An Iterative Approach for Web Catalog Integration with Support Vector Machines

Ing-Xiang Chen, Jui-Chi Ho, and Cheng-Zen Yang

Department of Computer Science and Engineering
Yuan Ze University, Taiwan, R.O.C.
{sean,ricky,czyang}@syslab.cse.yzu.edu.tw

Abstract. Web catalog integration is an emerging problem in current digital content management. Past studies show that more improvement on integration accuracy can be achieved with advanced classifiers. Because Support Vector Machine (SVM) has shown its supremeness in recent research, we propose an iterative SVM-based approach (SVM-IA) to improve the integration performance. We have conducted experiments of real-world catalog integration to evaluate the performance of SVM-IA and cross-training SVM. The results show that SVM-IA has prominent accuracy performance, and the performance is more stable.

1 Introduction

Web catalog integration is an emerging problem in current digital content management [1, 6–8]. For example, a B2C company such as Amazon may want to merge catalogs from several on-line vendors into its catalog to provide customers versatile contents. As noted in [1], catalog integration is more than a classification task because if some implicit source information can be exploited, the integration accuracy can be highly improved. In [1], an enhanced Naive Bayes classifier (NB-AS) is proposed and its improvements are justified.

Recently, several studies [2–5] have shown that Support Vector Machine (SVM) achieves better classification accuracy on average. In [2], a *cross-training* SVM (SVM-CT) approach is proposed to improve the accuracy by extracting the implicit relationships between the source and the destination catalogs. However, SVM-CT outperforms SVM in only nearly half the cases. In addition, the cross-training process is very time-consuming. In [4], a *topic restriction* approach is proposed to improve NB and SVM by restricting the classification of any document to a small set of candidate destination categories. A candidate category is decided if more than a predefined number of common documents appear in both source and destination categories. Although this approach can significantly improve the performance of NB, it only slightly improves the performance of SVM. In [5], Zhang and Lee propose a *Cluster Shrinkage* approach in which the documents of the same category are shrunk into the cluster center. The conducted transductive SVM called CS-TSVM can consistently outperform NB-AS. However, because the shrinking process is applied to all documents, it suffers from tentatively misclassifying a document into an improper destination category.

In this paper, we propose an iterative-adapting approach on SVM called SVM-IA for catalog integration with pseudo relevance feedback. In SVM-IA, the training set

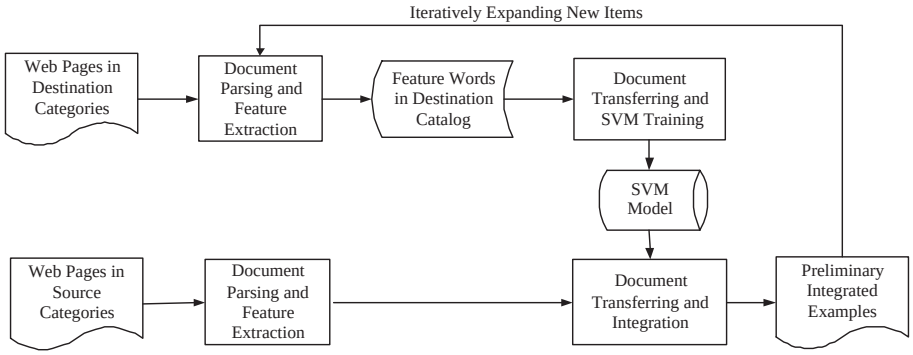

Fig. 1. The integration process of SVM-IA.

is iteratively expanded with newly integrated items to retrain the SVM classifier. With these adapted hyperplanes, the integration accuracy is thus improved. Since the expended features are classified first, the possibility of misclassification is reduced.

We have conducted several experiments with real-world catalogs from Yahoo! and Google. We have also compared SVM-IA with SVM and SVM-CT. The results show that SVM-IA outperforms SVM-CT on average, and the performance of SVM-IA is very stable in most cases.

2 Iterative-adapting SVM

In SVM-IA, the flattened source catalog S with a set of m categories S_1, S_2, ..., S_m is intended to be merged into the flattened destination catalog D with a set of n categories D_1, D_2, ..., D_n. Since a binary SVM can solve only two-class classification problems, we adopt a "one-against-all" strategy to decompose a multi-class problem into a set of binary SVM problems. Positive training data are composed of the feature information extracted from the destination class, and negative training data from other non-destination classes as in [9]. A set of binary SVM classifiers are then trained for the integration process of each destination category.

