Active Learning With Sampling by Uncertainty and Density for Data Annotations

Jingbo Zhu, *Member, IEEE*, Huizhen Wang, Benjamin K. Tsou, *Member, IEEE*, and Matthew Ma, *Senior Member, IEEE*

***Abstract*—To solve the knowledge bottleneck problem, active learning has been widely used for its ability to automatically select the most informative unlabeled examples for human annotation. One of the key enabling techniques of active learning is uncertainty sampling, which uses one classifier to identify unlabeled examples with the least confidence. Uncertainty sampling often presents problems when outliers are selected. To solve the outlier problem, this paper presents two techniques, *sampling by uncertainty and density (SUD)* and *density-based re-ranking*. Both techniques prefer not only the most informative example in terms of uncertainty criterion, but also the most representative example in terms of density criterion. Experimental results of active learning for word sense disambiguation and text classification tasks using six real-world evaluation data sets demonstrate the effectiveness of the proposed methods.**

***Index Terms*—Active learning, density-based re-ranking, sampling by uncertainty and density, text classification, uncertainty sampling, word sense disambiguation (WSD).**

I. Introduction

In machine learning approaches to natural language processing (NLP), supervised learning methods generally set their parameters using labeled training data. However, creating a large labeled training corpus is expensive and time-consuming in some real-world applications, and is often a bottleneck to build a supervised classifier for a new application or domain. For example, building a large-scale sense-tagged training corpus for supervised word sense disambiguation (WSD) tasks is a crucial issue, because validations of sense definitions and sense-tagged data annotation must be done by human experts, such as reported in OntoNotes project (Hovy *et al.* [10]). Our study aims to minimize the amount of human labeling efforts required for a supervised classifier to achieve a satisfactory performance by using *active learning*.

Among the techniques used to solve the knowledge bottleneck problem, active learning is a widely used framework in which the learner has the ability to automatically select the most informative unlabeled examples for human annotation (Cohn *et al.* [6], Seung *et al.* [21]). The ability of the active learner can be referred to as *selective sampling*, of which two major schemes exist: *uncertainty sampling* and *committee-based sampling*. Uncertainty sampling (Lewis and Gale [13]) uses only one classifier to identify unlabeled examples on which the classifier is least confident. Committee-based sampling (Seung *et al.* [21], McCallum and Nigam [14]) generates a committee of classifiers (always more than two classifiers) and selects the next unlabeled example by the principle of maximal disagreement among these classifiers. In this paper, we are interested in uncertainty sampling, which in recent years has been widely studied in natural language processing applications such as word sense disambiguation (Chen *et al.* [5], Chan and Ng [4]), text classification (TC) (Lewis and Gale [13], Zhu *et al.* [29]), statistical syntactic parsing (Tang *et al.* [23]), and named entity recognition (NER) (Shen *et al.* [22]).

From experimental results (Roy and McCallum [18], Tang *et al.* [23]), in uncertainty sampling many selected unlabeled examples having high uncertainty, namely outliers, cannot provide much help to the learner. Uncertainty sampling often fails by selecting such outliers. Previous studies have attempted to solve this problem. Cohn *et al.* [7] and Roy and McCallum [18] proposed a method that directly optimizes expected future error on future test examples. However, in real-world applications, their methods are almost intractable due to the high computational cost for selecting the most informative example from a large unlabeled pool. Tang *et al.* [23] adopted a sampling scheme of "most uncertain per cluster" for NLP parsing, in which the learner selects the sentence with the highest uncertain score from each cluster, and uses the density to weigh the selected examples. In fact, the scheme of using the most uncertain example per cluster still cannot solve the outlier problem faced by uncertainty sampling. Shen *et al.* ([22]) proposed to select examples based on informativeness, diversity and density criteria. In their work, the density of an unlabeled example is evaluated within a cluster, and multiple criteria are linearly combined with different coefficients. However, as different values of the coefficients are associated with various applications, it is difficult to determine those coefficients automatically.

This paper aims to overcome the shortcomings in related work with respect to the outlier problem by presenting two approaches based on the assumption that an unlabeled example with high density degree is less likely to be an outlier (Zhu *et al.* [29], Zhu *et al.* [27]). First, we present a *sampling by uncertainty and density* (SUD) technique in which a new

Manuscript received July 28, 2008; revised September 05, 2009. First published September 29, 2009; current version published July 14, 2010. This work was supported in part by the National Science Foundation of China under Grant 60873091. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

J. Zhu and H. Wang are with the Key Laboratory of Medical Image Computing (Northeastern University), Ministry of Education, and the Natural Language Processing Lab, Northeastern University Shenyang 110004, China (e-mail: zhujingbo@mail.neu.edu.cn; wanghuizhen@mail.neu.edu.cn).

B. K. Tsou is with the Language Information Sciences Research Center, City University of Hong Kong, Hong Kong, China (e-mail: rlbtsou@cityu.edu.hk).

M. Ma is with Scientific Works, Princeton Junction, NJ 08550 USA (e-mail: mattma@ieee.org).

Digital Object Identifier 10.1109/TASL.2009.2033421

Procedure: Active Learning Process
Input: initial small training set L, and pool of unlabeled data set U
Use L to train the initial classifier C
Repeat
- Use the current classifier C to label all unlabeled examples in U
- Select the most informative unlabeled example[1] from the unlabeled pool U, and ask oracle H for labeling
- Augment L with this new example, and remove it from U
- Use L to retrain the current classifier C

Until the predefined stopping criterion SC is met.

[1] To decrease the number times the learner is retrained during active learning process, we also can use a batch mode sampling selection to label m most informative unlabeled examples at each learning cycle, same as done in (Shen *et al.,* 2004; Tang *et al.,* 2002).

Fig. 1. General active learning algorithm.

uncertainty measure called *density*entropy* is adopted. Second, we propose a new *density-based re-ranking* technique to select the most representative example from the N-best output of entropy-based uncertainty sampling. It is noteworthy that these proposed techniques are easy to implement, and can be easily applied to several different learners, such as maximum entropy (ME), naïve Bayes (NB) and support vector machines (SVMs).

The rest of the paper is structured as follows. Section II begins with a discussion of the general active learning process and three issues in developing a high-accuracy active learning technique. In Sections III and IV, we provide an in-depth analysis of two proposed techniques, respectively. Experimental results are given in Section V, followed by some further discussions in Section VI. We conclude in Section VII with future work.

