Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 751646, 10 pages
http://dx.doi.org/10.1155/2015/751646

Research Article

Improving Classification of Protein Interaction Articles Using Context Similarity-Based Feature Selection

Yifei Chen, Yuxing Sun, and Bing-Qing Han

School of Technology, Nanjing Audit University, 86 W. Yushan Road, Nanjing 211815, China

Correspondence should be addressed to Yifei Chen; yifeichen91@nau.edu.cn

Received 20 October 2014; Revised 13 December 2014; Accepted 14 December 2014

Academic Editor: Fang-Xiang Wu

Protein interaction article classification is a text classification task in the biological domain to determine which articles describe protein-protein interactions. Since the feature space in text classification is high-dimensional, feature selection is widely used for reducing the dimensionality of features to speed up computation without sacrificing classification performance. Many existing feature selection methods are based on the statistical measure of document frequency and term frequency. One potential drawback of these methods is that they treat features separately. Hence, first we design a similarity measure between the context information to take word cooccurrences and phrase chunks around the features into account. Then we introduce the similarity of context information to the importance measure of the features to substitute the document and term frequency. Hence we propose new context similarity-based feature selection methods. Their performance is evaluated on two protein interaction article collections and compared against the frequency-based methods. The experimental results reveal that the context similarity-based methods perform better in terms of the $F1$ measure and the dimension reduction rate. Benefiting from the context information surrounding the features, the proposed methods can select distinctive features effectively for protein interaction article classification.

1. Introduction

An overwhelming number of biological articles are published daily online as a result of growing interest in biological research, especially relating to the study of protein-protein interactions (PPIs). It is essential to classify which articles describe PPIs, that is, to filter out those irrelevant articles from the whole collection of the biological literature. This allows a more efficient extraction of PPIs from the large amount of biological literature. Automated text classification is a key technology to rapidly find relevant articles. Text classification has been successfully applied to various domains such as text sentinel classification [1], spam e-mail filtering [2, 3], author identification [4], and web page classification [5]. Research on protein interaction article classification (IAC) is a text classification task with practical significance in the biological domain.

In the classic text classification framework, a feature extraction mechanism extracts features from raw articles, including all distinct terms (words). This is also known as bag-of-words (BOW) representation for text documents. Hence each article is represented by a multidimensional feature vector where each dimension corresponds to a term (feature) within the literature collection. Even a small literature collection would contain tens of thousands of features [6, 7]. The high dimensionality of the feature space not only increases computational time but also degrades classification performance. Hence, automated feature selection plays an essential role in making the text classification more efficient and accurate by selecting a subset of the most important features [8, 9]. Feature selection is an active research area in many fields such as data mining, machine learning, and rough sets [10–13].

The process of feature selection typically involves certain metrics that are designed for measuring the importance level of features, and the most important features are selected to help in efficient utilization of resources for large scale problems [14]. The existing feature selection methods are mostly based on the statistical information in documents, including term frequency and document frequency [7, 14–18]. Term frequency is the number of times a particular term appears in a document while document frequency

is the number of documents containing that term within the literature collection. One potential drawback of most of these frequency-based feature selection methods is that they treat each feature separately [19]. In other words, these approaches are context independent: they do not utilize the context information around the terms when judging their importance, such as word order, word cooccurrence, multiword chunks, and semantic relationships. However, this information is important for classifying which articles are PPI relevant or nonrelevant. For example, protein names exist in both PPI relevant and nonrelevant documents. So they could have great document frequency or term frequency. However, obviously they are not distinctive terms for the purpose of classification. Hence, it is difficult to measure the importance of all the terms just according to the document frequency or term frequency. After in-depth research we have noticed that, in the PPI relevant documents, the fact that proteins interact with each other is described through the context of those proteins. Meanwhile in nonrelevant documents, the fact that there are no interactions between the particular proteins is also depicted within the context of the documents. The above observation leads us to an interesting issue which is that the context of features in biological articles can be utilized to measure feature importance and to improve the feature selection process. Hence we propose context similarity-based feature selection methods.

This paper is organized as follows: we provide an overview of the existing frequency-based feature selection methods for text classification in Section 2, and this is followed by a definition of the proposed context similarity-based feature selection methods. Then in order to examine the two kinds of methods carefully, the experimental results and discussion are presented in Section 3 to find which one is more useful in the IAC task. This is followed by a conclusion in Section 4.

