Weakly-Supervised Occupation Detection for Micro-blogging Users

Ying Chen[1] and Bei Pei[2]

[1] China Agricultural University, China, 100083
chenying@cau.edu.cn
[2] Key Lab of Information Network Security, Ministry of Public Security, China, 200031
peibei@stars.org.cn

Abstract. In this paper, we propose a weakly-supervised occupation detection approach which can automatically detect occupation information for micro-blogging users. The weakly-supervised approach makes use of two types of user information (tweets and personal descriptions) through a rule-based user occupation detection and a MCS-based (MCS: a multiple classifier system) user occupation detection. First, the rule-based occupation detection uses the personal descriptions of some users to create pseudo-training data. Second, based on the pseudo-training data, the MCS-based occupation detection uses tweets to do further occupation detection. However, the pseudo-training data is severely skewed and noisy, which brings a big challenge to the MCS-based occupation detection. Therefore, we propose a class-based random sampling method and a cascaded ensemble learning method to overcome these data problems. The experiments show that the weakly-supervised occupation detection achieves a good performance. In addition, although our study is made on Chinese, the approach indeed is language-independent.

Keywords: occupation detection, sampling and ensemble learning.

1 Introduction

Micro-blogging platforms such as Twitter and Plurk, not only provide services for users to share information with friends, but also contain plenty of personalization business applications which usually are designed according to users' personal information (i.e. education experience, working experience and hobby). However, as personal information is usually not obligatorily provided by users, this kind of information is often incomplete or omitted. In this paper, we attempt to explore how to automatically detect personal information for micro-blogging users. To better explain our work, in this paper, we adopt the twitter's terminology. A "tweet" refers to a short message a user shares with others; a "following" is a user who the focused user is subscribed to; a "follower" is a user who subscribes to the focused user.

As far as we know, there have been few studies on personal information detection for micro-blogging users. Although intensive studies on personal information detection have been done in the past years, most of them focus on factual (objective) texts,

C. Zong et al. (Eds.): NLPCC 2014, CCIS 496, pp. 299–310, 2014.

such as news reports and homepages. Since those factual texts tend to introduce the focused person, the personal information is often given in the context of the mentions of the person of interest. In contrast, tweets are more like a type of subjective texts. Instead of introducing themselves, micro-blogging users like to express their opinions or describe their activities in their tweets. Therefore, those tweets sometimes do not provide enough explicit personal information, and personal information detection requires more deduction.

As personal information is rather broad, in this paper, we focus only on occupation information. In fact, occupation information can be considered as a hierarchical structure. The first level roughly contains three types of occupations (student, employee and un-employed), and each occupation then is further divided in the next level. In this paper, we explore occupation detection only according to the first-level division, and examine the research issues specifically for micro-blogging user.

In this paper, we first explore how to effectively integrate the micro-blogging information of a user (i.e. tweets and personal descriptions) for occupation detection. Furthermore, we investigate how to deduce the occupation of a user with the aid of some particular tweets of the user. Overall, there are three contributions of our user occupation detection.

First, we automatically construct a user occupation corpus with a set of rules and the rules can infer the occupations of some users according to their personal descriptions if exist. Although this user occupation corpus is very noisy, it can avoid costly human annotation and give a guide to the design of our user occupations detection. From the user occupation corpus, we find that few users explicitly release their un-employed status, and thus, we formulate the user occupation detection as a classification problem with three classes (student, employee and undetermined). Here, "undetermined" users refer to the ones whose occupations cannot be inferred from the given micro-blogging information even by humans.

