Notice of Violation of IEEE Publication Principles

"Business Blog Mining Based on Hierarchical SVM","
by Huiling Wang,
in the Proceedings of the 2008 International Symposium on Knowledge Acquisition and Modeling, December 2008, pp 837-841

After careful and considered review of the content and authorship of this paper by a duly constituted expert committee, this paper has been found to be in violation of IEEE's Publication Principles.

This paper contains substantial portions of the original text from the papers cited below. The original text was copied without attribution and without permission.

"Machine Learning Techniques for Business Blog Search and Mining,"
by Yun Chen, Flora S. Tsai and Kap Luk Chan,
in Expert Systems with Applications, 35 (2008) October 2004, pp 581-590, Elsevier

2008 International Symposium on Knowledge Acquisition and Modeling

Business Blog Mining Based on Hierarchical SVM

Huiling Wang
Management School,
Jinan University, Guangzhou, Guangdong, 510632, China
Wanghui_ling@yahoo.com.cn

Abstract

Blog has rapidly gained in popularity in the past decade. In particular, the growth of business blogs, written by professionals and providing commentary on businesses and companies, opens up new opportunities for developing blog knowledge mining techniques. In this paper, we propose hierarchical SVM model for blog mining. We implement the model in our database of business blogs with the aim of achieving better effect of blog mining. The hierarchical model is able to segment the business blogs into separate areas, which is useful for knowledge detection on the blogosphere. Various term-weighting schemes and factor values were also studied in detail, which reveal interesting patterns in our database of business blogs. Our multi-functional business blog mining system is found to be different from existing blog mining system, as it aims to provide better effects of business blog mining.

1. Introduction

In the past decade, blogs have rapidly gained in popularity through out the world. According to David Sifry of Technorati, over 37.3 million blogs were being tracked by Technorati in May 2006; on average, a new weblog is created every second. The business world has experienced significant influence by the blogosphere. A hot topic in the blogosphere may affect a product's life cycle[1]. An exposure of an inside story in the blogosphere may influence a company's reputation.

Focusing on mining in the individual business blogs written by professionals, as well as the corporate blogs, we propose a novel hierarchical classification method that generalizes support vector machine based on the results of support vector clustering method that are structured in a way that mirrors the class hierarchy. The grouping of the individual blogs into meta-class is determined by the class distributions described by the support vector clustering method.

2. Related works

This section analyses related work in developing blog-specific mining and extraction of useful information from blogs.

2.1. Information extraction from blogs

Current blog analysis focuses on extracting useful information from blog entry collections, and determining certain trends in the blogosphere. Natural Language Processing (NLP) algorithms have been used to determine the most important keywords and proper names within a certain time period from thousands of active blogs, which can automatically discover trends across blogs, as well as detect key persons, phrases and paragraphs[2]. A study on the propagation of discussion topics through the social network in the blogosphere developed algorithms to detect the long-term and short-term topics and keywords, which were then validated with real blog entry collections. On evaluating the suitable methods of ranking term significance in an evolving RSS feed

978-0-7695-3488-6/08 $25.00 © 2008 IEEE
DOI 10.1109/KAM.2008.96

corpus, three statistical feature selection methods were implemented: χ^2, mutual information (MI) and information gain (I), and the conclusion was that χ^2 method seems to be the best among all, but full human classification exercise would be required to further evaluate such a method.

By utilizing the hierarchical SVM structure, the classification can be decomposed into a set of smaller problems corresponding to hierarchical splits in the tree. As we show, each of these smaller problems can be solved accurately and efficiently. Moreover, each sub-problem is smaller than the original problem, and it is sometimes possible to use a much smaller set of features for each. We can select more proper domain term features in the smaller sub-problem.