2. Materials and Methods

2.1. Existing Feature Selection Methods for Text Classification. Feature selection is a process which selects a subset of the most important features. Such selection can help in building effective and efficient models for text classification. Normally, feature selection techniques can be divided into three categories: filters, wrappers, and embedded methods [19]. Filters measure feature importance using various scoring metrics that are independent of a learning model or classifier and select top-N features attaining the highest scores. Univariate filter techniques are computationally fast. However, they do not take feature dependencies into consideration, which was discussed as the motivation of this paper in Section 1. In addition, multivariate filter techniques incorporate feature dependencies to some degree, while they are slower and less scalable than univariate techniques. Wrappers evaluate features using a certain search algorithm together with a specific learning model or classifier. Wrapper techniques consider feature dependencies and provide interaction between features during the subset search processing but are computationally expensive compared with filters. Embedded methods integrate feature selection into the model learning phase. Therefore, they merge with the model or classifier much further than the wrappers. Nevertheless, they are also computationally more intensive than filters.

Considering the high dimensionality of the feature space for text classification tasks, the most frequently used approach for feature selection is the univariate filter method [7]. And among them four document frequency-based methods and two term frequency-based methods that will be discussed in the paper are illustrated as follows, where $P(t_k \mid c_i)$ is the percentage of documents belonging to a category c_i in which the term t_k occurs and $P(t_k \mid \bar{c}_i)$ is the percentage of documents not belonging to a category c_i in which the term t_k occurs. $|c|$ is the number of categories, which is 2 for the IAC task.

(1) Document Frequency (DF). Document frequency (DF) is a simple and effective feature selection method which is based on the assumption that infrequent terms are not reliable in text classification and may degrade the performance [7]. Hence, if the document frequency in which a term occurs is the largest, the term is retained [20]. The DF metrics of the term t_k can be computed as follows:

$$\mathrm{DF}(t_k) = \sum_{i=1}^{|c|} \mathrm{DF}(t_k, c_i) = \sum_{i=1}^{|c|} P(t_k \mid c_i), \tag{1}$$

where $\mathrm{DF}(t_k, c_i)$ is the DF measure of the term t_k in a category c_i and $\mathrm{DF}(t_k)$ is the sum of $\mathrm{DF}(t_k, c_i)$ across all the categories.

(2) Gini Index (GI). Gini Index (GI) was originally used to find the best attributes in decision trees. Shang et al. [15] proposed an improved version of the GI method to apply it directly to text feature selection. The $\mathrm{GI}(t_k, c_i)$ measures the purity of the feature t_k towards a category c_i. Its sum across categories, $\mathrm{GI}(t_k)$, is given as

$$\mathrm{GI}(t_k) = \sum_{i=1}^{|c|} \mathrm{GI}(t_k, c_i) = \sum_{i=1}^{|c|} P(t_k \mid c_i)^2 P(c_i \mid t_k)^2, \tag{2}$$

where $P(c_i \mid t_k)$ is the conditional probability of the feature t_k belonging to a category c_i given presence of the feature t_k.

(3) Class Discriminating Measure (CDM). Class discriminating measure (CDM) is a derivation of the odds ration introduced by Chen et al. [16]. The results in their paper indicate that CDM is a better feature selection approach than information gain (IG). The CDM calculates the effectiveness of the term t_k as follows:

$$\mathrm{CDM}(t_k) = \sum_{i=1}^{|c|} \mathrm{CDM}(t_k, c_i) = \sum_{i=1}^{|c|} \left| \log \frac{P(t_k \mid c_i)}{P(t_k \mid \bar{c}_i)} \right|, \tag{3}$$

where $\mathrm{CDM}(t_k, c_i)$ is the CDM measure of the term t_k in a category c_i and $\mathrm{CDM}(t_k)$ is the sum of $\mathrm{CDM}(t_k, c_i)$ across all the categories.

(4) Accuracy Balanced (Acc2). Accuracy balanced (Acc2) is a two-side metric (it selects both negative and positive features) that is based on the difference of the distributions of a term belonging to a category and not belonging to that category

in the documents. In Forman [14], the Acc2 is studied and claimed to have a performance comparable to the IG and chi-square statistical metrics. The Acc2 of the term t_k can be computed as follows:

$$\mathrm{Acc2}(t_k) = \sum_{i=1}^{|c|} \mathrm{Acc2}(t_k, c_i) = \sum_{i=1}^{|c|} \left| P(t_k \mid c_i) - P(t_k \mid \bar{c}_i) \right|, \tag{4}$$

where $\mathrm{Acc2}(t_k, c_i)$ is the Acc2 measure of the term t_k in a category c_i and $\mathrm{Acc2}(t_k)$ is the sum of $\mathrm{Acc2}(t_k, c_i)$ across all the categories.

