Automatic Microblog Summarization Based on Unsupervised Key-Bigram Extraction

Yufang Wu*, Heng Zhang, Bo Xu, Hongwei Hao, Chenglin Liu
Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Beijing, P.R. China.

* Corresponding author. Tel.: +86-10-82544483; email: yufang.wu@ia.ac.cn
Manuscript submitted September 19, 2014; accepted May 10, 2015
doi: 10.17706/ijcce.2015.4.5.363-370

Abstract: Microblog summarization can save large amount of time for users in browsing. In this paper, we propose an automatic microblog summarization method based on unsupervised key-bigram extraction. Firstly, we extract a key-bigram set to discover the key aspects of posts by three unsupervised key-bigram extractor based on Hybrid TF-IDF, TextRank and Latent Dirichlet Allocation (LDA). Secondly, we rank sentences by overlap similarity and mutual information strategies based on the key-bigram set. Top ranked salient sentences with redundancy removal are selected as summaries. Compared with some other text content based summarizers, the proposed method was shown to perform superiorly in experiments on SinaWeibo and Twitter datasets.

Key words: Automatic summarization, key-bigram extraction, microblog, sentence extraction.

1. Introduction

Microblog platforms such as Twitter and SinaWeibo make it convenient for us to get real-time information. However, they also lead to heavy information overload. Automatic microblog summarization can help us master the core content fleetly and roundly so that a large amount of time can be saved.

The goal of this paper is to automatically extract summary for a set of posts that are related to the same hot topic on microblog. It seems a bit like the multi-document summarization. However, it is more intractable to summarize microblog posts since they suffer from severe sparsity, bad normalization and heavy noise, while traditional documents are usually with nice writing style.

To overcome the above difficulties, we propose an efficient microblog summarization method based on unsupervised key-bigram extraction. Unlike most existing methods [1]-[4], which rank sentences directly or based on words, we fulfill summarization in two steps: 1) Extract a key-bigram set (KBS) to discover the key aspects of posts; 2) Rank sentences based on the KBS and extract the top ranked sentences as summary, which is supervised by a similarity threshold to keep the summary from redundancy. Two strategies, overlap similarity and mutual information, are proposed to rank sentences based on the idea that the sentence with appropriate length and containing more key-bigrams should get higher rank.

Our main contributions can be summarized as follows: 1) For all we know, it is the first work to exploit unsupervised key-bigram extraction to summarize microblog. Taking bigram instead of word as language concept makes the KBS more powerful to capture the key aspects of posts; 2) We propose two efficient sentence extraction strategies based on the KBS. Our method outperforms some existing text content based summarizers both on SinaWeibo and Twitter datasets.

The remainder of this paper is organized as follows: Section 2 briefly reviews the related work. Section 3

describes our proposed method in details. Section 4 presents experimental results on two datasets. Finally, conclusions are made in the last section.

2. Related Work

There have been a few recent efforts in summarizing microblog, while traditional text summarization methods show bad performances on microblog. Existing microblog summarization methods usually take into account text contents, social attributes and user influences [5], [6]. However, here we only review some related text content based summarization methods. Sharifi *et al*. [2] proposed Phrase Reinforcement (PR) to find the most commonly occurring phrases to be included in a summary. Later, they proposed a simpler statistical Hybrid TF-IDF algorithm to weight sentences, which even outperformed PR [3]. Inouye and Kalita [4] developed Hybrid TF-IDF to generate multiple post summary by introducing similarity threshold. Compared with several well-known traditional text summarizers, the results showed that Hybrid TF-IDF summarizer with similarity threshold, performed better than more complex traditional summarizers. However, previous works aim to use the whole Bag-of-Words (BoW) for scoring sentences directly, which may introduce much noise, especially for informal and conversational microblog posts.

Keyword extraction has a close connection to a number of text mining tasks. We focus on unsupervised keyword extraction methods in this paper. TF-IDF [7] is the most widely used method due to its simplicity and efficiency. TextRank [1], a kind of graph-based ranking methods, ranks words according to their centrality. Recently, more works [8], [9] focus on discovering latent semantic relationships between words to reduce vocabulary gap by LDA [10]. Ernesto *et al*. [11] exploited key phrase extraction to LAKE system at DUC-2005. Li *et al*. [12] summarized traditional multi-documents based on maximizing bigram weights by integer linear programing (ILP). However, no similar work has been applied to the heavy noisy and semantic sparse microblog. Hence, it is significant to investigate whether microblog summarization based on unsupervised key-bigram extraction can work efficiently.