II. General Active Learning Process

Formally, active learning is a two-stage process in which a small number of labeled samples and a large number of unlabeled examples are first collected in the initialization stage, and a closed-loop stage of query (i.e., selective sampling process) and retraining is adopted. The general active learning process can be summarized in Fig. 1.

As reported in previous studies (Tang *et al.* [23], Chen *et al.* [5], Zhu and Hovy [26]), active learning is a promising way to speed up data annotation while minimizing human labeling efforts. In practice, there are three crucial issues in developing a high accuracy active learning technique, as described in Fig. 1.

1) Construction of an Initial Training Data Set: Traditionally, the initial training data is generated at random, based on an assumption that random sampling is likely to build an initial training set with the same prior data distribution as that of the whole corpus. However, this situation seldom occurs in real-world applications, because random sampling technique cannot guarantee the selection of the most representative subset due to limited size of initial training set (e.g., 10). In our previous study (Zhu *et al.* [29]), a technique of *sampling by clustering* was applied to select some representative examples to form an initial training set. The building of a representative initial training data set is shown to be able to greatly improve active learning, particularly at its early stages.

2) Stopping Criterion: In principle, how to learn a stopping criterion is a problem of estimation of classifier effectiveness during active learning (Lewis and Gale [13]). Actually defining an appropriate stopping criterion for active learning is a tradeoff issue between labeling cost and effectiveness of the classifier. In our previous studies (Zhu and Hovy [26], Zhu *et al.* [28], Zhu *et al.* [30]), five simple and effective confidence-based stopping criteria including *max-confidence*, *min-error*, *overall-uncertainty*, *classification-change*, and *minimum-expected-error* were proposed to automatically determine when to stop the active learning process.

3) Selective Sampling Scheme: The third issue of active learning is how to select the most informative example for human annotation at each learning cycle, namely selective sampling scheme. In this paper, we are interested in uncertainty sampling (Lewis and Gale [13]) for pool-based active learning, in which an unlabeled example x with maximum uncertainty is selected for human annotation at each learning cycle. The maximum uncertainty implies that the current classifier (i.e., the learner) has the least confidence in its classification of this unlabeled example. In other words, an unlabeled example with maximum uncertainty is viewed as the most informative case in uncertainty sampling. However, in some scenarios, an unlabeled example with maximum uncertainty can be an outlier (Roy and McCallum [18], Tang *et al.* [23], Zhu *et al.* [29]), which is undesirable. In the following sections, we will discuss the outlier problem of uncertainty sampling, and present two effective techniques to solve this problem.

III. Sampling by Uncertainty and Density

A. Outlier Problem

In uncertainty sampling, the key issue of the selection of the most uncertain unlabeled example is how to measure the uncertainty of each unlabeled example x. The well-known *entropy* is a popular uncertainty measurement widely used in previous studies on active learning with uncertainty sampling (Tang *et al.* [23], Chen *et al.* [5], Zhu and Hovy [26]). The uncertainty measurement function based on the entropy can be expressed as follows:

$$H(x) = -\sum_{y \in Y} P(y|x) \log P(y|x) \tag{1}$$

where $P(y|x)$ is the *a posteriori* probability of the output class $y \in Y = \{y_1, y_2, \ldots, y_k\}$ given the input x. $H(\cdot)$ is the uncertainty measurement function based on the entropy estimation of the classifier's posterior distribution. In the remainder of this paper, uncertainty sampling based on entropy criterion is considered as the baseline method, and called traditional uncertainty sampling in the following sections.

As mentioned in previous studies on active learning (Roy and McCallum [18], Tang *et al.* [23], Zhu *et al.* [29]), uncertainty sampling often fails by selecting outliers. Schein and Unga [20] also reported that traditional uncertainty sampling sometimes shows unsatisfactory performance on data sets with noise. To

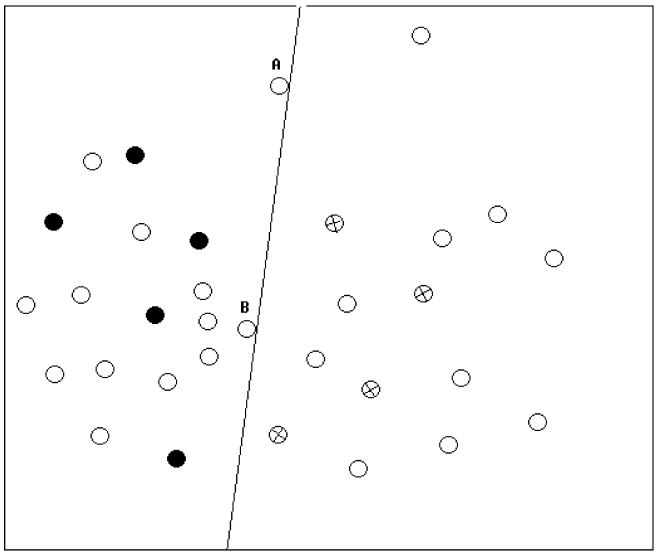

Fig. 2. Two unlabeled examples A and B with maximum uncertainty at the ith learning iteration. Solid circles and cross circles denote labeled samples with different labels in current training data. Blank circles denote unlabeled examples being queried. The solid line represents the corresponding decision boundary.

simplify our discussion, we give an example to explain the outlier problem, shown in Fig. 2.

The motivation behind uncertainty sampling is to find some unlabeled examples near decision boundaries, and use them to clarify the position of decision boundaries (Zhu *et al.* [29]). In uncertainty sampling, the current classifier considers unlabeled examples near decision boundaries as the most informative examples (i.e., the most uncertain cases). In other words, uncertainty sampling assumes that an unlabeled example with maximum uncertainty (i.e., very close to decision boundaries) has the highest chance to change the decision boundaries, and selecting such an unlabeled example for the next learning iteration can provide more help to the active learner.

In Fig. 2, two unlabeled examples, marked by A and B, have maximum uncertainty at the ith learning iteration and are to be considered. There are three unlabeled examples very close or similar to B but none for A. We think example B is more representative than example A, and A is likely to be an outlier. Adding B to the training set will thus help the learner more than adding A.