(5) Term Frequency Inverse Document Frequency (TFIDF). Term frequency inverse document frequency (TFIDF) is a numerical statistic that is intended to reflect how important a term is to a document in a collection or corpus. One of the simplest filter metrics is computed by summing the TFIDF. Wei et al. [21] introduced category information to TFIDF, which can be reformed using a notation of term frequency $\mathrm{tf}(t_k, c_i)$ that is the number of occurrences of a term t_k in documents from a category c_i. Consider

$$\mathrm{TFIDF}(t_k) = \sum_{i=1}^{|c|} \mathrm{tf}(t_k, c_i) \times \log\left(\frac{1}{P(t_k \mid c_i)} \right). \tag{5}$$

(6) Normalized Term Frequency-Based Gini Index ($GINI_{NTF}$). Normalized term frequency-based Gini Index ($\mathrm{GINI_{NTF}}$) revised the document frequency in the Gini Index metric with the term frequency by Azam and Yao [17]. Experimental results revealed that the term frequency-based metric was useful in feature selection. We reform the formula of $\mathrm{GINI_{NTF}}$ as follows:

$$\mathrm{GINI_{NTF}}(t_k) = \sum_{i=1}^{|c|} \left(\frac{\mathrm{tf_{norm}}(t_k, c_i)}{\mathrm{doc}(c_i)} \right)^2 \times \left(\frac{\mathrm{tf_{norm}}(t_k, c_i)}{\mathrm{tf_{norm}}(t_k, c_i) + \mathrm{tf_{norm}}(t_k, \bar{c}_i)} \right)^2, \tag{6}$$

where $\mathrm{tf_{norm}}(t_k, c_i)$ is the normalized term frequency of t_k in documents from a category c_i and $\mathrm{tf_{norm}}(t_k, \bar{c}_i)$ is the normalized term frequency of t_k in documents not from a category c_i. The normalized values of term frequency are used in the metric so that term frequencies are not influenced by varying lengths of documents.

2.2. Context Similarity-Based Feature Selection Methods. According to the bag-of-words document representation, each raw document in the article collection is transformed into a high-dimensional vector before the process of text classification. In order to address the issues of high dimensionality, the feature filter methods, such as the DF, GI, CDM, and Acc2, are utilized to select the most important features based on document frequency. One potential problem of these frequency-based methods is that they ignore the context relationships between features. As we have discussed in Section 1, context information is essential for the IAC task. When attempting to judge the importance levels of features, it may be advantageous to explicitly compare the similarity shared among contexts in PPI relevant articles or nonrelevant articles. Hence when building the feature selection metrics, we take the significance of context information of each feature into account through the context similarity.

Context Similarity Measure. $\mathrm{sim_{context}}(t_k, c_i)$ is designed to explicitly express the similarity shared by contexts of the term t_k in a certain category c_i. The measure is based on the word cooccurrences and chunks of a pair of context strings $\mathrm{context}_d(t_k, w)$ and $\mathrm{context}_{d'}(t_k, w)$ containing the term t_k within a category c_i. $\mathrm{context}_d(t_k, w)$ denotes a document d containing a term t_k within a context string $\{t_{-wk}, \ldots, t_{-1k}, t_k, t_{1k}, \ldots, t_{wk}\}$, where w is a window size that takes into account w terms before and after the term t_k. The term t_k is contained in another context string of document d', $\mathrm{context}_{d'}(t_k, w)$, which is $\{t'_{-wk}, \ldots, t'_{-1k}, t_k, t'_{1k}, \ldots, t'_{wk}\}$ with the window size w. Using $\mathrm{context}_d$, a multiword phrase chunk containing t_k and its word cooccurrence can be considered to measure the importance of t_k.

