https://doi.org/10.20965/jaciii.2014.p0818

Paper:

Label Propagation for Text Classification Using Latent Topics

Akiko Eriguchi and Ichiro Kobayashi

Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
E-mail: {g0920506, koba}@is.ocha.ac.jp

[Received December 15, 2013; accepted June 1, 2014]

The objective of this paper is to raise the accuracy of multiclass text classification through Graph-Based Semi-Supervised Learning (GBSSL). In GBSSL, it is essential to construct a proper graph which expresses the relation among nodes. We propose a method to construct a similarity graph by employing both surface information and latent information to express similarity between nodes. Experimenting on a Reuters-21578 corpus, we have confirmed that our proposal works well in raising the accuracy of GBSSL in a multiclass text classification task.

Keywords: graph-based semi-supervised learning, label propagation, text classification, latent Dirichlet allocation

1. Introduction

Semi-Supervised Learning (SSL) is a method to give a label to a large amount of unlabeled data with a small amount of labeled data, e.g., self-training [1], co-training [2], transductive support vector machine (TSVM) [3], and graph-based SSL methods [4–6]. Graph-Based Semi-Supervised Learning (GBSSL) is a SSL method with labeled data and unlabeled data represented on a similarity graph. It is reported that GBSSL is superior to the major machine learning methods such as SVM and TSVM in text classification task [4].

The accuracy of GBSSL relies on which datum is selected from a data set to be labeled and on how a graph is constructed [6, 7]. Regarding the former issue, it is important how the informative training data is selected. In general, Active Learning tackles the problem of selecting training data, i.e., labeled data. Some methods were proposed in order to raise the accuracy of GBSSL [6, 8]. In the latter graph construction, it matters how relations between the nodes of a graph are represented, since graph construction "is more of an art, than science" [7]. The sparse graph construction for GBSSL has recently attracted attention. The k-Nearest Neighbor (k-NN) graph, which is constructed with each node linked to its k-NN nodes, is known as a basic sparse graph construction. The b-matching graph [9], which has predefined degree b for all of its nodes, is the state-of-the-art graph for GBSSL, although it takes much more time to construct ($O(bn^3)$) than the k-NN graph ($O(n^2+kn\log n)$). Ozaki et al. [10] reported that nodes with a high degree, called hub nodes, lower the accuracy of GBSSL and proposed the *mutual* k-NN graph, which limits degrees of all nodes in a graph to k or less. They confirmed that the mutual k-NN graph achieves accuracy nearly equal to that of the b-matching graph in word sense disambiguation task.

In graph-based classification for text, a weighted graph is constructed based on a certain relation between nodes, i.e., documents, – similarity often used to express the relation between nodes in a graph. We use two types of similarity: the one is between surface information obtained by document vector [11] and the other is between latent information obtained by a topic model (Latent Dirichlet Allocation [12]). In this paper, we propose employing both surface information and latent information at a ratio of $(1-\alpha):\alpha$ $(0 \leq \alpha \leq 1)$ to construct a similarity graph for GBSSL and investigate the optimal α for raising the accuracy in GBSSL. We run GBSSL on a multiclass text classification task. The precision recall break even point (PRBEP) is introduced as a performance measure of our proposed method. Experimenting on the Reuters-21578 corpus, we confirm that using both surface information and latent information improves the accuracy of GBSSL rather than the methods using either of two types of information.

2. Graph-Based Semi-Supervised Learning for Text Classification

Our proposed GBSSL in multiclass text classification is detailed in this section.

2.1. Graph Construction

We use a weighted undirected graph $G=(V,E)$ whose node represents a document and whose edge represents similarity between nodes. Similarity is regarded as weight. V and E represent the nodes and edges of a graph, respectively. The graph G is represented as an adjacency matrix, and $w_{ij} \in \boldsymbol{W}$ represents similarity between the nodes i and j. In the case of the GBSSL, similarity between nodes is formed as shown in Eq. (1),

$$w_{ij} = sim(\boldsymbol{x}_i, \boldsymbol{x}_j) * \delta(j \in K(i)), \quad . \; . \; . \; . \; . \; . \quad (1)$$

where $K(i)$ is a set of i's k-NN nodes, and $\delta(z)$ is 1 if z is true, otherwise 0.