3. Microblog Summarization Based on Key-Bigram Extraction

3.1. Framework

Given a set of microblog posts related to the same topic, we extract salient sentences with redundancy removal to form a summary with appropriate length. Our method mainly consists of three parts, preprocessing & bigram formalization, key-bigram extraction, sentence extraction. Fig. 1 shows the whole framework: firstly, we generate bigrams based on the preprocessed posts, which are denoised and sentence-formalized; secondly, we extract bigrams that are highly related to the microblog topic as key-bigrams using three unsupervised techniques, Hybrid TF-IDF (HTI), TextRank (TR) and LDA; thirdly, we extract salient sentences by two strategies, Overlap Similarity (OS) and Mutual Information (MI). Combining a key-bigram extraction technique with a sentence extraction strategy, we obtain six instantiated summarizers, namely HTI-OS, TR-OS, LDA-OS, HTI-MI, TR-MI and LDA-MI.

3.2. Preprocessing & Bigram Formalization

We remove all topic hashtags, embedded URLs, symbol emotions, repost characters and user names to clean the posts. Then we split the posts into sentences, and consequently split sentences into unigrams. Then sentences are formalized by a bag of bigrams. Each bigram is generated by combining two adjacent unigram in each sentence. Bigram is a kind of language concepts like word and phrase. But it's more informative than word, so that it more powerful to convey the key aspects, and more convenient than phrase, since phrase extraction may need extern lexicon and complicated syntactic parsing.

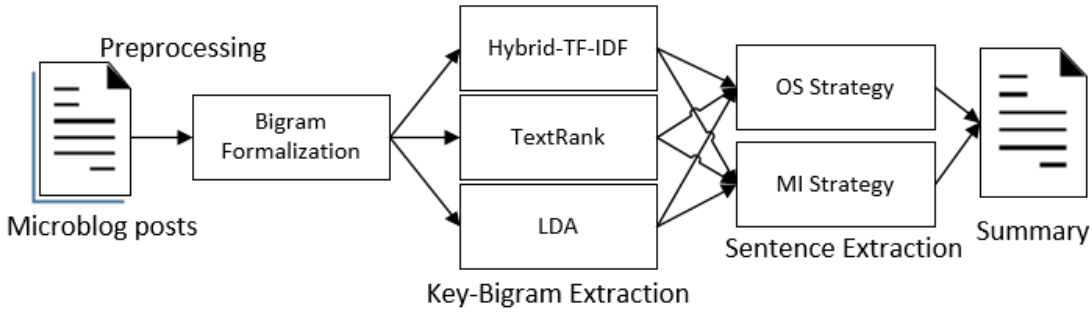


Fig. 1. Framework of key-bigram based automatic microblog summarization.

3.3. Key-Bigram Extraction

Hybrid TF-IDF Extractor. Let $tf(b_i)$ be the frequency of bigram b_i occurring in the sentence set, and $idf(b_i)$ be the proportion of the size of the sentence set to the number of sentences that b_i occurs. Hybrid TF-IDF [3] can be formally defined as follows:

$$S\_TF\_IDF(b_i) = tf(b_i) \times log_2\big(idf(b_i)\big). \quad (1)$$

Bigrams are ranked by their scores of (1). Then the top-*N* are extracted as key-bigrams of the posts.

TextRank Extractor. We construct a directed weighted graph $G(V, E)$ by taking each bigram as a vertex, and the co-occurring times of two ordered bigrams within a fixed length window (we set to 10) as the weight of edge, where *V* is the set of vertexes and *E* is the set of edges. Let $\mathrm{In}(v_i)$ be the set of vertexes that point to vertex v_i, and $\mathrm{Out}\,(v_j)$ be the set of vertexes that pointed by vertex v_j. Let w_{ji} be the weight of the edge from v_j to v_i. TextRank [1] computes the score of each vertex as follows:

$$S\_TR(v_i) = (1-d) + d \times \sum_{v_j \in In(v_i)} \Big(w_{ji} \times S\_TR\big(v_j\big) / \sum_{v_k \in Out(v_j)} w_{jk} \Big), \quad (2)$$

where d is the damping factor, whose value is usually set to 0.85. Recursively, we get the scores of each bigram and the influential top-*N* are selected as key-bigrams.