First $\mathrm{sim}(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w))$ is defined to measure the similarity between the context string pair $(\mathrm{context}_d, \mathrm{context}_{d'})$ as follows:

$$\mathrm{sim}\left(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w)\right) = \frac{\sum_{w=0}^{|w|} \mathrm{dis}\left(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w)\right)}{|w| + 1}. \tag{7}$$

The sum of all the context strings from 0 to maximum window size $|w|$ is utilized to incorporate word cooccurrence and phrase similarity comprehensively. $|w|$ is used to control the scope of the local information of term t_k involved in the measurement, and trials on the training data show that $|w| = 3$ is the optimal value. In this paper, Jaro-Winkler [22] distance is employed as the distance function of two context strings, $\mathrm{dis}(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w))$, because it was designed and best suited for short strings. The Jaro-Winkler distance is a measure of similarity between two strings, and it is a variant of the Jaro distance metric [23, 24]. The higher the Jaro-Winkler distance for two strings is, the more similar the strings are. The score is normalized such that 0 equates to no similarity and 1 is an exact match.

Then, $\mathrm{sim_{context}}(t_k, c_i)$ is defined to measure the similarity of context in the documents containing the term t_k belonging to a category c_i as follows:

$$\mathrm{sim_{context}}(t_k, c_i) = \sum_{\mathrm{context}_d, \mathrm{context}_{d'} \in c_k} \mathrm{sim}\left(\mathrm{context}_d(t_k, w), \mathrm{context}_{d'}(t_k, w)\right). \tag{8}$$

Context Similarity-Based Feature Selection Methods. In order to elaborate the context similarity-based feature selection metrics, the class discriminating measure (CDM) is considered as an example, which was very useful in reducing the feature set in some application domains. The metric of CDM has been defined in Section 2.1 based on $P(t_k \mid c_i)$ and $P(t_k \mid \bar{c}_i)$. Here $P(t_k \mid c_i)$, the percentage of documents with the term t_k belonging to the category c_i, can also be represented as $\mathrm{doc}(t_k, c_i)/\mathrm{doc}(c_i)$, where $\mathrm{doc}(t_k, c_i)$ is the document

frequency containing the term t_k in the category c_i and $\mathrm{doc}(c_i)$ is the total number of articles in the category c_i. $P(t_k \mid \bar{c}_i)$, the percentage of documents with the term t_k not belonging to the category c_i, can be represented as $\mathrm{doc}(t_k, \bar{c}_i)/\mathrm{doc}(\bar{c}_i)$, where $\mathrm{doc}(t_k, \bar{c}_i)$ is the document frequency containing the term t_k not in the category c_i and $\mathrm{doc}(\bar{c}_i)$ is the total number of articles not in the category c_i. Hence, we can have the following CDM metric:

$$\begin{aligned}\mathrm{CDM}(t_k) &= \sum_{i=1}^{|c|} \left| \log \frac{P(t_k \mid c_i)}{P(t_k \mid \bar{c}_i)} \right| \\ &= \sum_{i=1}^{|c|} \left| \log \left(\frac{\mathrm{doc}(t_k, c_i)}{\mathrm{doc}(c_i)} \cdot \frac{\mathrm{doc}(\bar{c}_i)}{\mathrm{doc}(t_k, \bar{c}_i)} \right) \right|.\end{aligned} \tag{9}$$

In order to make use of the context information of terms and not just the document frequency, we substitute the context similarity measure $\mathrm{sim}_{\mathrm{context}}(t_k, c_i)$ for the document frequency $\mathrm{doc}(t_k, c_i)$. Then the obtained metric with reformed definition is referred to as $\mathrm{CDM}_{\mathrm{cs}}$, class discriminating measure based on context similarity. If the context similarity of a term within a certain text category is greater, the term is more important for text classification. The definition of $\mathrm{CDM}_{\mathrm{cs}}$ is as follows:

$$\mathrm{CDM}_{\mathrm{cs}}(t_k) = \sum_{i=1}^{|c|} \left| \log \left(\frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)} \cdot \frac{\mathrm{doc}(\bar{c}_i)}{\mathrm{sim}_{\mathrm{context}}(t_k, \bar{c}_i)} \right) \right|. \tag{10}$$

The other three document frequency-based metrics defined in Section 2.1 can also be reformed in the same way based on the context similarity to $\mathrm{Acc2}_{\mathrm{cs}}$, $\mathrm{GI}_{\mathrm{cs}}$, and $\mathrm{DF}_{\mathrm{cs}}$:

$$\begin{aligned}\mathrm{Acc2}_{\mathrm{cs}}(t_k) &= \sum_{i=1}^{|c|} \left| \frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)} - \frac{\mathrm{sim}_{\mathrm{context}}(t_k, \bar{c}_i)}{\mathrm{doc}(\bar{c}_i)} \right|, \\ \mathrm{GI}_{\mathrm{cs}}(t_k) &= \sum_{i=1}^{|c|} \left| \left(\frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)} \right)^2 \left(\frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(t_k)} \right)^2 \right|, \\ \mathrm{DF}_{\mathrm{cs}}(t_k) &= \sum_{i=1}^{|c|} \frac{\mathrm{sim}_{\mathrm{context}}(t_k, c_i)}{\mathrm{doc}(c_i)},\end{aligned} \tag{11}$$

where $\mathrm{doc}(t_k)$ is the number of documents containing the term t_k in all the text categories.