Second, given the pseudo-annotated user occupation corpus, the following approach is intuitive for our user occupations detection: a supervised classification method (such as SVM, the Maximum Entropy model and so on) is chosen, and the features of an instance are extracted from all tweets of a user. However, this intuitive approach cannot work well because of the three data problems inherited in our pseudo-annotation corpus: data imbalance, data bias and data noise. The data imbalance refers to the imbalance between the three classes, and requires a specific classification approach (i.e., imbalance classification). The data bias refers to the class distribution in our corpus does not follow the real one. The data noise refers to noisy features and noisy pseudo tags in our corpus. In this paper, to overcome the data imbalance and the data bias, we choose a typical imbalance classification approach, which uses MCS (a multiple classifier system [12]) and a sampling method. The sampling method selects the balanced training datasets for the base classifiers of MCS. Furthermore, because the data noise is severe, we find that the typical sampling methods, such as random over-sampling [5] and random under-sampling [4], cannot perform well for our task. Thus, we propose a class-based random sampling method, which is an extension of random under-sampling. The empirical experiments show that our MCS-based user occupation detection system achieves a good performance.

Third, users are much different in terms of the scales of their tweets. For example, some users post several tweets and some have thousands. In fact, we observe that the occupation of a user can be determined only by several occupation-related tweets. Thus, it is better to detect user occupation only according to this kind of tweets. Unfortunately, how to select the occupation-related tweets is also a hard task. In this paper, we propose a cascaded ensemble learning method which selects some occupation-related tweets and uses them to further improve the user occupation detection.

2 Related Work

In this section, we first compare our user occupation detection with previous works on personal information detection, and then briefly present the state-of-the-art studies on imbalanced classification.

2.1 Occupation Detection

For occupation detection, several systems are presented in the bakeoff, searching information about entities in the Web (WePS). WePS works on the occupation detection for a web person, which extract the occupation information of a person from the given webpages. Artiles et al. [2] summarize these systems and find that a rule-based approach [6] is most effective because the approach can capture the structures of some kinds of webpages.

Furthermore, occupation detection can be considered as a sub-problem of Information Extraction (IE). The survey of Sarawagi [17] examines rule-based and statistical methods for IE, and point out that the different kinds of approaches attempt to capture the diversity of clues in texts [1,3,7,16]. Therefore, the properties of texts determine the approaches of occupation detection.

For the occupation detection for micro-blogging users, there are two types of textual information: personal descriptions and tweets. Moreover, these two types of textual information are complement for the user occupation detection. A personal description is a kind of structured texts, and occupation detection on those structured texts is well studied. In fact, the main challenge comes from tweets because tweets have their own characteristics, such as informal expressions, short texts and so on. In this paper, we explore the interaction of these two types of textual information for occupation detection.

2.2 Imbalanced Classification

The common understanding for data imbalance for multi-class classification is that the imbalance exists between the various classes. Because of severe class distribution skews, in most cases, classifiers trained with imbalanced data prefer to annotate test instances with majority class (MA) and ignore the minority class (MI). Thus, imbalanced data requires specific approaches, namely imbalanced classification.

Imbalanced classification has been widely studied in terms of data level and algorithmic level (see the comprehensive review [10]). From the data level, the most important approach is sampling which attempts to balance the class distribution, such as various over-sampling methods that replicate MI instances [5,8-9,18] and various under-sampling methods that remove MA instances[4,11,15,19]. From the algorithmic level, many approaches are proposed, such as cost-sensitive learning, one-class learning, and ensemble learning.

In this paper, we attempt to use a sampling method to solve both the data imbalance and the data bias in our corpus. Although the empirical study [13] shows that under-sampling is most effective for sentiment classification, it does not work well for our task because of the severe noise in our corpus. Thus, we explore how to extent the under-sampling method so that to handle the noisy data.

3 The User Occupation Corpus

In this section, we first introduce the construction of our user occupation corpus (including the corpus collection and the rule-based user occupation detection), and present an in-depth corpus analysis. According to the analysis, we then formulate our user occupation detection task.

3.1 The Corpus Collection

Our user occupation corpus indeed contains a set of users and a user has an information unit containing all information regarding the user (tweets, followings, followers, and personal descriptions). The whole corpus is collected through the following four steps.

1. Two hot topics are chosen from the Sina micro-blogging platform (a Chinese micro-blogging website): one is “College English Test” (a student activity), and the other is “the symptoms for going to work on Monday” (an employee activity). A user who posts a tweet for either of the two topics is considered as a seed. There are totally ~1,800 seeds.
2. Initial a user set which contains all seeds.
3. Beginning with the initial user set, the user set is iteratively increased by incorporating their friends. Here, a “friend” of a focused user refers to a user who is both a following and a follower of the user.
4. For each user in the user set, all related information is crawled from the Sina micro-blogging platform.