LDA Extractor. We extract key bigrams based on the topic-word (it is topic-bigram in our task) distribution matrix $\boldsymbol{\Phi} = (\boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_K)^T \in \mathbb{R}^{K \times V}$ of LDA [10], in which each column is the distribution of bigram b_v over the *K* topics, and each element $\hat{\varphi}_{k,v}$ is the probability of b_v belonging to topic z_k that measures the importance of b_v in z_k to some degree. We sum up $\hat{\varphi}_{k,v}$ by column as the global score of b_v, which is formally defined as the following equation:

$$S\_LDA(b_v) = \sum_{k=1}^{K} \hat{\varphi}_{k,v}. \quad (3)$$

We rank bigrams in descending order based on their global scores, and select the top-*N* as key-bigrams.

3.4. Sentence Extraction

We propose two strategies to rank sentences based on the straightforward idea that the sentences with appropriate length and containing more key-bigrams, are usually more salient.

Overlap Similarity (OS) Strategy. OS is a recall-liked score, which counts the overlap bigrams between the sentence and the KBS, and is divided by the size of KBS. In order to penalize the too long (or too short) sentences, we normalize the score by a factor that is the greater one between the average length of the

sentence set and the length of the candidate sentence. Specifically, the score of a sentence S_j computed by OS can be formally defined as follows:

$$S\_OS(S_j) = |\{b_i|b_i \epsilon S \& b_i \epsilon KBS\}|/(\max(AveLen, |S|) \cdot |KBS|), \quad (4)$$

where b_i is the co-occurring bigram, $|S_j|$ is the length of sentence, and $|KBS|$ is the size of KBS.

Mutual Information (MI) Strategy. MI measures the relevance between two variables. Therefore, we can measure the extent that how a sentence contains the KBS by MI. The higher MI score means the higher coverage degree of the sentence. Specifically, the score of a sentence S_j computed by MI can be formally defined as follows:

$$S\_MI(S_j) = \sum_{i=1}^{|KBS|} \log(p(b_i, S_j)/p(b_i)p(S_j))/\max(AveLen, |S_j|), \quad (5)$$

where $p(b_i, S_j)$ is the frequency of bigram b_i occurring in sentence S_j, $p(b_i)$ is the frequency of bigram b_i occurring in the sentence set, and $p(S_j)$ is the proportion of the length of sentence S_j to the length of the whole sentence set. The score is explicitly normalized by the same normalization factor defined in (4).

Top ranked sentences may be quite similar to each other. Therefore, we introduce a similarity threshold t when extracting sentences starting from the top ranked one.The current candidate sentence is chosen only when the similarities of it and the selected sentences all satisfy (6), or we discard the current one and move to the next ranked one until we extract M sentences.

$$Sim(S_i, S_j) = |\{b_i|b_i \epsilon S_i \& b_i \epsilon S_j\}|/(\log|S_i| + \log|S_j|) \leq t. \quad (6)$$

4. Experiments

4.1. Experimental Setup

We perform experiments on two datasets. One is the SinaWeibo posts, which covers50 topics from the SinaWeibo hot topic lists (http://huati.weibo.com/). Each topic contains about 2000 posts. Two volunteers are invited to extract 10 sentences from the posts of each topic to form the manual summary. The other is the Twitter posts from Inouye *et al.* [4], which consists of 25 topics. Each topic contains 100 posts and two manual summaries. Each manual summary consisted of four sentences.

As for performance evaluation, ROUGE-*N* [13] is one of the most popular automatic evaluation metrics. The Recall, Precision and F-measure of ROUGE-*N* can be computed as below:

$$\text{Recall} = \sum_{S \in MS} \sum_{n\_gram \in S} \text{Match}(n\_gram) / \sum_{S \in MS} \sum_{n\_gram \in S} \text{Count}(n\_gram), \quad (7)$$