B. Density*Entropy Measure

Based on above analysis, we introduce the concept of *density* to determine whether an unlabeled example is highly representative. The density degree of an unlabeled example can be evaluated based on how many unlabeled examples are similar or close to it. High-density degree examples are highly representative. That is, an example with high density degree is less likely to be an outlier.

Based on the above analysis, we make an assumption that unlabeled examples near the decision boundary and very close to other examples and are more important than those that are isolated (i.e., likely to be outliers). Therefore, the ideal selective sampling criterion should combine the best of uncertainty and density criteria together. We prefer not only the most informative example in terms of uncertainty, but also the most representative example in terms of density for active learning.

In obtaining the density degree, the traditional cosine measure is adopted to estimate the similarity between two examples, that is

$$\cos(w_i, w_j) = \frac{w_i \bullet w_j}{||w_i|| \cdot ||w_j||} \tag{2}$$

where w_i and w_j are the feature vectors of examples i and j.

In practice, the unlabeled corpus is often very large (e.g., more than tens of thousands of unlabeled examples). It is therefore unreasonable to exhaustively calculate similarities between any example and all the others in order to obtain the density degree. Tang *et al.* [23] and Shen *et al.* [22] applied a technique of evaluating the density of an example within a cluster. In their work, the unlabeled data set is first clustered into a predefined number of clusters using K-Means clustering. The density of an example can be defined as the average similarity between itself and the other examples within the same cluster (Shen *et al.* [22]).

Our first intuition in estimating density is to apply a clustering-based technique. However, there are three reasons that may prohibit us from doing so. First, it is difficult to answer how many clusters are appropriate for clustering-based density evaluation in a specific active learning task. Second, as done in Tang *et al.* [23], the average size of the resulting clusters is still very large. In this situation, the resulting density values of unlabeled examples are close to each other. Finally, experimental results show that the size distribution of the resulting clusters is very skewed. It causes density estimation to be biased towards small clusters.

An alternative approach would be the use of a similarity score threshold, for which a similarity score above a predefined threshold indicates similar examples. However, it is still an open question how to predefine a proper similarity score threshold for different active learning tasks.

To avoid these critical issues, in this work we present a new method called *K-Nearest-Neighbor-based density* (KNN-density) measure in which the density of an example is quantified by the average similarity between this example and the other K most similar examples (i.e., K nearest neighbors).

Given a set of K most similar examples $S(x) = \{s_1, s_2, \ldots, s_K\}$ of the unlabeled example x, the average similarity $\mathrm{AS}(\cdot)$ between example x and its K most similar examples can be calculated by

$$\mathrm{AS}(x) = \frac{\sum_{s_i \in S(x)} \cos(x, s_i)}{K}. \tag{3}$$

As discussed above, we prefer to select unlabeled examples with maximum uncertainty and highest density for human annotation, which will be of high value to the learner. Along this line of thinking, by considering uncertainty and density simultaneously, we propose a new uncertainty sampling method, named *sampling by uncertainty and density* (SUD). To combine uncertainty and density, we adopt a common approach named *density*entropy* in which the density $\mathrm{AS}(x)$ of an unlabeled ex-

ample x is used as the coefficient of its uncertainty $H(x)$, defined by

$$DSH(x) = \mathrm{AS}(x) \times H(x). \quad (4)$$

The motivation of our SUD method is to use the density factor to adjust the uncertainty $H(x)$ of an unlabeled example x. A more uncertain example with high density should be assigned with a higher uncertainty value. In other words, the SUD method favors the example with high uncertainty and high density at each learning iteration.

C. Optimization

Because the unlabeled corpus can be very large, optimization has to be used in order to make the sampling by uncertainty and density more practical. Several approaches are possible.

First, a general approach would re-estimate the density of each unlabeled example at each learning iteration. When the scale of the unlabeled pool is very large, such re-estimation would be prohibitive. The simplest approximation is that those K most similar examples of each example x can be calculated in advance and fixed during the active learning process. Unfortunately, this method may cause negative effects on performance, because it is problematic to assume that the density of each unlabeled example cannot be changed during active learning.

Second, an alternative approximation is to first calculate the similarity between any two different unlabeled examples in the beginning of active learning. For each example x, we can rank the other examples by the similarity score. In this way, the density estimation for each unlabeled example x is very efficient based on top-K examples in its rank list. In the following comparison experiments, we adopt this approach to make SUD more efficient for implementation. Some details will be described in Section V-C.

Finally, the pool of unlabeled examples can be reduced by random sub-sampling (Roy and McCallum [18]). The density can be estimated using only a subset of the unlabeled pool, especially when the unlabeled pool is very large.

IV. Density-Based Re-Ranking

In uncertainty sampling, the entropy-based uncertainty measure defined in (1) is based on posterior probabilities produced by a probabilistic classifier. It would not work for active learning with non-probabilistic classifiers. For example, a SVM-based classifier does not yield probabilistic output. Instead, it produces a decision margin. This same limitation is applied to SUD scheme.

To overcome this limitation, this section describes how the ideas from selective sampling can be extended to re-ranking tasks. N-best re-ranking techniques have been successfully applied in NLP tasks, such as machine translation (Zhang *et al.* [25]), syntactic parsing (Collins and Koo [8]), and summarization (Hovy and Lin [9]). To our knowledge, there has been no attempt to use re-ranking techniques for selective sampling in active learning.

There are two processing steps in active learning with re-ranking technique. In the first step, a basic learner is used to generate the N-best candidates in terms of the uncertainty criterion at each learning cycle. After that, in contrast to the SUD method, the density measure is used to rank these N-best candidates. The top candidate from this list, in terms of the density criterion, is selected for human annotation. The density criterion can be implemented based on the average similarity function $\mathrm{AS}(\cdot)$ defined in (3). This method is called *density-based re-ranking*. The density-based re-ranking stage aims to select the unlabeled example with the highest density from these N candidates generated by the basic learner. In this case, we think the re-ranked example is less likely to be an outlier.

Procedure: Active Learning with re-ranking
Input: initial small training set *L*, and unlabeled data set *U*
Use L to train the initial classifier *C*
Repeat
1. Use the current classifier *C* to label all unlabeled examples in *U*
2. Using uncertainty sampling technique to select *N* most uncertain example from *U*
3. Select the unlabeled example with maximum density as defined in Equation (3) from the top-*N* candidates, and ask oracle *H* for labeling
4. Augment *L* with this new labeled example, and remove it from *U*
5. Use *L* to retrain the current classifier *C*

Until the predefined stopping criterion *SC* is met.