2.1 Iterative Integration Process

Figure 1 shows the catalog integration process of SVM-IA. The set of documents in the destination catalog is parsed first to extract the feature words as the training input of the SVM. In feature extraction, the stopwords are removed and the remaining words are the features for training. The SVM classifier is trained with the positive and negative training examples extracted from the target category and other destination categories. After the training process, a cutting hyperplane is formulated for future classification tasks. When the classification is finished, an integration iteration is completed.

The integration process can be repeated to find a more suitable hyperplane. The adaptation is performed by iteratively adding the newly integrated source documents

Table 1. The experimental categories.

Category	Yahoo!	Y-G	Y Test	Google	G-Y	G Test
Autos	/Recreation/Automotive/	1732	436	/Recreation/Autos/	1090	451
Movies	/Entertainment/Movies_Film/	1801	211	/Arts/Movies/	612	222
Outdoors	/Recreation/Outdoors/	7266	1346	/Recreation/Outdoors/	5184	1381
Photo	./Visual_Arts/Photography/	1921	710	/Arts/Photography/	5721	727
Software	/Computers_Internet/Software/	1637	221	/Computers/Software/	2392	227
Doc Sum		14357	2924		14999	3017

into the training set. Since these integrated source documents may have implicit information of the source catalog, the hyperplane can be adapted to have better separation performance. In our study, the well-known linear kernel function was used in the SVM classifier. *SVM*^{light} [11] was used as our SVM tool.

2.2 Feature Expansion

In the integration phase, the feature words of the source documents that have been integrated are incorporated as the implicit catalog information to re-train the SVM classifiers. There are two thresholds to control the number of expanded feature words. One is the term frequency, the number of term occurrences in the integrated source documents. Another is the document frequency, the number of documents in which the term appears. If two documents belong to the same category in S, they may have strong semantic relationships and are more likely to belong to the same category in D. Therefore, iteratively expanding new features from the source documents will be beneficial for the SVM classifiers to learn the semantics between feature information and enhance the classifiers in the destination catalog.

An SVM-IA classifier constructs a hyperplane that separates the positive and negative examples by iteratively training new items from the source catalog with a maximum margin. After new items are iteratively added into the classifier and retrained, new support vectors are created to adjust the hyperplane. Since the hyperplane is supported by the combination of new source documents, the cutting hyperplane is automatically adjusted by the new support vectors and would be beneficial for catalog integration.

3 Experiments

We have conducted experiments with real-world catalogs from Yahoo! and Google to study the performance of SVM-IA with *SVM*^{light}. The experimental results show that SVM-IA consistently improves SVM in all cases, and outperforms SVM-CT on average.

3.1 Data Sets

Five categories from Yahoo! and Google were extracted in our experiments. Table 1 shows these categories and the number of the extracted documents after ignoring the

Table 2. The accuracy of catalog integration from Google to Yahoo!.

	SVM	CT1	CT2	CT3	IA1	IA2	IA3
Autos	89.43%	90.11%	90.80%	89.43%	93.79%	93.79%	93.79%
(435)	(389)	(392)	(395)	(389)	(408)	(408)	(408)
Movies	85.73%	90.09%	88.97%	87.98%	86.23%	85.95%	86.30%
(1423)	(1220)	(1282)	(1266)	(1252)	(1227)	(1223)	(1228)
Outdoors	91.16%	91.63%	90.70%	87.44%	94.42%	94.42%	94.42%
(215)	(196)	(197)	(195)	(188)	(203)	(203)	(203)
Photo	65.40%	63.29%	69.62%	63.71%	78.48%	81.01%	80.59%
(237)	(155)	(150)	(165)	(151)	(186)	(192)	(191)
Software	93.35%	95.05%	89.96%	94.06%	95.33%	95.47%	95.33%
(707)	(660)	(672)	(636)	(665)	(674)	(675)	(674)
Average	93.35%	95.05%	89.96%	94.06%	95.33%	95.47%	95.33%

Table 3. The accuracy of catalog integration from Yahoo! to Google.