2.2. Metric: Similarity Between Texts

In constructing a graph to represent a relation among documents, we use cosine similarity (sim_{cos}) of document vectors (*tf-idf* vectors) [11] as similarity based on surface information. Cosine similarity is often used as a similarity measure in text clustering [4, 13]. We also use similarity (sim_{latent}) of latent topic distributions by a topic model as similarity based on latent information. We use similarity (sim_{latent}) based on latent information and similarity (sim_{cos}) based on surface information in the proportion of $\alpha : (1-\alpha)$ $(0 \leq \alpha \leq 1)$. We define the sum of sim_{latent} and $sim_{surface}$ as sim_{nodes} (Eq. (2)).

In Eq. (2), P and Q represent latent topic distributions of documents S and T, respectively. We estimate the latent topic distribution of a document by means of Latent Dirichlet Allocation (LDA) [12]. We use $L2$ norm distance (Eq. (5)) for the metric between topic distributions. If they are similar, the value of $L2$ norm distance becomes 0, which contradicts the idea of similarity. We introduce a standard sigmoid function to get the value of $L2$ norm distance ranging between $[0,1]$. The sim_{latent} in Eq. (2) is expressed by Eq. (4).

$$\begin{aligned} sim_{nodes}(S,T) \equiv\ & \alpha * sim_{latent}(P,Q) \\ & + (1-\alpha) * sim_{surface}(S,T), \quad (2) \end{aligned}$$

$$sim_{surface}(S,T) = \cos(tfidf(S), tfidf(T)), \quad (3)$$

$$sim_{latent}(P,Q) = \frac{2}{1+\exp^{L^2(P,Q)}}, \quad (4)$$

$$L^2(P,Q) = \int (P(\boldsymbol{x}) - Q(\boldsymbol{x}))^2 d\boldsymbol{x}. \quad (5)$$

2.3. Label Propagation

We use label propagation [5, 6] – a type of GBSSL – to classify texts. It estimates the value of a label based on the cluster assumption that nodes mutually linked in a graph should belong to the same category.

The goal of the learning procedure is to estimate the values $\boldsymbol{f}$ for given n nodes. The values are obtained as the solution (Eq. (8)) to the following objective function of an optimal problem (Eq. (6)). The first and second terms in Eq. (6) express the deviation between an estimated value and a correct value of training data and the difference between the estimated values of nodes that are next to one another in the adjacency graph, respectively. Here, λ (>0) is a parameter balancing both terms. l indicates the number of training data among all n nodes in a graph. $y^{(i)}$ is a label of node i. $f^{(i)}$ is an estimation value for a label to be given to node i . $w^{(i,j)}$ is a weight between nodes i and j. Eq. (6) is transformed into Eq. (7) through Laplacian matrix $\boldsymbol{L}$ $(\equiv \boldsymbol{D}-\boldsymbol{W})$. Here, $\boldsymbol{W}$ indicates an adjacency matrix of texts. $\boldsymbol{D}$ is a diagonal matrix, each of whose diagonal elements is equal to the sum of elements in each row (or column) of $\boldsymbol{W}$. Minimizing the first and second terms in Eq. (6) and solving it, we get a closed-form solution (Eq. (8)) using $\boldsymbol{L}$. $\boldsymbol{I}$ is an identity matrix.

$$\begin{aligned} J(\boldsymbol{f}) &= \sum_{i=1}^{l} \left(y^{(i)} - f^{(i)}\right)^2 + \lambda \sum_{i<j} w^{(i,j)} \left(f^{(i)} - f^{(j)}\right)^2 \quad (6) \\ &= \|\boldsymbol{y} - \boldsymbol{f}\|_2^2 + \lambda \boldsymbol{f}^T \boldsymbol{L} \boldsymbol{f}, \quad (7) \end{aligned}$$

$$\boldsymbol{f} = (\boldsymbol{I} + \lambda \boldsymbol{L})^{-1} \boldsymbol{y}. \quad (8)$$