2.2 Support vector machines for blog mining

Support vector (SV) clustering has been recently derived from the single-class support vector machine for estimating the underlying probability distribution. The mathematical formulation of the SV clustering is as follows. Given a set of input patterns, $\{\mathbf{x}_i\} \subseteq X$, the support vector method for clustering is to find

$$\min imize \ \ R^2 + C\sum_i \zeta_i$$

$$subject\ to \ \ \|\Phi(x_i) - a\| \le R^2 + \zeta_i \quad (1)$$

where R is the radius, a is the center of the enclosing sphere, ξ_i is a slack variable, and C is a constant controlling the penalty of noise. To solve this problem, we introduce the Lagrangian

$$L(R, a, \zeta_i, \alpha_i) = R^2 + Csum_i \zeta_i - \sum_i \alpha_i \zeta_i$$

$$-\sum_i \beta_i (R^2 + \zeta_i - \|\Phi(x_i) - a\|^2) \qquad (2)$$

where $B = (\beta_1, \beta_2, \ldots, \beta_\ell)$ and $A = (\alpha_1, \alpha_2, \ldots, \alpha_\ell)$ are the ℓ nonnegative Lagrange multipliers associated with the two constrains in Eq. (1). Differentiating L with respect to R, $\mathbf{a}$, and ξ_i, and setting them to zero, we obtain[3]

$$C - \alpha_i - \beta_i = 0, \ \sum_i \beta_i = 1 \quad (3)$$

$$\alpha = \sum_i \beta_i \Phi(x_i) \qquad (4)$$

Substituting Eqs. (3) and (4) into (2), we see this problem reduces to its dual problem

$$\max imize \quad \mathrm{L} = \sum_i \Phi(x_i) \cdot \Phi(x_i) \beta_i$$

$$-\sum_{i,j} \beta_i \beta_j \Phi(x_i) \cdot \Phi(x_j)$$

$$subject\ to \quad \sum_i \beta_\mathrm{i} = 1 \ \ and \ \ 0 \le \beta_\mathrm{i} \le C, \forall i$$

（5）

The functional form of mapping $\Phi(\mathbf{x}_i)$ does not need to be known since it is implicitly defined by the choice of *kernel* function

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j) \qquad （6）$$

where ‘·’ represents the dot product. Throughout this paper we use the Gaussian kernel

$$\mathrm{K}(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j}) = \exp(-\mathrm{q}\|\mathbf{x}_\mathrm{i} - \mathbf{x}_\mathrm{j}\|^2) \qquad （7）$$

We can further compute the distance between input pattern and spherical center as

$$R^2(x) = \|\Phi(x) - a\|^2 = K(x, x) - 2\sum_i \beta_i K(x_i, x)$$

$$+\sum_{i,j} \beta_i \beta_j K(x_i, x) \qquad (8)$$

Appling the KKT conditions, we yield

$$\xi_i \alpha_i = 0$$

$$\beta_i (R^2 + \xi_i - \|\Phi(\mathbf{x}_i) - \mathbf{a}\|^2) = 0 \qquad （9）$$

Any point with $\beta_i > 0$ will be referred to as a *support vector*. The radius of the sphere R can be computed as

$$\mathrm{R} = \{\mathrm{R}(\mathrm{x}_\mathrm{i})| \ \ \beta_\mathrm{i} \text{ is in } (0，\mathrm{C})\} \quad (10)$$

Let us denote n_{sv} and n_{out} as the number of support vectors and outliers, respectively, and note the following results that are a consequence of the above constraints:

$$\mathrm{n}_{\mathrm{out}} + \mathrm{n}_{\mathrm{sv}} \ge 1/\mathrm{C}, \qquad \mathrm{n}_{\mathrm{out}} \le 1/\mathrm{C} \quad (11)$$

In order to label data points into clusters, we need to identify the connected components. We define an adjacency matrix $\lfloor A_{ij} \rfloor$ between pairs of points $\mathbf{x}_i$ and $\mathbf{x}_j$. $A_{ij} = 1$ (if for all y on the line segment connecting x_i and x_j, $R(y) \leq R$), while $A_{ij}=0$ (otherwise)

Clusters are then defined as the connected components of the graph induced by *A*. This labeling procedure is justified by the observation the nearest neighbors in the data space can be connected by a line segment that is contained in the high dimensional sphere. Checking of the line segment is implemented by sampling a number of points on the segment (for example, a value of 20). Note that the outliers are not classified by this procedure; they can be left unlabeled, or classified e.g., according to the cluster to which they are closest to. Here we go through a set of examples demonstrating the parameters used in SVC[4].

2.3 The proposed hierarchical SVM in blog mining

The proposed hierarchical blog mining model consists of three stages: (i) data preprocessing stage, (ii) unsupervised learning stage-hierarchical structure construction, and (iii) supervised learning stage-tree-node classifier training.

