPML achieves 30% to 50% more No of EM iteration when compared with existing schemes.

V. Conclusion

The Recursive Product Catalog Pattern Matching and Learning scheme for product categorization of commercial portal based on target product coming from other providers. To prove the Product taxonomy has high positive correlation between generative model probability and classification accuracy. Finally, the analysis results to the design of a RPCPML to identify and apply the best design parameter settings in Java are applied. The proposed scheme is implemented, and conducted comprehensive performance analysis and evaluation, which showed its efficiency and advantages over existing schemes.

Acknowledgment

The authors wish to thank the Computer Science Department of Paavai Engineering College for theirl support to carry out this project successfully..

References

[1] V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M. Kan, and K.R. McKeown, "SIMFINDER: A Flexible Clustering Tool for Summarization," Proc. NAACL Workshop Automatic Summarization, 2001 pp. 41-49.

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

[3] Jeffrey Junfeng Pan, Sinno Jialin Pan, Jie Yin, Lionel M. Ni and Qiang Yang, "Tracking Mobile Users in Wireless Networks via Semi-Supervised Co-Localization", 2011 IEEE.

[4] Wei Zhang, Zhouchen Lin and Xiaoou Tang, "Learning Semi-Riemannian Metrics for Semisupervised Feature Extraction", Ieee Transactions On Knowledge And Data Engineering,April 2011, Vol. 23, No. 4.

[5] Emilio Miguela´n ez, Pedro Patro´ n, Keith E. Brown, Yvan R. Petillot, and David M. Lane , "Semantic Knowledge-Based Framework to Improve the Situation Awareness of Autonomous Underwater Vehicles", Ieee Transactions On Knowledge And Data Engineering, May 2011 Vol. 23, No. 5.

[6] Yun Yang and Ke Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations", Ieee Transactions On Knowledge And Data Engineering, February 2011, Vol. 23, No. 2.

[7] Renchu Guan, Xiaohu Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang, "Text Clustering with Seeds Affinity Propagation", IEEE Transactions On Knowledge And Data Engineering, April 2011,Vol. 23, No. 4.

[8] Inmaculada Dópido, Jun Li, Prashanth Reddy Marpu, Antonio Plaza, José M. Bioucas Dias, and Jon Atli Benediktsson, "Semisupervised Self-Learning for Hyperspectral Image Classification", IEEE Transactions On Geoscience And Remote Sensing, July 2013, Vol. 51, No. 7.

[9] Rajeev Gupta and Krithi Ramamritham, "Query Planning for Continuous Aggregation Queries over a Network of Data Aggregators", IEEE Transactions On Knowledge And Data Engineering, June 2012, Vol. 24, No. 6.

[10] Brian Quanz, Jun (Luke) Huan, and Meenakshi Mishra, "Knowledge Transfer with Low-Quality Data: A Feature Extraction Issue", IEEE Transactions On Knowledge And Data Engineering, October 2012,Vol. 24, No. 10.

[11] Suat Ozdemir and Hasan Çam, "Integration of False Data Detection With Data Aggregation and Confidential Transmission in Wireless Sensor Networks", IEEE/ACM Transactions On Networking, June 2010,Vol. 18, No. 3.

[12] Tak-Lam Wong and Wai Lam, "Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach', IEEE Transactions On Knowledge And Data Engineering, April 2010, Vol. 22,No.4.

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

[14] D.R. Radev, H. Jing, M. Stys, and D. Tam, "Centroid-Based Summarization of Multiple Documents," Information Processing and Management: An Int'l J.2004, vol. 40, pp. 919-938.

[15] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," Proc. 14th Int'l Joint Conf. Artificial Intelligence (IJCAI), pp. 448-453, 1995.

[16] P. Liang, H. Daume´ III, and D. Klein, "Structure Compilation: Trading Structure for Features," Proc. Int'l Conf. Machine Learning (ICML), 2008.