3. Experimental Settings

We use a Reuters-21578 corpus data set[1] collected from the Reuters newswire in 1987 as target documents for the multiclass text classification. It consists of English news articles classified into 135 categories. We use the "ModApte" split to get training documents, i.e., labeled data, and test documents, i.e., unlabeled data, extract documents that have only a title and text body, and apply the stemming and the stop-word removal processes to the documents. We use the 10 most frequent out of the 135 potential topic categories – *earn, acq, grain, wheat, money-fx, crude, trade, interest, ship,* and *corn* [3, 4].

We prepare 15 data sets, each of which consists of 3,299 common test data and 20 training data. We use 11 types of categories for training data: the above mentioned 10 categories and the category (*other*) that indicates 125 categories. The categories of 20 training data are chosen randomly only if one of the 11 categories is chosen at least once.

In a graph construction for GBSSL, we estimate latent topics in target documents with LDA using Collapsed Gibbs sampling. Iteration number of the sampling is 200. We regard the average of topic distributions after 5 trials as topic distribution in target documents. The optimal number of latent topics in target documents is decided by averaging 5 trials of perplexity values (Eq. (9)). Here, N is the number of all words in target documents. w_{mn} is the n-th word in the m-th document. θ is the occurrence probability of latent topics for documents. ϕ is the occurrence probability of words for each latent topic.

$$Perplexity(\boldsymbol{w}) = \exp\left(-\frac{1}{N}\sum_{m,n} \log\left(\sum_{z} \theta_{mz}\phi_{zw_{mn}}\right)\right). \quad (9)$$

In label propagation, we propagate labels on the above graph. The number of nodes in a graph is $|V_{l+u}| = n$ $(=3,319)$, and similarity between nodes is based on both surface information and latent information at a ratio of $(1-\alpha) : \alpha (0 \leq \alpha \leq 1)$ (Eq. (2)). Parameter α in Eq. (2) is varied from 0.0 to 1.0 every 0.1. We choose parameter k in the k-NN graph from $\{2, 10, 50, 100, 250, 500, 1000, 2000, n\}$ and parameter λ in the label propagation method from $\{1, 0.1, 0.01, 1e-4, 1e-8\}$. Using 5 of the 15 data sets, we decide on a pair of optimal parameters (k, λ) for each category. We categorize the remaining 10 data sets with the predeter-

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

Table 1. The optimal parameters (k, λ) for each category.

Category\\α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
earn	(50, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)	(1000, 1)
acq	(500, 0.1)	(250, 0.1)	(250, 0.01)	(100, 0.01)	(100, 1e-8)	(50, 0.1)	(10, 1e-8)	(10, 1e-8)	(10, 1e-4)	(250, 0.01)	(500, 1e-4)
money-fx	(2, 1)	(2, 1)	(10, 0.1)	(2, 0.1)	(2, 1)	(2, 1)	(50, 1e-4)	(50, 0.01)	(2, 1e-8)	(50, 0.01)	(10, 0.1)
grain	(100, 0.1)	(50, 1)	(50, 1)	(10, 1)	(50, 1e-8)	(10, 1)	(10, 1)	(50, 1e-8)	(50, 1e-8)	(50, 1)	(50, 1)
crude	(10, 1)	(50, 0.1)	(50, 0.01)	(100, 1e-8)	(10, 0.01)	(10, 1e-8)	(50, 1e-8)	(2, 1e-4)	(50, 1e-8)	(2, 1e-8)	(50, 1e-8)
trade	(10, 1)	(10, 1e-8)	(10, 1e-8)	(10, 1e-4)	(10, 1e-8)	(10, 1e-4)	(10, 1e-8)	(2, 0.01)	(10, 1e-8)	(10, 1e-8)	(10, 0.1)
interest	(10, 0.1)	(10, 1)	(10, 0.1)	(10, 1e-8)	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(100, 1e-8)	(100, 1e-8)
ship	(10, 1)	(100, 1e-8)	(50, 0.1)	(10, 1e-8)	(10, 0.1)	(10, 0.1)	(10, 0.1)	(10, 0.1)	(2, 1)	(10, 0.1)	(10, 0.1)
wheat	(100, 0.01)	(100, 1e-8)	(100, 1e-8)	(50, 1e-4)	(50, 1e-4)	(50, 1e-4)	(100, 1e-8)	(50, 1e-8)	(50, 1e-8)	(50, 1e-8)	(50, 1e-8)
corn	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(10, 1)	(10, 0.01)	(10, 0.01)	(10, 0.1)	(10, 0.1)	(2, 1e-8)	(10, 1e-8)