Although our corpus is scalable through the iteration of Step 3, due to the limited time, our corpus collect only totally 30,840 users, which is ~0.015% of the whole Sina micro-blog users. However, we can still gain enlightenments on the user occupation detection through our pilot study on this rather small-scale corpus.

3.2 The Rule-Based User Occupation Detection

Because human annotation is time-costing, in this paper, we propose a rule-based user occupation detection system, which automatically detects the occupations of some users according to their personal descriptions if exist.

Although the Sina micro-blogging platform provides templates for users to input their occupation information, we observe that users often do not exactly follow them. For example, a user lists only his working company "Lenovo" without time information. Because of the incompleteness, the rule-based user occupation detection becomes challenging.

In the rule-based user occupation detection, for a user, his/her working experience firstly is examined. If the job information is provided, the user is tagged as "employee". Otherwise, go to next. Secondly, the education experience is examined. If the college information is provided, the user is tagged as "student". Otherwise, the user is tagged as "undermined".

3.3 The Corpus Analysis

In our user occupation corpus, ~74% instances (users) provide their personal descriptions, and however, only ~36% instances prefer to publish their occupation information. Furthermore, our rule-based occupation detection detects the occupations only for 31% users (~17% are students and ~14% are employees). This indicates that only some of Sina micro-blogging users present useful occupation information through their personal descriptions.

After the rule-based occupation detection, a user in our corpus is either an instance with an occupation (namely, a rule-determined instance, which is either a student or an employee) or an "undetermined" instance (namely, a rule-undetermined instance, whose occupation cannot be detected by our rules). In the following section, we examine the rule-determined data and the rule-undetermined data, which contains all rule-determined instances and all rule-undetermined instances, respectively.

***The rule-determined data*:** for the rule-determined data, we reserve ~1000 instances for the development data (namely, the rule-determined dev) and ~1000 instances for the test data (namely, the rule-determined test). These two datasets then are annotated by humans as follows.

An instance is tagged with one of the four tags: student, employee, un-employed and undetermined. For a user, an annotator reads his/her tweets one by one in chronological order (beginning with the most recent tweet). If a tag can be confidently given to the user according to the present tweet, the annotator stops. Finally, if the annotator cannot assign a tag after reading all tweets of the focused user, the "undetermined" tag is given to the user.

According to the human-annotated data, we find that the overall accuracy of the two rule-determined datasets is ~72%. This indicates that ~72% users deliver the same occupation information both in their personal descriptions and in their tweets. Moreover, the real occupation distribution in the two rule-determined datasets is: student (50.8%), employee (36.5%), un-employed (1.2%) and undetermined (11.5%).

***The rule-undetermined data*:** for the rule-undetermined data, we reserve ~1000 instances for the development data (namely, the rule-undetermined dev) and ~1000 instances for the test data (namely, the rule-undetermined test). Similar to the rule-determined data, these two datasets are annotated by humans. For the two rule-undetermined datasets, the overall accuracy is ~28%, and the real occupation distribution is: student (40.2%), employee (31.4%), un-employed (0.4%) and undetermined (28.0%).

3.4 Task Formulation

According to the real occupation distribution for the rule-determined data and the rule-undetermined data, we formulate the user occupation detection task as follows. Because the tag "un-employed" occupies such a low percentage (~1%) that we decide to ignore them in our current work, the occupation detection becomes a classification problem with three classes (student, employee, and undetermined). The low percentage of "un-employed" may be due to the fact that most of un-employed users do not like to mention their job status.

Regarding the data setting for our user occupation detection, except for the human-annotated instances in Section 3.3, we use all instances with the tags outputted from the rule-based user occupation detection as the training data (namely, the pseudo-training data). In the pseudo-training data, 15.2% instances are students, 12.6% are employees, and 69.2% are undetermined instances.