Fig. 3. Active learning with re-ranking technique.

This technique can be applied to active learning with probabilistic or non-probabilistic classifiers, because re-ranking technique is independent of the type of classifier used as the basic learner in active learning. For example, Shen *et al.* [22] applied an SVM-based classifier in active learning for named entity recognition task in which the uncertainty of each unlabeled example is estimated based on margin. The example with minimum margin is viewed as the most uncertain case. The N-best candidates selected by the SVM-based active learner can be used for re-ranking.

The procedure of active learning with re-ranking is summarized in Fig. 3

V. Evaluation

A. Datasets

To evaluate the effectiveness of various active learning methods including *uncertainty sampling*, *sampling by uncertainty and density (SUD)*, and *density-based re-ranking*, in this section we constructed some comparison experiments of active learning for two typical tasks: word sense disambiguation and text classification, using six publicly available real-world datasets as shown Table I.

1) Word Sense Disambiguation Task: Three publicly available real-world data sets are used in this task: *Interest*, *Line*, and *OntoNotes*. The *Interest* data set was developed by Bruce and Wiebe ([3]). It consists of 2369 sentences containing the noun "*interest*" with its correct sense manually labeled. The noun "*interest*" has six different senses in this data set. Interest data set has been previously used for WSD study (Ng and Lee [16]).

TABLE I
DESCRIPTIONS OF DATA SETS USED IN ACTIVE LEARNING EVALUATION, INCLUDING: THE NUMBER OF CLASSES AND CLASS DISTRIBUTION

Data sets		Class	Class distribution
OntoNotes	Rate	2	208/934/
	President	3	936/157/17/
	People	4	815/67/7/5/
	Part	4	102/454/16/75/
	Point	7	471/37/88/19/12/3/7/
	Director	2	637/35/
	Revenue	2	517/23/
	Bill	4	349/122/40/3/
	Future	3	23/82/409/
	Order	8	342/6/61/54/4/2/6/3/
Interest		6	500/1252/178/66/361/11/
Line		6	373/376/349/429/2218/404
Comp2a		2	983/1000/
Comp2b		2	999/1000/
WebKB		4	504/930/1641/1124/

In the *Line* data set, each instance of the word "*line*" has been tagged with one of six WordNet senses. The Line data set has been used in some previous studies on WSD (Leacock *et al.* [11]). The *OntoNotes* project (Hovy *et al.* [10]) uses the WSJ part of the Penn Treebank. The senses of noun words occurring in OntoNotes are linked to the Omega ontology (Philpot *et al.* [17]). In this experiment, we focus on the ten most frequent nouns[1] previously used for active learning (Zhu and Hovy [26]): *rate*, *president*, *people*, *part*, *point*, *director*, *revenue*, *bill*, *future*, and *order*.

2) Text Classification Task: Three publicly available data sets are used in this active learning comparison experiment: *Comp2a*, *Comp2b*, and *WebKB*. The Comp2a consists of the *comp.os.ms-windows.misc* and *comp.sys.ibm.pc.hardware* subsets of 20-NewsGroups. The Comp2b dataset consists of *comp. graphics* and *comp.windows.x* categories from 20-NewsGroups. The WebKB dataset was widely used in text classification research (McCallum and Nigam [15]). We used the four most populous categories: *student, faculty, course*, and *project*. These datasets have been previously used in active learning for text classification (Roy and McCallum [18]; Schein and Ungar [20]).

B. Experimental Settings

We utilize a maximum entropy (ME) model (Berger *et al.*, [2]) to design the basic classifier for WSD and TC tasks. The advantage of the ME model is its ability to freely incorporate features from diverse sources into a single, well-grounded statistical model. A publicly available ME toolkit[2] was used in our experiments. To build the ME-based classifier for WSD, three knowledge sources are used to capture contextual information: *unordered single words in topical context*, *POS of neighboring words with position information*, and *local collocations*, which are the same as the knowledge sources used in (Lee and Ng, [12]). The first type of features corresponds to a set of words occurring with the disambiguated word w in the same sentence. The second type of features denotes a set of part-of-speech (POS) tags of m words to left or right of w. Local collocation features are words adjacent to the word w to be disambiguated. In the design of the text classifier, the maximum entropy model is also utilized, and no feature selection technique is used.

In the following comparison experiments, the algorithm starts with an initial training set of ten labeled examples, and selects the most informative example at each learning iterationPT. A tenfold cross-validation was performed. All results reported are the average of ten trials in each active learning process.

Since the class distributions of most data sets shown in Table I are skewed, we adopted the F1 measure as the performance evaluation metric, because it combines the precision and recall estimated for each class separately. To globally compare different active learning methods, we adopted the *deficiency* metric (Baram *et al.* [1]) that has been widely used in previous studies (Schein and Unga [20]). The deficiency metric between two active learning methods REF and AL is defined by

$$\mathrm{Def}_n(\mathrm{AL},\mathrm{REF}) = \frac{\sum_{t=1}^{n}(\varphi_n(\mathrm{REF}) - \varphi_t(\mathrm{AL}))}{\sum_{t=1}^{n}(\varphi_n(\mathrm{REF}) - \varphi_t(\mathrm{REF}))} \quad (5)$$

where REF is uncertainty sampling method (i.e., the baseline active learning method in this work), and AL is one of our active learning methods such as SUD or density-based re-ranking. $\varphi_t(\mathrm{REF})$ and $\varphi_t(\mathrm{AL})$ denote the evaluation performance (i.e., F1 value in this work) at tth learning iteration of active learning methods REF and AL, respectively. n refers to the evaluation stopping point which is represented as the number of annotated examples at the stopping point. Smaller deficiency value (i.e., < 1.0) indicates AL method is better than REF method. Conversely, a larger value (i.e., > 1.0) indicates a negative result.

C. Results of Different K Values for Density Estimation

To determine an appropriate K value for density estimation, we designed some experiments on the estimation of density for active learning with SUD on Interest dataset. In these experiments, K varies between 5 and 200.