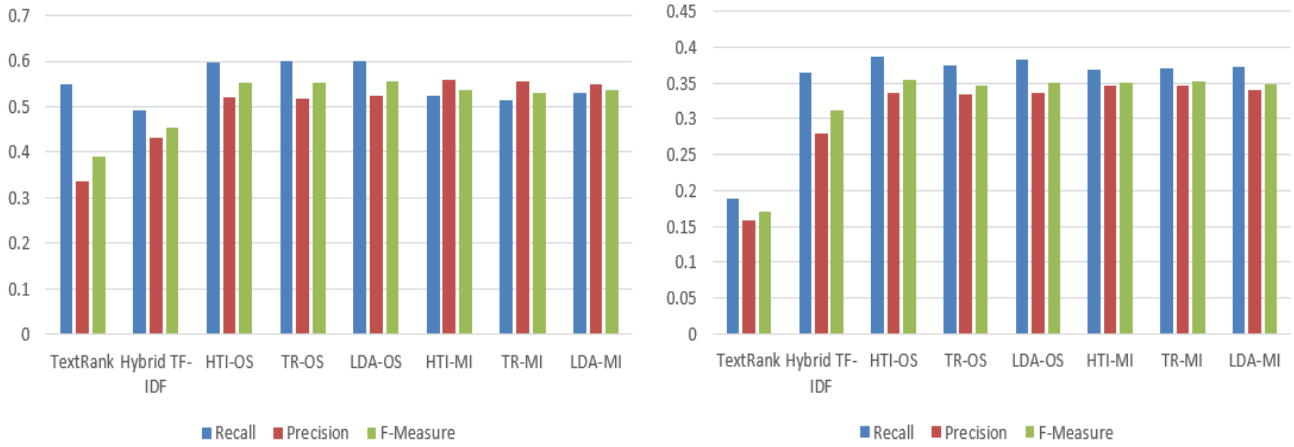
$$\text{Precision} = \sum_{S \in MS} \sum_{n\_gram \in S} \text{Match}(n\_gram) / \left(|MS| \times \sum_{n\_gram \in AS} \text{Count}(n\_gram)\right), \quad (8)$$

$$F - Measure = 2 \times \text{Recall} \times \text{Precision}/(\text{Recall} + \text{Precision}), \quad (9)$$

where MS is the manual summaries, AS is the automated summary, $\text{Match}(n\_gram)$ is the number of co-occurring n-grams between the manual and automated summaries, $\text{Count}(n\_gram)$ is the number of n-grams in the manual summaries, and $|MS|$ is the number of manual summaries. To keep the comparability with Inouye's work, we take ROUGE-1 as the metric, in which n-grams go back to unigrams.

4.2. Results and Discussions

We compare our method with two baselines, the Hybrid TF-IDF with similarity threshold summarizer [4] and the TextRank summarizer [1]. The ROUGE-1 performance on two datasets is shown in Fig. 2. Generally, our six key-bigram-based summarizers outperform baselines obviously on both datasets, especially the former one, which gains about 10% improvements of F-measure. Specifically, 1) TextRank summarizer shows bad performance on microblog summarization, which also proves that it is unwise to directly apply traditional summarizer to summarize microblog. However, TR-OS and TR-MI summarizers, which use TextRank to extract key-bigrams instead of sentences, show obvious improvementson both datasets, especially the great enhancements of precision. 2) Hybrid TF-IDF summarizer, scoring sentences with the whole BoW, is much better than TextRank but still with low precision. However, our six key-bigram-based summarizers, which only use less than 200 bigrams, show much better results. This is mainly because the KBS filters out the noisy and trivial words, so that our method generates more precise summaries on noisy microblog. 3) Three key-bigram extractors, namely HTI, TR and LDA, show similar F-measure scores under the same sentence extraction strategy, among which LDA and HTI slightly outperform others on SinaWeibo and Twitter datasets respectively. 4) The OS strategy is generally superior to MI strategy according to the F-measure scores. While the former one gets higher recall scores because it is a recall-designed strategy, and the latter one shows better precision values because it penalizes long sentences more severely.



(a) (b)

Fig. 2. ROUGE-1 comparison of baselines and our methodon (a) SinaWeibo and (b) Twitter datasets.