4 The MCS-Based User Occupation Detection

In this section, we first examine the data problems in the pseudo-training data, and then introduce the MCS-based user occupation detection. Particularly, we present our class-based random sampling method and cascaded ensemble learning method.

4.1 The Overview of the MCS-Based User Occupation Detection

Our user occupation detection is a three-class classification problem. Given the pseudo-training data, many common supervised classification methods cannot work effectively because of the following three data problems.

1. Data imbalance: the data imbalance problem is severe in the pseudo-training data. For example, the imbalance ratio between "undetermined" and "employee" is ~4.6 (69.2% vs. 15.2%) although only some "undetermined" instances are real undetermined by humans.
2. Data bias: our user occupation corpus is somewhat biased to the users with occupations ("student" or "employee") because of the selection of topics during the corpus collection (see Section 3.1). Thus, the pseudo-training data may not reflect the real occupation distribution even it can be annotated by humans. In the other hand, it is difficult to capture the real occupation distribution of micro-blogging users because it is changeable.

3. Data noise: there are two kinds of data noises in the pseudo-training data: noisy features and noisy pseudo tags. First, micro-blog is popular because it has fewer restrictions on writing styles. Thus, tweets themselves are intrinsically noisy, and furthermore, the features based on tweets are likely to be noisy. Second, since the rule-based user occupation detection achieves only a decent performance (~72% for the rule-determined data and ~28% for the rule-undetermined data, see Section 3.3), the tags in the pseudo-training data are severely noisy, particularly for "undetermined".

Although data imbalance and data bias seem different from each other, both of them involve the skewed class distribution in the pseudo-training data, and they can be solved with the same approach – a sampling method which can select a balanced training dataset for a classifier. After taking the data noise into account, we propose a class-based random sampling method. Moreover, considering that users have various-scale tweets, we propose a cascaded ensemble learning method which integrates the occupation information of all tweets and the occupation information of individual tweets to do user occupation detection.

In this paper, we choose MCS as the framework of our system. There are two stages in MCS: training and test. During the training stage, a base classifier is trained with a supervised classification method as well as some training instances (users) selected by the class-based random sampling method. For each training instance, all of the tweets are catenated into a document on which feature extraction works. During the test stage, the cascaded ensemble learning method is used. The class-based random sampling method and the cascaded ensemble learning method are described as follows.

4.2 The Class-Based Random Sampling

Random under-sampling is a typical sampling method for imbalance classification. It randomly selects a subset of the MA instances from the initial training data and then combines them with all of the MI instances to form the training dataset for a base classifier. Our analysis shows that random under-sampling or its variations perform well because they satisfy the following two conditions: (1) the MI instances are high-quality, such as human-annotated data [13]; (2) all of the MI instances are used in a base classifier. Unfortunately, our pseudo-training data indeed is very noisy, and a base classifier will be confused if its training dataset includes all of the MI instances. In this paper, we propose a class-based random sampling method.

In general, the class-based random sampling method (shown in Figure 1) guarantees that the instances belonging to the same class are equally likely to be chosen. The class-based random sampling has only one parameter K, and our empirical experiments (see Section 5.2) show that the parameter has a big impact to the effect of the sampling. In addition, random under-sampling is actually an extreme case of the class-based random sampling method where K is chosen the maximum.

Although our class-based random sampling looks simple, it can effectively solve the aforementioned three data problems. First, the training dataset for a base classifier

is balanced because each class contributes exact K instances. Thus, the problem of data imbalance and data bias can be avoided. Second, our class-based random sampling selects different subsets of the initial training data for base classifiers. Given the noisy pseudo-training data, the more possible the subsets are, the more possible it is for an effective feature to be selected by a base classifier. This is also the fundamental difference between our class-based random sampling and various under-sampling methods.

4.3 The Cascaded Ensemble Learning

We propose a cascaded ensemble learning method, as shown in Figure 2. In general, there are two steps. Firstly, for a test user, two tags are gotten from the two ensemble learning methods, the whole-tweets-based ensemble learning and the individual-tweet-based ensemble learning. Secondly, according to the two tags, a rule-based ensemble learning is used to get the final tag of the test user. The three ensemble learning methods are explained as follows. Notice, "user" and "instance" are not exchangeable in this section.