mined parameters. In classification for multiclass texts, we apply the one-versus-the-rest method to give a category label to each test document. Labels are given when estimation values of each document label exceed each of the predefined thresholds. We obtain the value of the precision recall break even point (PRBEP) and the average of PRBEP in each category. PRBEP is the point at which the precision corresponds to the recall, and it is used as an index for measuring information retrieval performance. Each PRBEP depends on each predefined threshold. So, we find thresholds at the best PRBEPs and let them predefined thresholds.

4. Results

Table 1 shows a pair of the optimal parameters (k, λ) for each category corresponding to the value of α a range from 0.0 to 1.0 every 0.1. **Figs. 1–10** show the experimental results in using these parameters in each category. The horizontal axis indicates the value of α and the vertical axis that value of PRBEP. Each figure shows the average of PRBEP in each category after 10 trials for each α. **Fig. 11** shows the macro average of PRBEP after 10 trials in the overall category corresponding to each α. **Fig. 12** shows how the relative ratio of PRBEP changes corresponding to the number of test data in each category, i.e., *earn*, *acq*, and *money-fx* at $\alpha = 0$, 0.2, and 1.0, and **Fig. 13** also shows the correlation between PRBEP and the number of test data in each category, i.e., *money-fx*, *grain*, *crude*, *trade*, *interest*, *ship*, *wheat*, and *corn* at $\alpha = 0$, 0.2, and 1.0. The horizontal axis indicates the number of test data in each category and the vertical axis the value of PRBEP. The dotted line shows the results for each category at $\alpha = 0$, the dashed-dotted line at $\alpha = 1$, and the solid line at $\alpha = 0.2$.

In all figures, the cases at $\alpha = 0$ mean that only surface information is used in the graph for GBSSL, and the cases at $\alpha = 1$ mean that only latent information is used. The results at $\alpha \neq 0$ or 1 indicate the cases of latent information and surface information mixed at a ratio of $\alpha : (1 - \alpha)$ $(0 < \alpha < 1)$. We consider the result at $\alpha = 0$ to be the baseline.

We start by explaining **Figs. 1–10**. The PRBEPs at $\alpha \in (0, 1]$ $(\alpha \neq 0)$ are greater than that at $\alpha = 0$ in **Figs. 1–3** and **Figs. 6–8**. Some PRBEPs at $\alpha \in (0, 1]$ $(\alpha \neq 0)$ are greater than at $\alpha = 0$, and other PRBEPs are less in **Figs. 4**, **5**, **9**, and **10**. The PRBEP in **Fig. 10** tends to decline as α increases.

Second, the macro average at $\alpha = 0$ is 45.2 in **Fig. 11**. The maximum value of the macro average is 51.0 at $\alpha = 0.2$ and the minimum is 44.5 at $\alpha = 1$. The macro average increases monotonically from 45.2 to 51.0 as α increases from 0.0 to 0.2. When α exceeds 0.2, the macro average decreases monotonically from 51.0 to 44.5.