For all the experiments, we set similarity threshold t in (6) to 0.77 for Twitter dataset as Inouye's work; and set it to 0.5 for SinaWeibo dataset considering that a sentence carrying less than 50% new information does not deserve to be included in the summary especially on a big candidate set. Many existing keyword extraction researches propose to filter words by part-of-speech (POS) and remove stop words in preprocessing [11]. However, we find that SinaWeibo dataset shows the best results for all six summarizers with stop words removed and all POS maintained. And Twitter dataset performs best with all POS and stop words maintained. This may because of the small scale of Twitter dataset. For LDA-OS and LDA-MI summarizers, we compute the average ROUGE-1 results of 20 times to weaken the effects of random seeding. Besides, our exhaustive experimental results lead us to set the number of topics of LDA to 10 on SinaWeibo dataset and set it to 5 on Twitter dataset. Interestingly, the two values are much close to the number of sentences of manual summary on each dataset. It also means it is appropriate to summarize each hot topic with 10 subtopics on SinaWeibo dataset and 4 subtopics on Twitter dataset.

In order to measure the contribution of bigram, we also compared its performance with key-unigrams

(namely keywords). The results on two datasets are shown in Table 1 and Table 2. As we can see, two datasets gain more than 3% improvement by key-bigrams. Especially on Twitter dataset, the contribution of bigram is rather significant, since unigram-based summarizers only show 0.3% improvement than the result of Hybrid TFIDF in Fig. 2(b). This can be explained by that bigrams are more efficient to discover useful information from the more noisy Twitter posts since we keep all words during preprocessing.

Table 1. ROUGE-1 Comparison of Bigram and Unigram on SinaWeibo Dataset

Summarizers	Unigram			Bigram		
	Recall	Precision	F-Measure	Recall	Precision	F-Measure
HTI-OS	0.5481	0.5077	**0.5229**	0.5975	0.5198	0.5512
TR-OS	0.5337	0.5025	0.5137	0.6007	0.5183	0.5515
LDA-OS	0.5460	0.5057	0.5205	0.5997	0.5246	**0.5550**
HTI-MI	0.4937	0.5221	0.5042	0.5237	0.5595	0.5371
TR-MI	0.4913	0.5151	0.4997	0.5133	0.5539	0.5293
LDA-MI	0.4948	0.5163	0.5021	0.5310	0.5506	0.5371

Table 2. ROUGE-1 Comparison of Bigram and Unigram on Twitter Dataset

Summarizers	Unigram			Bigram		
	Recall	Precision	F-Measure	Recall	Precision	F-Measure
HTI-OS	0.3469	0.2807	0.3042	0.3875	0.3366	**0.3534**
TR-OS	0.3547	0.2819	0.3070	0.3751	0.3336	0.3471
LDA-OS	0.3547	0.2905	0.3135	0.3825	0.3354	0.3501
HTI-MI	0.3335	0.2896	0.3025	0.3691	0.3466	0.3498
TR-MI	0.3498	0.3010	**0.3162**	0.3709	0.3467	0.3530
LDA-MI	0.3431	0.3028	0.3146	0.3714	0.3402	0.3476

5. Conclusions

This paper presents an automatic microblog summarization method based on key-bigram extraction, which summarizes a set of microblog posts from a specific topic in two steps: key-bigram set (KBS) extraction and sentence ranking based on the KBS. We implemented three unsupervised key-bigram extraction techniques based on Hybrid TF-IDF (HTI), TextRank and LDA, and two sentence ranking measures overlap similarity (OS) and mutual information (MI). Compared with the Hybrid TF-IDF summarizer that using BoW for scoring sentences and the TextRank summarizer that uses direct sentence ranking for summarizing traditional single document, our proposed method yielded superior performance on SinaWeibo and Twitter datasets. Specifically, our instantiated method HTI-OS performs best among the six summarizers when we synthetically consider the ROUGE-1 F-measure and the simplicity.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China Grants 61203281 and 61303172. Besides, we would like to thank Inouye *et al*. for their Twitter dataset.

References

[1] Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. *EMNLP* (pp. 404–411).

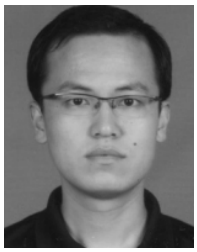
[2] Sharifi, B., Hutton, M. A., & Kalita, J. K. (2010). Summarizing microblogs automatically. *Proceedings of HLT/NAACL: Vol. 10* (pp. 685-688).