Fig. 4 and Table II summarize the results of SUD methods with different k values for density estimation on the Interest dataset. Table II shows that the best performance (0.510 deficiency) was achieved by SUD with $K = 10$. In the following comparison experiments, we set K value to 10 for density estimation.

The second optimization approach mentioned in Section III-C is used to estimate the density for each unlabeled example during active learning. To implement this approach, we first calculate the similarity of each example pair in the beginning of active learning. For each unlabeled example, all other examples in the unlabeled pool can be first ranked in the decreasing order of similarity to the current example. The similarity of each example pair is estimated only one time during the whole active learning process. If an example is chosen at the ith learning iteration, we remove it immediately from the ranked list of each of the rest of unlabeled examples for the next learning iterations. In practice, it only incurs little computational cost to calculate the density of an unlabeled example x by looking up top-K examples in its ranked list in each learning iteration. Therefore, the implementation of

[1]See http://www.nlplab.com/ontonotes-10-nouns.rar

[2]See http://homepages.inf.ed.ac.uk/s0450736/maxent_ toolkit.html

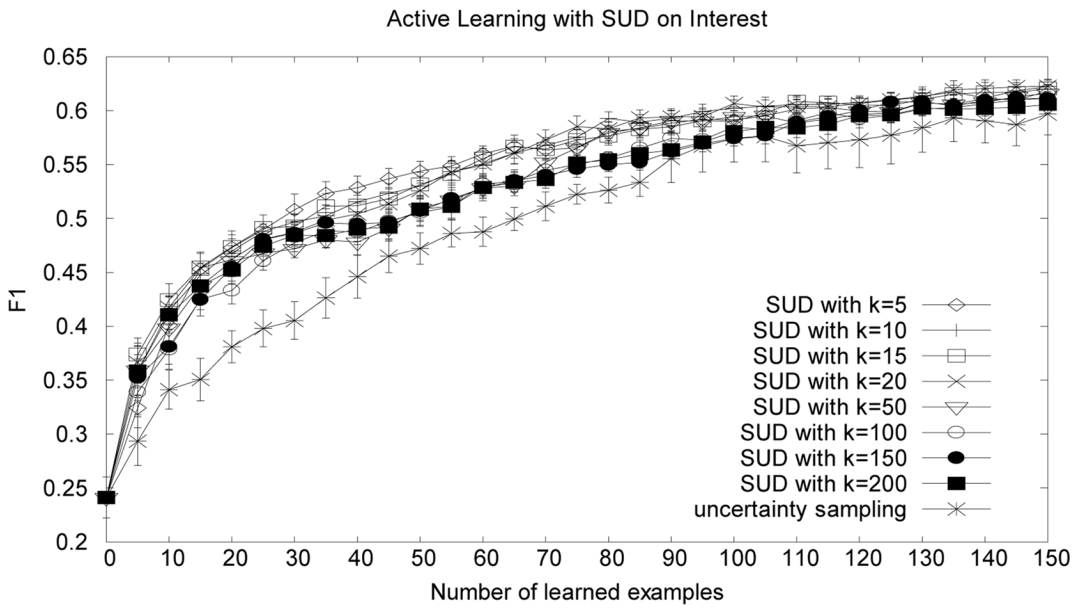

Fig. 4. Results of uncertainty sampling and SUD methods for density estimation in active learning on the *Interest* dataset, with K varying between 5 and 200. Confidence bars (with confidence at 95% level) indicate the variability of competing active learning techniques.

TABLE II
AVERAGE *DEFICIENCY* VALUES ACHIEVED BY SUD METHODS WITH DIFFERENT K VALUES, COMPARED WITH UNCERTAINTY SAMPLING (REF). THE STOP POINT IS 150. THE BOLDFACE NUMBER INDICATES THE BEST PERFORMANCE

Methods	SUD	SUD	SUD	SUD
(K)	(K=5)	(K=10)	(K=15)	(K=20)
Deficiency	0.539	**0.510**	0.517	0.514
Methods	SUD	SUD	SUD	SUD
(K)	(K=50)	(K=100)	(K=150)	(K=200)
Deficiency	0.635	0.702	0.688	0.694

density estimation used by SUD and density-based re-ranking methods is very efficient. In the density-based re-ranking algorithm, the number of candidates (i.e., N in step 2) is set to 10 based on analysis of the experimental results shown in Fig. 4.

D. Experimental Results

Fig. 5 and Table III show the results of various active learning methods for WSD and TC tasks. Fig. 5 shows that in comparison to uncertainty sampling, SUD achieves statistically significant improvement on 4 out of 5 datasets: Interest, Comp2a, Comp2b, and WebKB. On the Line dataset, SUD achieves a deficiency of 0.98 which indicates slightly better performance than that of uncertainty sampling. Similarly, as seen from Fig. 5, compared to uncertainty sampling, density-based re-ranking achieves statistically significant improvement on the Line, Comp2a, Comp2b, and WebKB data sets, and similar performance on the Interest dataset. It is noteworthy that SUD outperforms density-based re-ranking on the Interest, Comp2b and WebKB datasets. However, density-based re-ranking outperforms SUD on the Line dataset.

Comparing results on 10 subsets of OntoNotes shown in Table III, SUD and re-ranking methods on most subsets achieve similar or slightly better performance than uncertainty sampling. The anomaly for SUD lies on the *point* and *order* subsets,

TABLE III
AVERAGE *DEFICIENCY* VALUES ACHIEVED BY VARIOUS ACTIVE LEARNING METHODS, COMPARED WITH UNCERTAINTY SAMPLING. THE STOPPING POINT IS 150. THE BOLDFACE NUMBERS INDICATE THE WORSE PERFORMANCE

Data sets		SUD	Re-ranking
Interest		0.49	0.91
Line		0.98	0.59
Comp2a		0.62	0.57
Comp2b		0.30	0.56
WebKB		0.54	0.79
OntoNotes	rate	0.98	0.87
	president	0.99	**1.74**
	people	0.92	0.91
	part	0.71	0.85
	point	**1.35**	0.96
	director	0.97	0.98
	revenue	0.93	**1.28**
	bill	0.97	0.96
	future	0.81	0.93
	order	**1.22**	0.84

and for density-based re-ranking on the *president* and *revenue* subsets. From Table I we can see that those disambiguated words in OntoNotes have very skewed sense distributions. By analyzing the experimental results we found that density criterion makes the learner tend to select the examples belonging to the predominant class. In such a case, the resulting classifier would achieve unsatisfactory F1 performance. In other words, the density criterion seems to have possibly negative effects on active learning performance in terms of F1 metric on data sets that have very skewed class distribution.