We lastly find the direct proportion between the PRBEP value and the number of test data of the categories: *earn*, *acq*, and *money-fx* in **Fig. 12**. We do not find any significant correlation between PRBEP and the number of test data for the categories: *money-fx*, *grain*, *crude*, *trade*, *interest*, *ship*, *wheat*, and *corn* in **Fig. 13**.

5. Discussions

Looking at **Figs. 1–10**, each optimal α at which PRBEP is the maximum is different and not uniform in the respective categories, so we cannot simply tell a specific ratio for balancing both types of information – i.e., surface information and latent information – that gives the best accuracy.

From a total viewpoint, however, we see a definite trend. The upward trend for PREBP appears in more than half of the categories in **Figs. 1–10**. The macro average for the overall category is shown in **Fig. 11**. Regarding the macro average at $\alpha = 0$ as a baseline, the maximum macro average at $\alpha = 0.2$ exceeds at $\alpha = 0$ by 5.8% and is still greater than that at $\alpha = 1$ by 6.5%. Differences between the values at $\alpha = [0.1, 0.5]$ and that at $\alpha = 0$ are also significant in a t-test at the 0.05 significant level. We can thus say that using both types of information gives higher accuracy than using only surface information, and, of course, than using only latent information. If an appropriate α is decided, we will get better accuracy in GBSSL.

Regarding the correlations between PRBEP and the number of test data in a category, the more test data for each category we have, the higher the accuracy of GBSSL we probably get. The Reuters-21578 corpus does not have

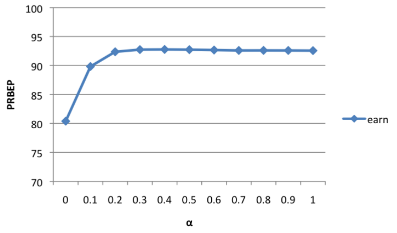


Fig. 1. *earn*.

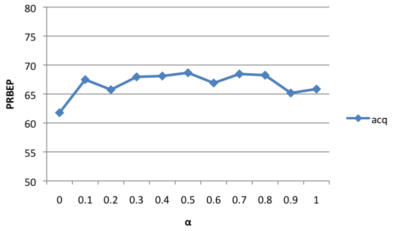


Fig. 2. *acq*.

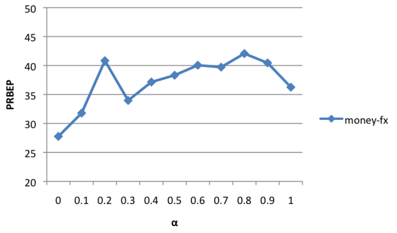


Fig. 3. *money-fx*.

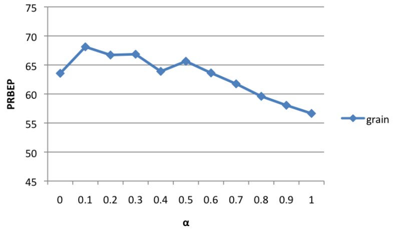


Fig. 4. *grain*.

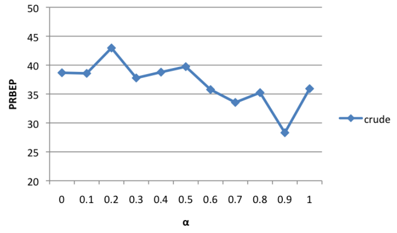


Fig. 5. *crude*.

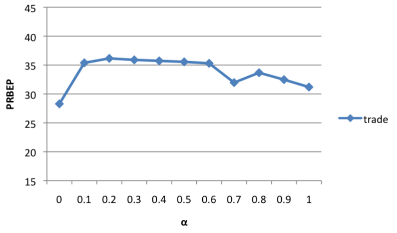


Fig. 6. *trade*.

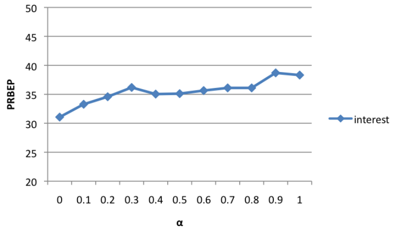


Fig. 7. *interest*.