[3] Sharifi, B., Hutton, M. A., & Kalita, J. K. (2010). Experiments in microblog summarization. *Proceedings of IEEE 2nd Int. Conf. Social Computing* (pp. 49-56).

[4] Inouye, D., & Kalita, J. K. (2011). Comparing twitter summarization algorithms for multiple post summaries. *Proceedings of IEEE 3nd Int. Conf. Social Computing* (pp. 298-306).

[5] Duan, Y. J., *et al.* (2012). Twitter topic summarization by ranking tweets using social influence and content quality. *Proceedings of COLING* (pp. 763–780). Mumbai.
[6] Huang, H. Z., *et al.* (2012). Tweet ranking based on heterogeneous networks. *Proceedings of COLING* (pp. 1239–1256). Mumbai.
[7] Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. *Proceedings of EMNLP* (pp. 216-223).
[8] Liu, Z. Y., Huang, W. Y., Zheng, Y. B., & Sun, M. S. (2010). Automatic keyphrase extraction via topic decomposition. *Proceedings of EMNLP* (pp. 366-376).
[9] Zhao, W. X., Jiang, J., He, J., Song, Y., Palakorn, A., Lim, E. P., & Li, X. M. (2011). Topical keyphrase extraction from twitter. *Proceedings of HLT/NAACL: Vol. 1* (pp. 379-388).
[10] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.
[11] D'Avanzo, E., & Magnini, B. (2005). A Keyphrase-based approach to summarization: The LAKE system at DUC-2005. *Proceedings of DUC*.
[12] Li, C., Qian, X., & Liu, Y. (2013). Using supervised bigram-based ILP for extractive summarization. *Proceedings of ACL* (pp. 1004-1013).
[13] Lin, C. Y., & Eduard, H. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of HLT/NAACL: Vol. 1* (pp. 71-78).

Yufang Wu received the B.S. degree in automation from Central South University, Changsha, China, in 2012. She is currently pursuing her master's degree in pattern recognition and intelligent systems at the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. Her research interests include natural language processing especially text analysis, machine learning and data mining.



Heng Zhang received the B.S. degree in electronic and information engineering from University of Science and Technology of China, Hefei, China, in 2007, and got the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2013. Currently, he is an assistant professor at the CASIA. His research interests include short text classification and analysis, speech recognition, handwriting recognition, document analysis and information retrieval.

Bo Xu is currently an associate professor at Institute of Automation, the Chinese Academy of Sciences from July 2011. He received M.E. degree in Automation in 2006 from Xi'an Jiao Tong University and the Ph.D. degree in pattern recognition and artificial intelligence from the Institute of Automation, Chinese Academy of Sciences, in 2011. His research interests include pattern recognition, image processing, machine learning and especially the applications to character recognition.

Hongwei Hao is currently a professor in the Institute of Automation, Chinese Academy of Sciences (CASIA), China. He received the Ph.D. degree in Pattern Recognition and Intelligent Systems from CASIA in 1997. From 1999 to 2011, he worked as an associate professor and then a professor at the University of Science and Technology Beijing, China. From 2002 to 2003, he was a visiting researcher in the Central Research Laboratory, Hitachi Ltd., Tokyo, Japan. His research interests cover large-scale semantic computing, large-scale machine learning theory, and intelligent massive information processing.

Chenglin Liu is a professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of Chinese Academy of Sciences (CASIA), Beijing, China, and is now the deputy director of the laboratory. He received the B.S. degree in electronic engineering from Wuhan University, Wuhan, China, the M.E. degree in electronic engineering from Beijing Polytechnic University, Beijing, China, the Ph.D. degree in pattern recognition and intelligent control from CASIA, Beijing, China, in 1989, 1992 and 1995, respectively. He was a postdoctoral fellow at Korea Advanced Institute of Science and Technology (KAIST) and later at Tokyo University of Agriculture and Technology from March 1996 to March 1999. From 1999 to 2004, he was a research staff member and later a senior researcher at the Central Research Laboratory, Hitachi, Ltd., Tokyo, Japan. His research interests include pattern recognition, image processing, neural networks, machine learning, and especially the applications to character recognition and document analysis. He has published over 180 technical papers at prestigious international journals and conferences. He is on the editorial board of journals Pattern Recognition, Image and Vision Computing, and International Journal on Document Analysis and Recognition. He is a fellow of the IAPR, and a senior member of IEEE.