As discussed in Section IV, our density-based re-ranking technique can be applied to active learning with non-probabilistic classifiers such as SVMs. We design some experiments applying density-based re-ranking for active learning with SVMs to the *Interest* dataset, as shown in Fig. 6. In these experiments, we adopted the one-against-all strategy to build

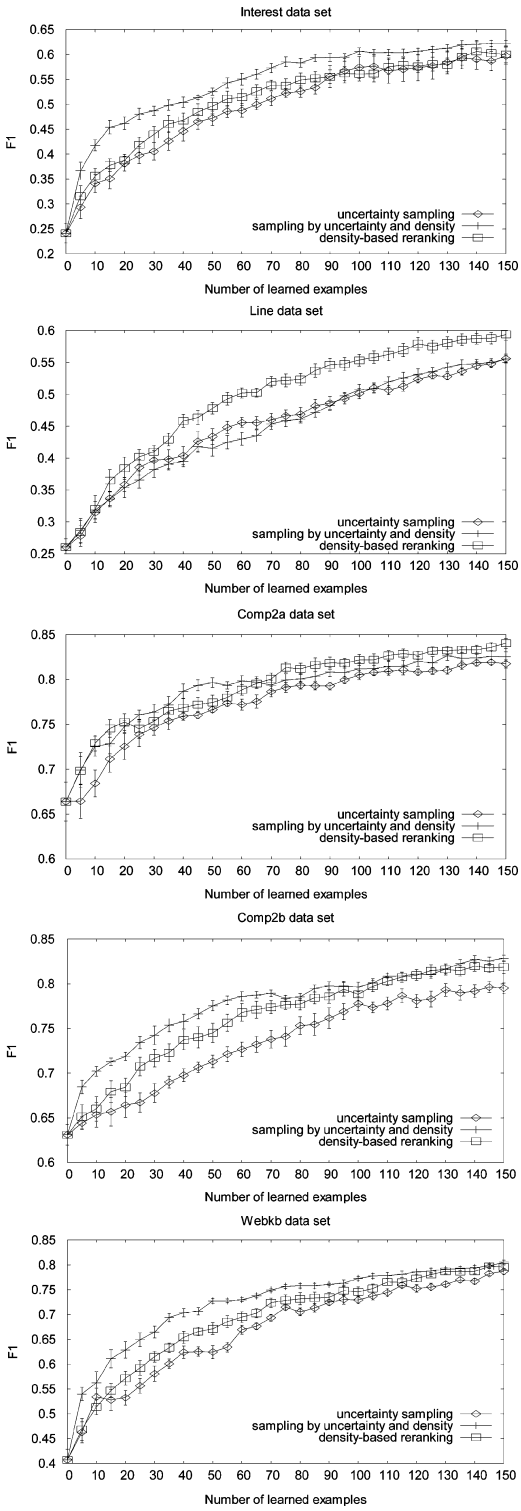

Fig. 5. Effectiveness of various selective sampling techniques in active learning for WSD and TC tasks on five evaluation data sets. Confidence bars (with confidence at 95% level) indicate the variability of competing active learning techniques.

a multi-class SVM classifier which is an ensemble of binary classifiers. In this case, the uncertainty of an unlabeled example x can be defined as the absolute value of the difference between the two largest outputs of the decision functions (Schapire *et al.* [19], Shen *et al.* [22], Vlachos [24]).

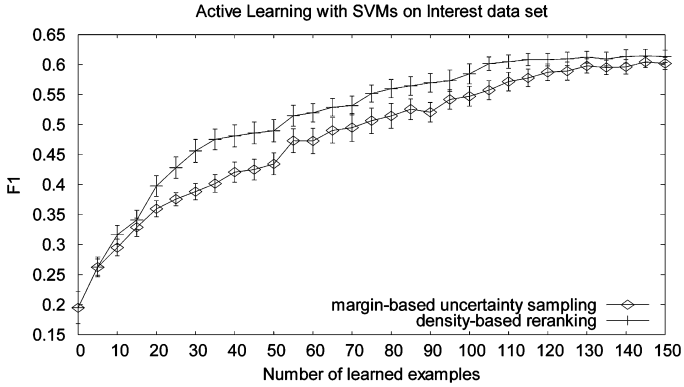

Fig. 6. Results of margin-based uncertainty sampling and density-based re-ranking methods for active learning with SVMs on the Interest data set.

Fig. 6 shows the effectiveness of applying uncertainty sampling and density-based re-ranking for active learning with SVMs. Compared with margin-based uncertainty sampling, density-based re-ranking achieves statistically significant improvement on the Interest dataset.

In SUD and density-based re-ranking, the density criterion is used to avoid selecting unlabeled examples with low density degree, which are likely to be outliers. It is noteworthy to investigate how much agreement there is between the examples selected by uncertainty sampling and one of our density-based methods. In the following experiments, the agreement percentage (AP) between two methods M1 and M2 is estimated by

$$\mathrm{AP}(M1, M2) = \frac{N_{\mathrm{same}}(M1, M2)}{N} \tag{6}$$

where N is the total number of learnt examples until the current iteration. $N_{\mathrm{same}}(M1, M2)$ is the total number of same examples learnt by M1 and M2 until the current iteration.

Figs. 7 and 8 show that the agreement between uncertainty sampling and our density-based methods for active learning on the Interest dataset, respectively. Fig. 7 shows that the highest agreement of 31% is achieved between SUD and density-based re-ranking, and the lowest agreement of 11% is obtained by the pair of uncertainty sampling and SUD at the 150th iteration. For the pair of uncertainty sampling and SUD, the first same chosen example appears in the 75th learning iteration. The first same chosen example for the pair of uncertainty sampling and density-based re-ranking methods appears at the 25th iteration. However, since SUD and density-based re-ranking methods use the density criterion, the first example selected by both methods is the same.