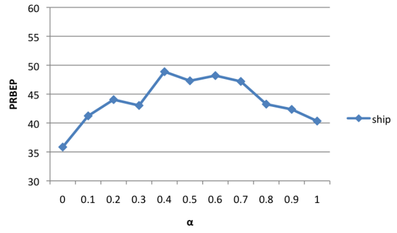


Fig. 8. *ship*.

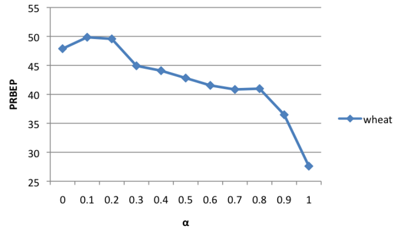


Fig. 9. *wheat*.

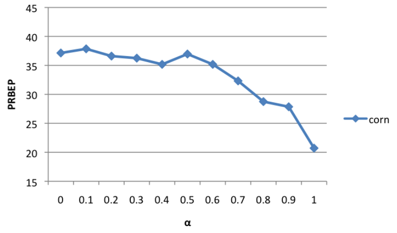


Fig. 10. *corn*.

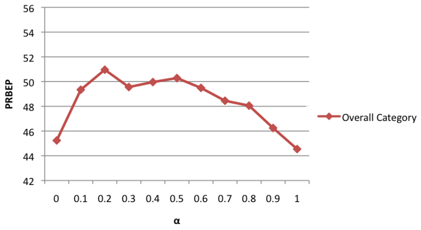


Fig. 11. Macro average.

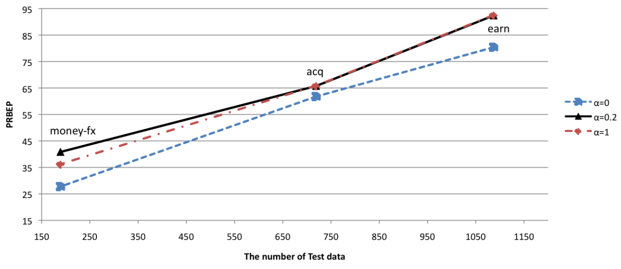


Fig. 12. The correlation between the PRBEPs and the number of test data of categories in the range of test data [150,1150].

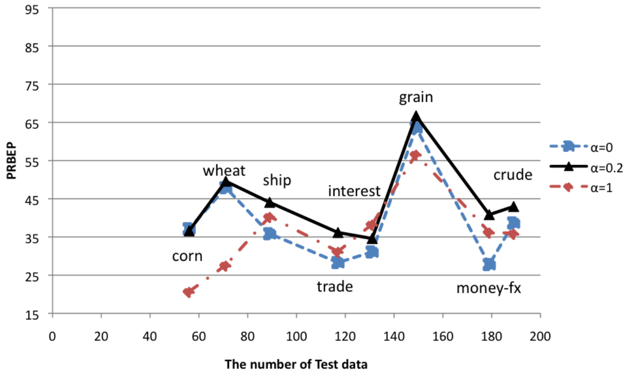


Fig. 13. The correlation between the PRBEPs and the number of test data of categories in the range of test data [0,200].

enough test data in each of the 10 categories, so it will be necessary to confirm the above findings using another corpus data set. We also can say that we need a large amount of data to extract better latent information from them. We confirmed this for the categories: *ship*, *wheat* and *corn*, where the number of test data is less than 100. PRBEPs of the categories at $\alpha = 1$ are worse than those at $\alpha = 0$ or 0.2. The shortage of test data prevented us from estimating latent topics well in the three categories: *ship*, *wheat*, and *corn*.

6. Conclusions

To improve the accuracy of GBSSL, more studies have been done from the viewpoint of the selection of training data and the graph construction. In the latter graph construction, some types of sparse graph constructions have been proposed, but the sparse graph construction methods simply decide which edge is cut off from a graph.