3. Results and Discussion

3.1. Experimental Settings

Classification Model $\mathrm{Model}_{\mathrm{SVM\_poly}}$. Support vector machines (SVMs) pioneered by Vapnik [25] are suitable for complex classification problems. Their power comes from the combination of the kernel trick and maximum margin hyperplane separation. SVMs are one of the most successful approaches for classification in text mining [26, 27]. Hence, in this paper, we employ the SVMs with a polynomial kernel as a classification model, $\mathrm{Model}_{\mathrm{SVM\_poly}}$, which is trained and tested using the LIBSVM toolbox [28]. A 10-fold cross-validation is adopted to tune parameters.

Data Sets. An in-depth investigation will be carried out to compare the performances of the four proposed context similarity-based methods and the six existing frequency-based feature selection methods. Two data sets ($\mathrm{Data}_{\mathrm{BCII}}$ and $\mathrm{Data}_{\mathrm{BCIII}}$) are used in our experiments to evaluate the performance, which are both extracted from the BioCreAtIvE (the Critical Assessment of Information Extraction in Biology) challenges. The challenges were set up to evaluate the state of the art of text mining and information extraction in the biological domain.

In the data preprocessing step, all words are converted to lower case, punctuation marks and stop words are removed, and no stemming is used. Consider the following.

(1) $\mathrm{Data}_{\mathrm{BCII}}$: we obtain the $\mathrm{Data}_{\mathrm{BCII}}$ from the Protein Interaction Article Subtask (IAS) of the BioCreAtIvE II challenge [29]. The $\mathrm{Data}_{\mathrm{BCII}}$ is composed of abstracts of 6,172 articles in total, which are taken from a set of MEDLINE articles that are annotated as interaction articles or not according to the guidelines used by the MINT and IntAct databases. There are 5,495 abstracts used as training data and 677 ones as test data. And there are 3,536 and 338 interaction articles, that is, positive examples, in the training and test set, respectively.

(2) $\mathrm{Data}_{\mathrm{BCIII}}$: we obtain the $\mathrm{Data}_{\mathrm{BCIII}}$ from the PPI Article Classification Task (ACT) of the BioCreAtIvE III challenge [30]. The training set (TR) consists of a balanced collection of 2,280 articles classified through manual inspection, divided into PPI relevant and nonrelevant articles. The annotation guidelines for this task were refined iteratively based on the feedback from both annotation databases and specially trained domain experts. The development (DE) and test (TE) set take into account PPI relevant journals based on the current content of collaborating PPI databases. Random samples of abstracts from these journals were taken to generate a development set of 4,000 abstracts (628 PPI relevant and 3,318 nonrelevant abstracts) in total and a test set of 6,000 abstracts (918 PPI relevant and 5,090 nonrelevant abstracts). These two disjointed sets were drawn from the same sample collection.

Performance Measures. Since the applications are restricted to IAC, which is a binary classification task, we measure the performance in terms of $F1$ measure [20]. The $F1$ is determined by a combination of precision and recall. Precision is the percentage of documents that are correctly classified as being positive. Recall is the percentage of positive documents that are correctly classified. The precision, recall, and $F1$ are obtained as

$$\begin{aligned}\mathrm{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \\ \mathrm{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \\ F1 &= \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},\end{aligned} \tag{12}$$

where TP is the number of positive documents that are correctly classified as positive ones, FP is the number of negative documents that are misclassified as positive ones, TN is the number of negative documents that are correctly classified as negative ones, and FN is the number of positive documents that are misclassified as negative ones.

3.2. Experimental Results on the $Data_{BCII}$. First, we test all the feature selection methods when $Model_{SVM\_poly}$ is applied on the $Data_{BCII}$ data set, where there are 29,979 total features extracted using the bag-of-words document representation. The proposed context similarity-based methods, GI_{cs}, DF_{cs}, CDM_{cs}, and $Acc2_{cs}$, are compared with the frequency-based methods, GI, DF, CDM, Acc2, TFIDF, and $GINI_{NTF}$, when the number of the selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%. Figure 1 shows the trend curves of all the feature selection methods, and the optimal parameter value of the window size of context information is 3, which is tuned through 10-fold cross-validation.