In active learning with SVMs, margin-based uncertainty sampling and density-based re-ranking obtains only 15% agreement at the 150th iteration as shown in Fig. 8. It is interesting to see that the first chosen example by margin-based uncertainty sampling and density-based re-ranking methods is the same. However, the second same example chosen by both methods appears at the 20th iteration.

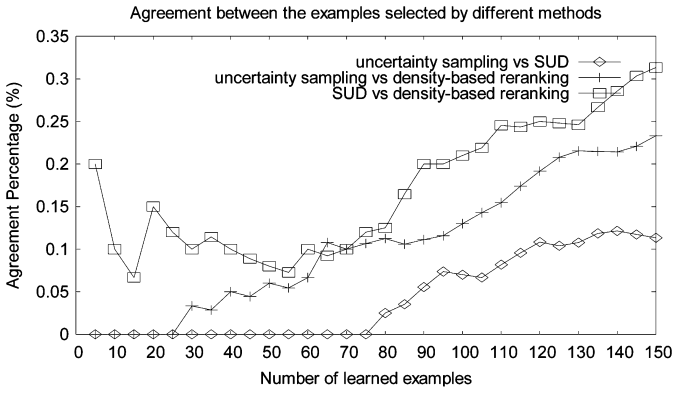

Fig. 7. Results of agreement between the examples selected by different selective sampling methods for active learning with ME-based classifier on the Interest data set.

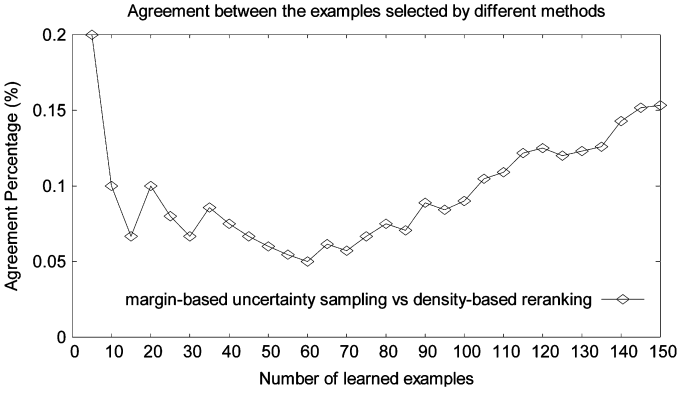

Fig. 8. Results of agreement between the examples selected by margin-based uncertainty sampling and density-based re-ranking methods for active learning with SVMs on the Interest data set.

VI. Conclusion and Future Work

In this paper, we addressed the outlier problem of uncertainty sampling, and proposed two approaches, SUD and density-based re-ranking, in which density and uncertainty criteria are considered simultaneously to select the most informative unlabeled example for human annotation at each learning cycle. The density-based re-ranking techniques can be applied for committee-based sampling for active learning, but for the purpose of applying our SUD to committee-based sampling, we should adopt other uncertainty measurements such as *vote entropy* (Seung *et al.* [21]) to measure the uncertainty of each unlabeled example.

Misclassified unlabeled examples may convey more information than correctly classified unlabeled examples, because learning such misclassified examples has high chance of affecting the position of decision boundaries in the next learning cycle (Chen *et al.* [5], Zhu *et al.* [30]). However, in practice, it is almost impossible to exactly recognize which unlabeled example is misclassified, because the true label of each unlabeled example is unknown in advance. In our future work, we plan to study how to automatically determine whether an unlabeled example has been misclassified. One line of thinking is to make an assumption that an unlabeled example may be misclassified if an unlabeled example is automatically assigned to two different labels during two recent consecutive learning cycles.[3] We think that the switching of a labeling decision at the decision boundary may be attributable to the fact that the unlabeled example is misclassified. In future work, we will further study how to make use of misclassified information to select the most useful examples for human annotation, and how to apply our proposed techniques in the setting of online active learning with probabilistic and non-probabilistic classifiers.

References

[1] Y. Baram, E. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *J. Mach. Learn. Res.*, vol. 5, pp. 255–291, 2004, (2004).
[2] A. L. Berger, V. J. Della Pietra, and S. A. Della Pietra, "A maximum entropy approach to natural language processing," *Comput. Linguist.*, vol. 22, no. 1, pp. 39–71, 1996.
[3] R. Bruce and J. Wiebe, "Word sense disambiguation using decomposable models," in *Proc. 32nd Annu. Meeting Assoc. Comput. Linguist.*, 1994, pp. 139–146.
[4] Y. S. Chan and H. T. Ng, "Domain adaptation with active learning for word sense disambiguation," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguist.*, 2007, pp. 49–56.
[5] J. Chen, A. Schein, L. Ungar, and M. Palmer, "An empirical study of the behavior of active learning for word sense disambiguation," in *Proc. Main Conf. Human Lang. Technol. Conf. North American Chap, Assoc. Comput. Linguist.*, 2006, pp. 120–127.
[6] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Mach. Learn.*, vol. 15, no. 2, pp. 201–221, 1994.
[7] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, "Active learning with statistical models," *J. Artif. Intell. Res.*, vol. 4, pp. 129–145, 1996, (1996).
[8] M. Collins and T. Koo, "Discriminative re-ranking for natural language parsing," in *Proc. 17th Int. Conf. Mach. Learn.*, 2000, pp. 175–182.
[9] E. Hovy and C.-Y. Lin, , M. Maybury and I. Mani, Eds., "Automated Text Summarization in SUMMARIST," in *Advances in Automatic Text Summarization*. Cambridge, MA: MIT Press, 1998, pp. 18–24.
[10] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel, "Ontonotes: The 90% solution," in *Proc. Human Lang. Technol. Conf. North Amer. Chap. ACL*, 2006, pp. 57–60.
[11] C. Leacock, G. Towell, and E. Voorhees, "Corpus-based statistical sense resolution," in *Proc. ARPA Workshop Human Lang. Technol.*, 1993, pp. 260–265.
[12] Y. K. Lee and H. T. Ng, "An empirical evaluation of knowledge sources and learning algorithm for word sense disambiguation," in *Proc. ACL-02 Conf. Empirical Methods in Natural Lang. Process.*, 2002, pp. 41–48.
[13] D. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," in *Proc. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1994, pp. 3–12.
[14] A. McCallum and K. Nigam, "Employing EM in pool-based active learning for text classification," in *Proc. 15th Int. Conf. Mach. Learn.*, 1998a, pp. 350–358.
[15] A. McCallum and K. Nigam, "A comparison of event models for naïve bayes text classification," in *Proc. AAAI-98 Workshop Learning for Text Categorization*, 1998b, pp. 41–48.
[16] H. T. Ng and H. B. Lee, "Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach," in *Proc. 34th Annu. Meeting Assoc. Comput. Linguist.*, 1996, pp. 40–47.
[17] A. Philpot, E. Hovy, and P. Pantel, "The omega ontology," in *Proc. OntoLex 2005—Ontologies and Lexical Resources*, 2005, pp. 59–66.
[18] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 441–448.
[19] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," in *Proc. 14th Int. Conf. Mach. Learn.*, 1997, pp. 322–330.
[20] A. I. Schein and L. H. Ungar, "Active learning for logistic regression: An evaluation," *Mach. Learn.*, vol. 68, no. 3, pp. 235–265, 2007.
[21] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proc. 5th Annu. ACM Conf. Comput. Learn. Theory*, 1992, pp. 287–294, ACM Press.