The whole-tweets-based ensemble learning: it is very simple ensemble learning based on the majority vote. Firstly, for a test user, a set of tags are obtained from the base classifiers. Secondly, the majority vote is applied to the set of tags to get the final tag of the test user. Notice, similar to the training of the base classifiers, a test instance inputted to a base classifier is a document which contains all tweets posted by the test user.

The individual-tweet-based ensemble learning: although different users post various-scale tweets, it is often a case that the occupation of a user is determined only by several occupation-related tweets. Therefore, we propose an individual-tweet-based ensemble learning, explained as follows. Step 1 and 2 attempt to select some occupation-related tweets, and Step 3 and 4 try to use the occupation-related tweets for user occupation detection.

1. Given a test user, detect the occupation tag for each of his/her tweets. The procedure is similar to the whole-tweets-based ensemble learning except the following two things: a test instance inputted to a base classifier is an individual tweet, and the tag is attached with the support rate which is calculated during the majority vote.
2. For each individual tweet, if its tag is "student" or "employee", examine its support rate. If the support rate is greater than the given threshold, the individual tweet is considered as an occupation-related tweet. Since the base classifiers do not perform very well, we treat only the tweets whose tags have high confidences as occupation-related tweets.
3. If the test user has an occupation-related tweet, calculate the votes of "student" and "employee", and go to next. Otherwise, tag the user "undetermined" and stop. The vote of "student" ("employee") is the number of the tweets whose tags are "student" ("employee").

4. With the votes of "student" and "employee", the majority vote is used to get the final tag of the user.

The rule-based ensemble learning: From the error analysis of the whole-tweets-based ensemble learning, we observe that the following two kinds of confusions often occur: "undetermined vs. student", and "undetermined vs. employee". Therefore, we attempt to correct the "undetermined" tags outputted from the whole-tweets-based ensemble learning as follows. For a test user, if the tag from the whole-tweets-based ensemble learning is "student" or "employee", use this tag as the final one. Otherwise, use the tag from the individual-tweet-based ensemble learning as the final one.

Input: the initial training data, K (a value which controls the size of the outputted training dataset)
Output: a training dataset for a base classifiers
Procedure:
1. For each class c_i (i = 0 to M, M is the class number), K instances are randomly selected from the instances whose tags are c_i in the initial training data.
2. $M*K$ instances are combined to form a training dataset for a base classifier.

Notice:

$1 \le K \le Min_{i=0}^{M}$ (the number of the intances tagged with c_i in the initial training data)