Figure 1 indicates that all these feature selection methods have a similar trend on the $Data_{BCII}$, and the proposed methods are more effective. The context similarity-based methods and the term frequency-based methods achieve the best performance when around 4% top important features are selected, while the document frequency-based methods obtain the best performance when around 7-8% features are used. Moreover, the proposed methods outperform the other methods on selecting the top important features to achieve the best $F1$ measure. Among the context similarity-based feature selection methods, when the top 1300 features (4.3% of total number of features) are selected, GI_{cs} acquires the highest $F1$ measure 77.07, which effectively improves the $F1$ measure of the $Model_{SVM\_poly}$ when all the features are used (73.55) by 3.52.

Further, in order to study the performance of all these feature selection methods in more detail, a small feature set in the scope of the top 2000 is used. The corresponding $F1$ measure results are shown in Table 1 when the top 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900 features are selected. The best result for each feature set is shown in bold. It can be seen from Table 1 that the context similarity-based methods outperform those methods based on the document frequency or term frequency. The last column of Table 1 presents the best performance of the $Model_{SVM\_poly}$ that various feature selection methods can achieve, and the size of selected features when the best performance is achieved is illustrated in the parentheses. It can be seen that, compared with the four document frequency-based methods, the TFIDF and the $GINI_{NTF}$ perform better, which shows that term frequency is a relatively more important factor than document frequency. Moreover, all the context similarity-based methods achieve better performance with fewer selected features, and among them the GI_{cs} performs the best on the $Data_{BCII}$. Hence, the proposed method can extract more effective information from context similarity measure of term cooccurrences and chunks than just calculating the document frequency or term frequency. This context information is helpful when measuring the importance of features to boost the performance.

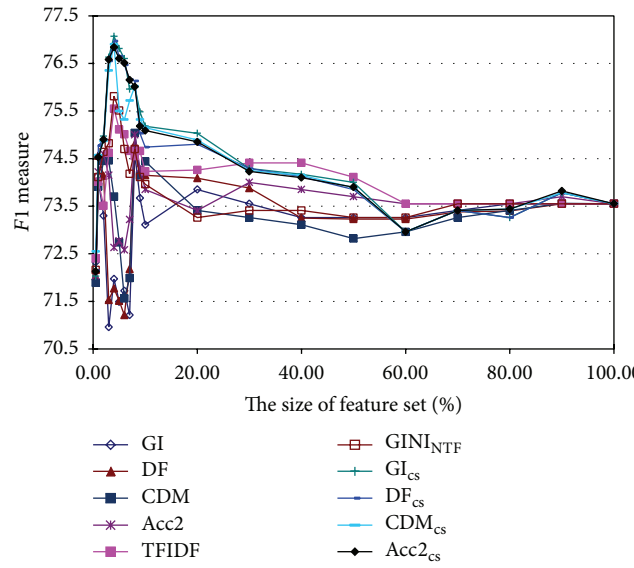

Figure 1: The $F1$ performance curves of all the feature selection methods on the $Data_{BCII}$ when the number of the selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%.

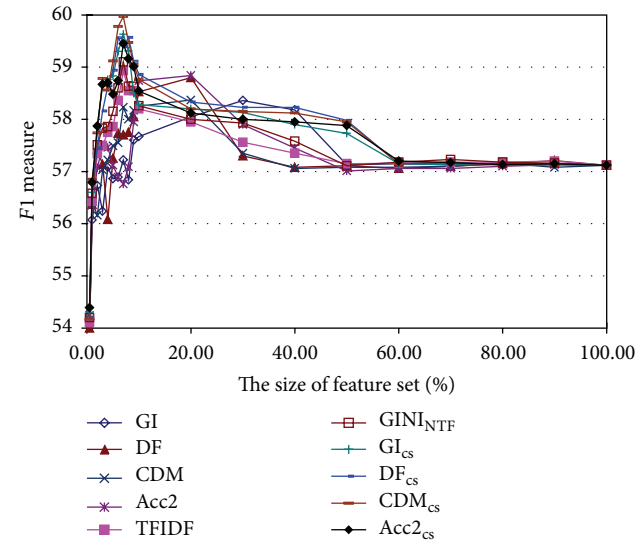

Figure 2: The $F1$ performance curves of all the feature selection methods on the $Data_{BCIII}$ when the number of selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%.