	SVM	CT1	CT2	CT3	IA1	IA2	IA3
Autos	80.96%	88.30%	85.78%	86.70%	84.86%	85.78%	85.78%
(436)	(353)	(385)	(374)	(378)	(370)	(374)	(374)
Movies	93.39%	91.83%	88.11%	92.05%	95.54%	95.62%	95.62%
(1346)	(1257)	(1236)	(1186)	(1239)	(1286)	(1287)	(1287)
Outdoors	82.81%	91.40%	87.33%	90.50%	86.43%	86.43%	86.43%
(221)	(183)	(202)	(193)	(200)	(191)	(191)	(191)
Photo	81.52%	94.79%	82.94%	92.89%	86.73%	87.20%	88.15%
(211)	(172)	(200)	(175)	(196)	(183)	(184)	(186)
Software	90.28%	96.06%	96.20%	95.77%	93.80%	93.94%	93.94%
(710)	(641)	(682)	(683)	(680)	(666)	(667)	(667)
Average	89.12%	92.51%	89.30%	92.10%	92.20%	92.44%	92.51%

documents that could not be retrieved and removing the documents with error messages. As in [1, 2], the documents appearing in only one category were used as the destination catalog D, and the common documents were used as the source catalog S. The number of distinct common documents is 2870. However, because some documents may appear in more than one category of the same catalog, the number of test documents may slightly vary in Yahoo! and Google. Thus, we measured the accuracy by the following equation.

$$\frac{\text{Number of docs correctly classified into } D_i}{\text{Total number of docs in the test dataset}}$$

In the processing, we used the stopword list in [10] to remove the stopwords.

3.2 Experimental Settings

In our experiments, both the cross-training and iterative-adapting techniques were employed on SVM to test how much they can enhance a purely text-based SVM learner. In [2], the label attributes extracted from the D_A catalog are considered useful predictors for the D_B catalog by adding extra $|A|$ labels. Therefore, in the SVM-CT implementation, a document $d \in D_B - D_A$ is submitted to the SVM ensemble $S(A, 0)$, which gives a score $w_{c_A} \cdot d + b_{c_A}$ for each class $c_A \in A$. These scores are inserted into the $|A|$ columns as label attributes. To convert the scores into the term attributes, ordinary term attributes are scaled by a factor of f ($0 \leq f \leq 1$) and label attributes are scaled by $1 - f$. We followed the origin SVM-CT settings with $f = 0.95$ and $1 - f = 0.05$.

After the transformation of label attributes, every document $d \in D_A - D_B$ gets a new vector representation with $|T|$ + $|A|$ columns where $|T|$ is the number of term features. Then, these new term vectors are trained as $S(B, 1)$ to classify the test documents. As the algorithm reported in [2], the cross-training process can be repeated like a ping-pong way.

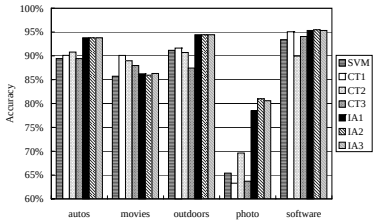

Fig. 2. The accuracy performance from Google to Yahoo!.

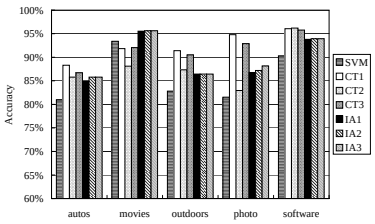

Fig. 3. The accuracy performance from Yahoo! to Google.

3.3 Results

Table 2 lists the experimental results of integrating Google's pages into Yahoo!'s categories. Table 3 lists the experimental results of reversely integrating Yahoo!'s pages into Google's categories. As listed in the two tables, we have measured the accuracy achieved by the following classifiers: SVM, cross-training SVM (SVM-CT), and iterative-adapting SVM (SVM-IA). IA1, IA2, and IA3 separately represent the result by first, second, and third iterations of adding new features from the source catalog and retraining. Similarly, CT1 is the result of first cross-training with the label attributes extracted from the source catalog. The result of CT2 is based on the SVM-CT1 classifiers proceeding with the second cross-training, and so is the result of CT3 based on the SVM-CT2 classifiers.