Fig. 1. The class-based random sampling method

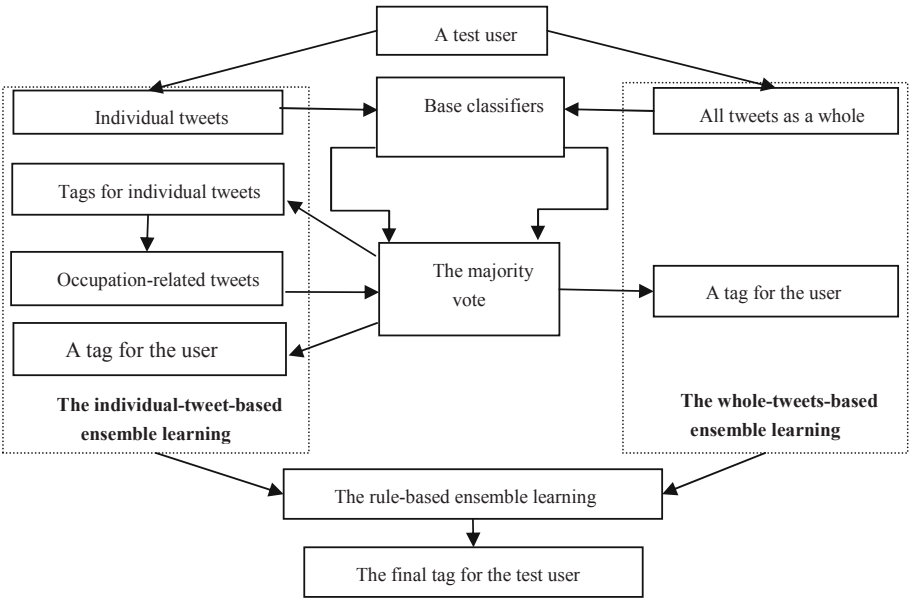

Fig. 2. The cascaded ensemble learning method

5 Experiments

5.1 Experiment Settings

Regarding the experiment data setting, we use the pseudo-training data as the initial training data, the rule-determined test and the rule-undetermined test as the test datasets, and the rule-determined dev and the rule-undetermined dev as the development datasets (see Section 3.3). In addition, in the dev/test data, tag "un-employed" annotated by humans is replaced with tag "undetermined".

To examine the performances of our MCS-based occupation detection, we implemented two baselines for comparisons. One is a common classification (SC), which uses the following dataset to train one and only one classifier: all of "student" instances (4594 instances), all of "employee" instances (3805 instances) and some of "undetermined" instances (4594 instances). The other baseline (UndSamp+WTEnsem) is the under-sampling used in [13]. Similar to our occupation detection, it is also MCS-based and uses random under-sampling and the whole-tweets-based ensemble learning. Moreover, four common measures are chosen for evaluation, i.e. precision (Prec), recall (Rec), F-score (Fs), and accuracy (Acc).

For any supervised classifiers, such as SC and the base classifiers in a MCS-based framework, the Maximum Entropy model implemented by the package Mallet1 is chosen as the classification method, and the bag of words is used as features.

Regarding the parameter setting in the experiments, all of them are learned from the development datasets. In particular, two parameters, L and K, are very important. L is the number of the base classifiers in a MCS-based framework, and K is the parameter of our class-based random sampling. In our experiments, K is 500 and L is100.

5.2 The Performances of Different Occupation Detection Models

Table 1 and 2 list the performances of the four occupation detection models for the two test datasets, the rule-determined test (rule-det) and the rule-undetermined test (rule-undet), respectively. To examine our sampling and ensemble learning separately, we develop two MCS-based occupation detection models: RanSamp+WTEnsem and RanSamp+CasEnsem. The former uses the class-based random sampling and the whole-tweets-based ensemble learning, and the latter uses the class-based random sampling and the cascaded ensemble learning.

From Table 1 and 2, first, we find that the final model, RanSamp+CasEnsem, significantly outperforms the SC model with 9.9% for "rule-det" and 11.4% for "rule-undet" in F score. This indicates that our class-based random sampling and cascaded ensemble learning work very well.

Second, from SC to UndSamp+WTEnsem, a significant improvement is achieved (4.0% for "rule-det" and 4.6% for "rule-undet" in F score). This indicates that a MCS-based framework with a sampling method can effectively overcome the data imbalance and the data bias in our pseudo-training data. Moreover, when the under-sampling (UndSamp+WTEnsem) is replaced by our class-based random sampling

[1] http://mallet.cs.umass.edu/

Table 1. The performances of the different models for the rule-determined test

	Prec	Rec	Fs	Acc
SC	62.2	65.6	60.7	67.6
UndSamp+WTEnsem	64.1	66.8	64.7	73.6
RanSamp+WTEnsem	68.6	73.2	69.8	77.0
RanSamp+CasEnsem	69.5	72.6	***70.6***	***77.7***

Table 2. The performances of the different models for the rule-undetermined test

	Prec	Rec	Fs	Acc
SC	57.9	54.4	50.3	50.0
UndSamp+WTEnsem	60.2	56.1	54.9	54.6
RanSamp+WTEnsem	63.3	59.9	58.4	58.2
RanSamp+CasEnsem	63.3	62.8	***61.7***	***61.9***

(RanSamp+WTEnsem), the performances are further improved (5.1% for "rule-det" and 3.5% for "rule-undet" in F score). This indicates that our class-based random sampling not only can overcome the skewed class distribution problem, but also can effectively reduce the bad effect from the data noise.