3.3. Experimental Results on the $Data_{BCIII}$. Then, we test the proposed feature selection methods on the $Data_{BCIII}$ when the number of selected features is the top 0.5%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100%, where there are 23,084 features extracted using the bag-of-words representation in total. Figure 2 shows the trend curves of the $F1$ measure versus different sizes of selected features. From Figure 2 we can see that when around 7% top important features are used, the

Table 1: The $F1$ measure results when the $\text{Model}_{\text{SVM\_poly}}$ is applied to the $\text{Data}_{\text{BCII}}$ when the top 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900 features are selected. In each column, the bold value indicates the best performance for each feature set when various feature selection methods are used, respectively. The "best" column presents the best performance that various feature selection methods can achieve, and the numbers in the parentheses are the corresponding sizes of feature sets.

Number of features	100	300	500	700	900	1100	1300	1500	1700	1900	best
GI	72.32	74.04	73.30	73.34	70.96	71.79	71.97	71.49	71.27	71.21	74.60(2300)
DF	72.02	74.56	74.14	72.50	71.53	72.21	71.77	71.51	71.21	72.18	74.80(2300)
CDM	71.89	73.91	74.45	74.49	74.46	74.25	73.70	72.73	71.56	71.99	75.04(2100)
Acc2	72.40	74.22	74.60	73.49	74.15	74.01	72.63	72.77	72.58	73.22	75.00(2100)
TFIDF	72.40	74.01	73.51	73.61	74.62	75.11	75.55	75.11	75.01	74.66	75.55(1300)
GINI_{NTF}	72.16	74.11	74.64	74.70	74.82	75.41	75.81	75.51	74.70	74.18	75.81(1300)
GI_{cs}	72.03	74.52	**74.97**	**76.14**	**76.63**	**76.66**	**77.07**	**76.81**	**76.60**	75.96	**77.07**(1300)
DF_{cs}	72.15	**74.77**	74.90	76.09	76.55	76.60	76.97	76.65	76.47	**76.16**	76.97(1300)
CDM_{cs}	**72.55**	74.57	74.87	75.99	76.35	76.47	76.90	75.50	75.38	75.81	76.90(1300)
Acc2_{cs}	72.13	74.53	74.90	76.06	76.58	76.61	76.84	76.60	76.51	76.15	76.84(1300)

Table 2: The $F1$ measure results when the $\text{Model}_{\text{SVM\_poly}}$ is used on the $\text{Data}_{\text{BCIII}}$ when the top 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900 features are selected. In each column, the bold value indicates the best performance for each feature set when various feature selection methods are used, respectively. The "best" column presents the best performance that various feature selection methods can achieve, and the numbers in the parentheses are the corresponding sizes of feature sets.

Number of features	100	300	500	700	900	1100	1300	1500	1700	1900	best
GI	52.16	56.07	56.73	56.24	57.08	56.86	56.89	57.34	57.22	56.84	58.36(5900)
DF	50.91	56.37	57.42	57.14	56.98	57.25	57.73	57.09	57.70	57.75	58.80(4300)
CDM	52.13	56.43	56.16	57.03	57.22	57.49	57.55	58.27	58.23	58.00	58.37(4500)
Acc2	52.24	56.35	57.07	57.48	57.46	57.12	56.89	57.14	56.76	57.09	58.84(3900)
TFIDF	52.10	56.43	57.34	57.50	57.75	57.86	58.26	58.51	58.93	58.55	58.93(1700)
GINI_{NTF}	52.20	56.58	57.51	57.83	57.86	58.05	58.60	58.83	59.10	58.63	59.10(1700)
GI_{cs}	52.30	56.62	57.80	58.66	58.76	58.83	59.30	59.51	59.63	59.31	59.63(1700)
DF_{cs}	52.21	56.80	57.45	58.16	58.64	58.94	59.56	59.39	59.39	**59.57**	59.57(1900)
CDM_{cs}	52.17	**56.85**	57.74	**58.78**	**59.06**	**59.12**	**59.78**	**59.81**	**59.97**	59.47	**59.97**(1700)
Acc2_{cs}	**52.39**	56.79	**57.87**	58.67	58.70	58.48	58.74	59.06	59.45	59.16	59.45(1700)

proposed methods and term frequency-based methods can achieve the best performance, while document frequency-based methods need to utilize more than 15% top features to achieve their best performance, which is less effective.