(i) Data preprocessing stage

The first stage in blog mining is to transform blog entries, which typically are the strings of characters into a representation suitable for learning algorithm and classification. Information retrieval research suggests that word stems work well as representation units and that their ordering in a document is of minor importance for many tasks[5]. This leads to an attribute value representation of text. Each distinct word corresponds to a feature. To avoid unnecessarily large feature vectors, words are considered as features only if they occur in the training dataset at least three times and if they are not “stop-words”.

(ii) Unsupervised learning stage

In the second stage, we introduce hierarchical structure construction, and perform the unsupervised support vector clustering method to construct the hierarchical structure. In our proposed model, we use the iterative approach to construct the hierarchical structure with the results of support vector clustering. It starts from a small value of *q* where one cluster occurs, and increases it to detect cluster splitting. When single point clusters start to break off or a large number of support vectors is obtained, then *C* is decreased. For a decision about when to stop dividing, we used the CS measure that is used to determine the cluster validity in the traditional clustering method. The spirit of CS measure is to hope that the distance is smaller the better between points belonging to the same cluster, and the distance is larger the better between points belonging to a different cluster. The definitions of compact measure and separated measure are stated below.

(iii) Supervised learning stage

This stage is tree-node classifier training. The hierarchical classification offers a lot of flexibility in designing the classifier system. For instance, one can replace the classifier at the internal nodes of the generated hierarchical structure with stronger ones. In the proposed approach, if an internal node has two branches, we use the binary SVM as the tree-node classifier. When an internal node has more than two branches, we evaluate several multi-class SVM approaches, such as one-against-one and one-against-rest, and choose the one with best accuracy as the tree-node classifier. Moreover, different feature selection methods can be used at each tree-node that is specific to the domain of the input data. Each sub-problem is smaller than the original problem, and it is sometimes possible to use a much smaller set of features for each.

3. Experiments

We have created a business blog data set, and five hundred and twenty nine business blog entries were collected in two phases. First, a small number of blog entries are manually collected from various CEOs's blog sites and business blog sites. Meaningful blog entries from these blog sites were extracted and stored into our database. Each blog entry is saved as a text file in its corresponding category, for further text preprocessing. For the preprocessing of the blog data, we performed lexical analysis by removing stopwords and stemming using the Porter stemmer. The text files are then used as the input for the Text to Matrix Generator (TMG) to generate the term-document matrix for input to the blog search and mining system[6]. The overview of the system is shown in Figure 1.

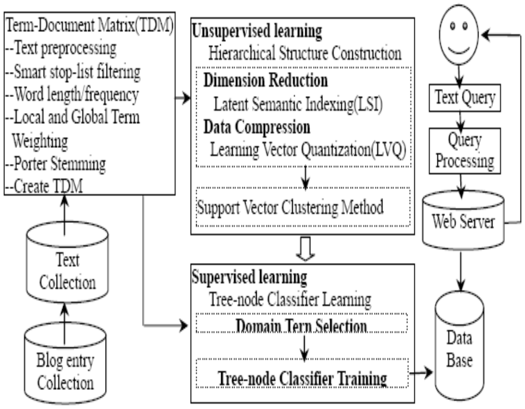
Term-Document Matrix(TDM)
--Text preprocessing
--Smart stop-list filtering
--Word length/frequency
--Local and Global Term Weighting
--Porter Stemming
--Create TDM
Text Collection
Blog entry Collection
Unsupervised learning
Hierarchical Structure Construction
Dimension Reduction
Latent Semantic Indexing(LSI)
Data Compression
Learning Vector Quantization(LVQ)
Support Vector Clustering Method
Supervised learning
Tree-node Classifier Learning
Domain Tern Selection
Tree-node Classifier Training
Text Query
Query Processing
Web Server
Data Base

Figure 1 Schematic diagram of the proposed hierarchical SVM for blog mining

The hierarchical structure we constructed is a category tree. In the category tree, a text will first be classified by the classifier at the root level into one or more lower level categories. It will then be further classified by the classifier(s) of the lower level category(ies) until it reaches a final category which could be a leaf category or an internal category. Due to the limited space, we eliminate the sub-tree where the number of branches from its root is more than nine[7]. To find the advantage of proposed hierarchical SVM classifier, we compare our approach with both non-hierarchical classifiers as well as hierarchical classifiers, and pay attention to the following two questions: (1) Does our hierarchical classifier based on support vector clustering method improve the classification performance when compared to a non-hierarchical SVM classifier? (2) How does our hierarchical method compare with other famous text categorization methods?