Third, from RanSamp+WTEnsem to RanSamp+CasEnsem, a significant improvement (3.3% in F score) is achieved for "rule-undet", and however, a slight improvement (0.8% in F score) for "rule-det". We observe that the improvement of the RanSamp+CasEnsem model is from the significant improvements of "employee" (4.8% in F score) and "undetermined" (4.2% in F score). This indicates that our individual-tweet-based ensemble learning can effectively solve the confusion of "employee vs. undetermined". Moreover, from the error analysis for the RanSamp+WTEnsem model, we find that this kind of confusions occur much more often in "rule-undet" than in "rule-det". Therefore, "rule-undet" takes more benefit from our individual-tweet-based ensemble learning than "rule-det" does.

6 Conclusion

In this paper, we make a pilot study for occupation detection for micro-blogging users, and find that even a simple occupation detection task, which detects only the three types of occupation information, is a difficult research issue.

According to the available micro-blogging resources, we proposed a weakly-supervised user occupation detection which achieves a significant improvement. Through the experiments, we realized the main challenges in the user occupation detection, and examine the contributions of different kind of user information to the occupation detection. We believe that the current study should lay ground for future research on occupation detection for micro-blogging users.

Acknowledgement. This work was supported by Key Lab of Information Network Security, Ministry of Public Security.

References

1. Agichtein, E., Gravano, L.: Snowball: Extracting relations from large plaintext collections. In: Proceedings of the 5th ACM International Conference on Digital Libraries (2000)
2. Artiles, J., Gonzalo, J., Sekine, S.: WePS 2 Evaluation Campaign: Overview of the Web People Search Attribute Extraction Task. In: 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference (2009)
3. Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: IJCAI (2007)
4. Barandela, R., Sanchez, J., Garcia, V., Rangel, E.: Strategies for Learning in Class Imbalance Problems. Pattern Recognition (2003)
5. Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research (2002)
6. Chen, Y., Lee, S.Y.M., Huang, C.: PolyUHK: A Robust Information Extraction System for Web Personal Names. In: 2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference (2009)
7. Gruhl, D., Nagarajan, M., Pieper, J., Robson, C., Sheth, A.: Context and domain knowledge enhanced entity spotting in informal text. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 260–276. Springer, Heidelberg (2009)
8. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Proc. Int'l J. Conf. Intelligent Computing, pp. 878–887 (2005)
9. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In: Proc. Int'l J. Conf. Neural Networks, pp.1322–1328 (2008)
10. He, H., Garcia, E.: Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. Knowledge and Data Engineering 21(9), 1263–1284 (2009)
11. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: Proc. Int'l Conf. Machine Learning (1997)
12. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. John Wiley & Sons, Inc., Hoboken (2004)
13. Li, S., Wang, Z., Zhou, G., Lee, S.Y.M.: Semi-supervised Learning for Imbalanced Sentiment Classification. In: Proceedings of IJCAI (2011)
14. Liu, X., Zhang, S., Wei, F., Zhou, M.: Recognizing Named Entities in Tweets. In: ACL (2011)
15. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory Under Sampling for Class Imbalance Learning. In: Proc. Int'l Conf. Data Mining, pp. 965–969 (2006)
16. Minkov, E., Wang, R.C., Cohen, W.W.: Extracting personal names from emails: Applying named entity recognition to informal text. In: HLT/EMNLP (2005)
17. Sarawagi, S.: Information Extraction. Foundations and Trends in Databases (2008)
18. Wang, B.X., Japkowicz, N.: Imbalanced Data Set Learning with Synthetic Samples. In: Proc. IRIS Machine Learning Workshop (2004)
19. Zhang, J., Mani, I.: KNN Approach to Unbalanced Data Distributions: A Case Study Involving Information Extraction. In: Proc. Int'l Conf. Machine Learning (ICML 2003), Workshop Learning from Imbalanced Data Sets (2003)