We have proposed new graph construction for GBSSL from the viewpoint of measuring similarity between nodes, rather than of selecting important edges. We have used a topic model (LDA) and estimated latent information for documents. We then employed both latent information and surface information, which is usually used in text classification task, in the proportion of $\alpha : (1-\alpha)$ $(0 \leq \alpha \leq 1)$. The proposal has been applied to a multi-class text classification task using GBSSL. We have confirmed that our proposal for constructing the graph for GBSSL based on both surface information and latent information gives higher accuracy than that based on either surface information or latent information alone.

References:

[1] H. Scudder, "Probability of error of some adaptive pattern-recognition machines," IEEE Trans. on Information Theory, Vol.11, No.3, pp. 363-371, 1965.

[2] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," In Proc. of the eleventh Annual Conf. on Computational Learning Theory, pp. 92-100, 1998.

[3] T. Joachims, "Transductive Inference for Text Classification using Support Vector Machines," In Proc. of the Sixteenth Int. Conf. on Machine Learning, pp. 200-209, 1999.

[4] A. Subramanya and J. Bilmes, "Soft-Supervised Learning for Text Classification," In Proc. of the 2008 Conf. on Empirical Methods in Natural Language Processing, pp. 1090-1099, 2008.

[5] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with Local and Global Consistency," Advances in Neural Information Processing Systems, Vol.16, pp. 321-328, 2004.

[6] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions," In Proc. of the Twentieth Int. Conf. on Machine Learning, pp. 912-919, 2003.

[7] X. Zhu, "Semi-supervised Learning with Graphs," Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2005.

[8] Q. Gu and J. Han, "Towards Active Learning on Graphs: An Error Bound Minimization Approach," IEEE Int. Conf. on Data Mining, pp. 882-887, 2012.

[9] T. Jebara, J. Wang, and S.-F. Chang, "Graph construction and *b*-matching for semi-supervised learning," In Proc. of the 26th Annual Int. Conf. on Machine Learning, pp. 441-448, 2009.

[10] K. Ozaki, M. Shimbo, M. Komachi, and Y. Matsumoto, "Using the mutual *k*-nearest neighbor graphs for semi-supervised classification of natural language Data," In Proc. of the Fifteenth Conf. on Computational Natural Language Learning, pp. 154-162, 2011.

[11] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, 1983.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," Machine Learning Research, Vol.3, pp. 993-1022, 2003.

[13] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization," In Proc. of HLT-NAACL 2006 Workshop on TextGraphs: Graph-based Algorithms for Natural Language Processing, pp. 45-52, 2006.

Name:
Akiko Eriguchi

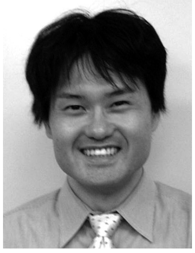
Affiliation:
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Address:
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Brief Biographical History:
2013 Received B.Sc. degree from Ochanomizu University
2013- Master Student, Ochanomizu University

Main Works:
- "High-quality Training Data Selection using Latent Topics for Graph-based Semi-supervised Learning," Sofia, Bulgaria, August 4-7, 2013.

Membership in Academic Societies:
- The Association for Computational Linguistics (ACL)
- The Japanese Society for Artificial Intelligence (JSAI)
- Information Processing Society of Japan (IPSJ)



Name:
Ichiro Kobayashi

Affiliation:
Professor, Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Address:
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

Brief Biographical History:
1995- Assistant Professor, Hosei University
1996-2003 Associate Professor, Hosei University
2003-2010 Associate Professor, Ochanomizu University
2011- Professor, Ochanomizu University.

Main Works:
- "Everyday Language Computing Project Overview," J. of Advanced Computational Intelligence and Intelligent Informatics, Vol.10, No.6, pp. 773-781, 2006.

Membership in Academic Societies:
- The Association for Computational Linguistics (ACL)
- Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)
- The Japanese Society for Artificial Intelligence (JSAI)
- Japan Association for Systemic Functional Linguistics (JASFL)
- The Association for Natural Language Processing (NLP)