Table 2 and Table 3 both show that SVM-IA consistently improves SVM after three iterations. In Table 2, the SVM-IA classifiers not only have sustaining improvements but also outperform SVM-CT in most categories. In /Recreation/Outdoors and /Arts/Photography, SVM-CT is even worse than pure SVM and the improvements are very unstable. Although in Table 3 SVM-CT have effective improvements in most categories after CT3, the overall improvements are not stable, and the accuracy in /Entertainment/Movies_Film is even worse than pure SVM. Figure 2 and Figure 3 further indicate that the accuracy of SVM-IA is stably improved, but SVM-CT has unstable accuracy performance. The reason of vastly unstable performance is that a large number of label attributes are altered in the subcategories of /Entertainment/Movies_Film in Yahoo! after cross-training process. The same situation also happened in /Recreation/Outdoors and /Arts/Photography in Google. These label changes resulted in wrong mappings between the subcategories, and would thus decreased the accuracy. Moreover, we found that the cross-training process was very time-consuming. This makes SVM-CT less feasible for large catalog integration.

4 Conclusions

In this paper, we have studied the effects of iterative-adapting approach to enhance the integration accuracy. We compared our approach with SVM and SVM-CT. The experimental results are very promising. It shows that our approach consistently achieves improvements on SVM classifiers and is on average superior to cross-training that has been proposed to improve SVM.

Several issues still need to be further discussed. First, generalizing the flat catalog assumption to the hierarchical catalog model is of the major interest for the catalog integration because hierarchical catalogs are more practical in real cases. Second, how to construct a systematical mechanism combining effective auxiliaries to enhance the power of SVM is a more difficult problem but needs further investigation. To conclude, we believe that the accuracy of catalog integration can be further improved with the assistance of more effective auxiliary information.

5 Acknowledgement

We would like to especially thank S. Godbole for his great help in our SVM-CT implementation.

References

1. Agrawal, R., Srikant., R.: On Integrating Catalogs. Proc. the 10th WWW Conf. (WWW10), (May 2001) 603–612
2. Sarawagi, S., Chakrabarti S., Godbole., S.: Cross-Training: Learning Probabilistic Mappings between Topics. Proc. the 9th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, (Aug. 2003) 177–186
3. Chen, I.-X., Shih, C.-H., Yang, C.-Z.: Web Catalog Integration using Support Vector Machines. Proc. 1st IWT, (Oct. 2004) 7–13
4. Tsay, J.-J., Chen, H.-Y., Chang, C.-F., Lin, C.-H.: Enhancing Techniques for Efficient Topic Hierarchy Integration. Proc. the 3rd Int'l Conf. on Data Mining (ICDM'03), (Nov. 2003) (657–660)
5. Zhang, D., Lee W. S.: Web Taxonomy Integration using Support Vector Machines. Proc. WWW2004, (May 2004) 472–481
6. Kim, D., Kim, J., Lee, S.: Catalog Integration for Electronic Commerce through Category-Hierarchy Merging Technique. Proc. the 12th Int'l Workshop on Research Issues in Data Engineering: Engineering e-Commerce/e-Business Systems (RIDE'02), (Feb. 2002) 28–33
7. Marron, P. J., Lausen, G., Weber, M.: Catalog Integration Made Easy. Proc. the 19th Int'l Conf. on Data Engineering (ICDE'03), (Mar. 2003) 677–679
8. Stonebraker, M., Hellerstein, J. M.: Content Integration for e-Commerce. Proc. the 2001 ACM SIGMOD Int'l Conf. on Management of Data, (May 2001) 552–560
9. Zadrozny., B.: Reducing Multiclass to Binary by Coupling Probability Estimates. In: Dietterich, T. G., Becker, S., Ghahramani, Z. (eds): Advances in Neural Information Processing Systems 14 (NIPS 2001). MIT Press. (2002)
10. Frakes, W., Baeza-Yates, R.: Information Retrieval: Data Structures and Algorithms. Prentice Hall, PTR. (1992)
11. Joachims, T.: Making Large-Scale SVM Learning Practical. In Scholkopf, B., Burges, C., Smola, A. (eds): Advances in Kernel Methods: Support Vector Learning. MIT Press. (1999)