[3]For example, an unlabeled example x was classified into class A at ith iteration, and class B at $i+1$th iteration.

[22] D. Shen, J. Zhang, J. Su, G. Zhou, and C.-L. Tan, "Multi-criteria-based active learning for named entity recognition," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguist.*, 2004, pp. 589–596.
[23] M. Tang, X. Luo, and S. Roukos, "Active learning for statistical natural language parsing," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 120–127.
[24] A. Vlachos, "A stopping criterion for active learning," *Comput. Speech Lang.*, vol. 22, no. 3, pp. 295–312, 2008.
[25] Y. Zhang, A. S. Hildebrand, and S. Vogel, "Distributed language modeling for N-best list re-ranking," in *Proc. 2006 Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 216–223.
[26] J. Zhu and E. Hovy, "Active learning for word sense disambiguation with methods for addressing the class imbalance problem," in *Proc. 2007 Joint Conf. Empirical Methods in Natural Lang. Process. Comput. Natural Lang. Learn.*, 2007, pp. 783–790.
[27] J. Zhu, H. Wang, and B. K. Tsou, "A density-based re-ranking technique for active learning for data annotations," in *Proc. ICCPOL09*, Hong Kong, 2009, pp. 1–10.
[28] J. Zhu, H. Wang, and E. Hovy, "Learning a stopping criterion for active learning for word sense disambiguation and text classification," in *Proc. 3rd Int. Joint Conf. Natural Lang. Process.*, 2008a, pp. 366–372.
[29] J. Zhu, H. Wang, T. Yao, and B. Tsou, "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification," in *Proc. 22nd Int. Conf. Comput. Linguist.*, 2008b, pp. 1137–1144.
[30] J. Zhu, H. Wang, and E. Hovy, "Multi-criteria-based strategy to stop active learning for data annotation," in *Proc. 22nd Int. Conf. Comput. Linguist.*, 2008c, pp. 1129–1136.

Jingbo Zhu (M'09) received the Ph.D. degree in computer science from Northeastern University, Shenyang, China, in 1999.

He has been with the Institute of Computer Software and Theory, Northeastern University, since 1999. Currently, he is a Full Professor in the Department of Computer Science, and is in charge of research activities within the Natural Language Processing Laboratory. He was a Visiting Scholar at ISI, University of Southern California at Los Angeles, from 2006 to 2007. He has published more than 100 papers, and holds four U.S. patents. His current research interests include natural language parsing, machine translation, text topic analysis, knowledge engineering, machine learning, and intelligent systems.

Huizhen Wang received the Ph.D. degree in computer science from Northeastern University, Shenyang, China, in 2008.

She has been with the Institute of Computer Software and Theory, Northeastern University, since 2008. Currently, she is a Lecturer in the Department of Computer Science. She has published more than 30 papers. Her current research interests include knowledge engineering, opinion mining, and machine learning for natural language processing.

Benjamin K. Tsou (M'97) received the M.A. degree from Harvard University, Cambridge, MA, and the Ph.D. degree from the University of California, Berkeley.

He is a member of the Royal Academy of Overseas Sciences of Belgium and Chair Professor of Linguistics and Asian Languages, as well as Director of the Language Information Sciences Research Center of the City University of Hong Kong. Since 1995, he has developed and cultivated the largest (350 million characters, 1.5 million word types by 2008) synchronous corpus of Chinese LIVAC (http://www.livac.org) which makes unique provisions for monitoring linguistic and related trends for application in NLP. He has authored several books and monographs, and over 100 articles, mostly on computational linguistics and Chinese linguistics. He has served on the editorial boards of *Natural Language Processing* (Japan), *International Journal of Computational Linguistics and Chinese Language Processing* (IJCLCLP) (Taipei), *International Journal of Computer Processing of Oriental Languages* (Singapore), *Monograph series in Natural Language Processing*, John Benjamins (Amsterdam), among others. His current research interests include sentiment analysis, computational lexicography and lexicology, and resource developing and natural language processing.

Dr. Tsou is the Founding President of the Asian Federation of Natural Language Processing (AFNLP) and was the Chairman of SIGHAN, ACL. He also serves on the Executive Board of the Chinese Information Processing Society of China.

Matthew Y. Ma (SM'00) received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, the M.S. degree in electrical engineering from the State University of New York at Buffalo, and the Ph.D. degree in electrical and computer engineering from Northeastern University, Boston, MA.

He is currently with Scientific Works, Princeton Junction, NJ, as a Chief Scientist. Prior to that, he has had 11 years tenure as a Senior Scientist at Panasonic R&D Company of America focusing on mobile and imaging research and two other intellectual property firms focusing on patent research. He has 13 granted U.S. patents and is the author of 30 or so conference and journal publications. He is an Associate Editor of the *International Journal of Pattern Recognition and Artificial Intelligence* (IJPRAI). He is the coauthor and author of two books: *Personalization and Recommender Systems* (World Scientific, 2008) and *Fundamentals of Patenting and Licensing for Scientists and Engineers* (World Scientific, 2009). He has been an Affiliated Professor at Northeastern University, Shenyang, China, since 2001. His primary research interest includes patent research, image analysis, pattern recognition, and natural language processing.