Table 1 shows the performance comparison of our proposed hierarchical SVM classification and the non-hierarchical SVM in the five most frequent categories. Our proposed hierarchical SVM classification has a better performance, but poor in the categories like "client", "customer", "business" and "firm". We now try to explain why the performance of those four categories is poorer than other categories. The topic "client" is a child node of the topic "user", which contains five sub-categories. We build a multi-class SVM classifier to distinguish these five sub-categories. As mentioned above, we apply a multi-class SVM using one-against-rest approach, each binary SVM distinguishes only one category with the other four categories. In our experiments, we find that these five categories are somewhat overlapped where they share many of the same entries[8]. So to distinguish these five categories becomes difficult.

Table 1 Comparison of the proposed hierarchical SVM with non-hierarchical SVM

	non-hierarchical SVM					hierarchical SVM
	1	2	3	4	5	
Market	97.2	96.4	98.5	97.4	99.1	98.6
Advertise	93.6	95.6	94.2	94.2	95.5	98.9
Product	67.7	73.1	74.9	75.2	75.8	96.4
Client	92.1	92.8	91.8	92.0	90.1	89.1
customer	85.9	86.3	88.4	88.2	88.1	82.6
Trade	68.5	75.2	77.1	76.9	76.8	98.9
Corporate	70.1	62.6	68.2	72.9	75.8	94.7
Business	82.7	84.3	84.8	85.6	83.4	85.1
Firm	86.3	87.1	84.9	86.1	84.2	83.3

4. Conclusions

This paper presents results using hierarchical SVM models for search and analysis of business blogs. To our knowledge, is the first such study focusing on business blogs. A multi-functional business blog directory has been successfully developed in mining the business blogs. The system implements a different approach from the existing blog search engines. The employed measures have proved to be appropriate for improving the precision and recall for our business blog data set. Our experiments on our data set of business blogs demonstrate how our blog-mining model can present the blogosphere in terms of topics with measurable keywords, hence tracking popular conversations and topics in the blogosphere. We hope that this work will contribute to the growing need and importance for search and mining of business blogs.

Acknowledgement

It is a project supported by National Social Science Foundation of China (04BTQ015), Natural Science (6300278)，and Philosophy and Social Science in Guangdong Province (07SJY020), the results of Management School of Jinan University.

References

[1] Chiang and Hao, 2003 J.-H. Chiang and P.-Y. Hao, A new kernel-based fuzzy clustering approach: Support vector clustering with cell growing, *IEEE Transactions on Fuzzy Systems* 11 (2003) (4), pp. 518–527.

[2] Kumar et al., 2002 S. Kumar, J. Ghosh and M.M. Crawford, Hierarchical fusion of multiple classifiers for hyperspectral data analysis, *Pattern Analysis and Applications* **5** (2002) (2), pp. 210–220 Spl. Issue on Fusion of Multiple Classifiers.

[3] Weiss et al., 1999 S.M. Weiss, C. Apte, F.J. Damerau, D.E. Johnson, F. J Oles and H. Goetz *et al.*, Maximizing text-mining performance, *IEEE Intelligent Systems* 14 (1999) (4), pp. 2–8.

[4] Ben-Hur et al., 2001 A. Ben-Hur, D. Horn, H.T. Siegelmann and V.N. Vapnik, Support vector clustering, *Journal of Machine Learning Research* 2 (2001), pp. 125–137.

[5] Berry and Browne, 2005 M.W. Berry and M. Browne, Understanding search engines: Mathematical modeling and text retrieval (2nd ed.), SIAM, Philadelphia, PA (2005).

[6] Chen and Hsieh, 2006 R.-C. Chen and C.-H. Hsieh, Web page classification based on a support vector machine using a weighted vote schema, *Expert Systems with Applications* 31 (2) (2006), pp. 427–435.

[7] Mishne and de Rijke, 2006 Mishne, G., & de Rijke, M. (2006). A Study of blog search. In *Proceedings of the ECIR '06*.

[8] Avesani et al., 2005 Avesani, P., Cova, M., Hayes, C., & Massa, P. (2005). Learning contextualised weblog topics. In *Proceedings of the WWW '05 workshop on the weblogging ecosystem: aggregation, analysis and dynamics*.