Then, for the purpose of more detailed study on a small feature set, Table 2 shows the $F1$ measure results when the size of the selected features is 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900. The best result for each feature set is shown in bold. It can be seen that on the $\text{Data}_{\text{BCIII}}$ the performance of the context similarity-based methods is also better than that of their corresponding frequency-based methods. And when the size of the feature set is 1700 (7.4% of the total number of features), CDM_{cs} acquires the highest $F1$ measure value 59.97, which improves the $F1$ measure of the $\text{Model}_{\text{SVM\_poly}}$ when all the features are used (57.12) by 2.85. Hence the context information of terms is helpful for the feature selection in IAC applications.

We notice that there is a significant drop in performance from the $\text{Data}_{\text{BCII}}$ to $\text{Data}_{\text{BCIII}}$, which suffered from the fact that the training article collection is extracted from different online article sources compared with the test data sets, and that the test data sets have the high class skew problem [30].

3.4. Analysis and Discussion

Comparison of the Selected Features. Besides the $F1$ measure results, we also analyze the effectiveness of feature selection methods through studying the profile of the selected features. The sorted lists of the top-10 features picked by each method are given in Tables 3 and 4 on the $\text{Data}_{\text{BCII}}$ and $\text{Data}_{\text{BCIII}}$, respectively. The features that are selected commonly by all the methods are indicated in bold. These common features make the same contribution to the classification performance, such as "interact" in Table 3 and "interaction" in Table 4. Hence we compare the special features selected by different methods. We note that there are two categories of special selected features according to two different feature selection principals. The first category features are the ones

Table 3: The top 10 features on the $Data_{BCII}$ selected by various feature selection methods. The terms that are selected commonly by all the methods are indicated in bold.

Number	GI	DF	CDM	Acc2	TFIDF	$GINI_{NTF}$	GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
1	protein	protein	hybrid	bind	proteins	protein	**interact**	**interact**	bind	**interact**
2	bind	bind	**interact**	interaction	cell	bind	bind	hybrid	**interact**	hybrid
3	interaction	proteins	protein	domain	receptor	domain	hybrid	bind	hybrid	bind
4	domain	cell	cell	complex	cells	proteins	binds	interaction	binds	interaction
5	proteins	**interact**	proteins	**interact**	kinase	**interact**	identified	binds	analyse	binds
6	complex	cells	spots	proteins	domain	cell	analyse	analyse	complex	analyse
7	cell	domain	binds	cell	bind	complex	interaction	domain	interaction	expression
8	terminal	analyse	cells	hybrid	beta	cells	activation	human	activity	human
9	**interact**	complex	spot	protein	protein	kinase	function	activity	domain	activity
10	cells	interaction	domains	interacts	**interact**	receptor	activity	identified	identified	identified

Table 4: The top 10 features on the $Data_{BCIII}$ selected by various feature selection methods. The terms that are selected commonly by all the methods are indicated in bold.

Number	GI	DF	CDM	Acc2	TFIDF	$GINI_{NTF}$	GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
1	protein	protein	interacts	protein	cells	protein	hybrid	interact	binds	binds
2	bind	cell	**interaction**	bind	cell	cells	interact	binds	interact	interact
3	results	cells	interact	**interaction**	expression	cell	binds	study	bind	bind
4	cell	bind	binds	domain	**interaction**	**interaction**	expression	bind	expression	expression
5	cells	results	gene	complex	bind	expression	bind	expression	**interaction**	study
6	study	**interaction**	domain	proteins	protein	proteins	study	**interaction**	activity	**interaction**
7	**interaction**	activity	cell	kinase	proteins	complex	subunit	activity	complex	complex
8	using	proteins	terminal	gene	gene	domain	**interaction**	subunit	domain	activity
9	gene	study	interacting	cell	genes	gene	activity	results	results	terminal
10	use	function	ubiquitin	interacts	human	human	increase	complex	activity	results

selected based on the statistical frequency. These features obtain higher scores because more documents contain them or they occur more. However, the term cooccurrences and chunks within the document are ignored. For example, the terms "protein" and "cell" are selected by all the frequency-based methods but the context similarity-based methods on both $Data_{BCII}$ and $Data_{BCIII}$. Considering "protein," it is just used to describe different protein names, which can appear anywhere in biological articles with the result of high document frequency or term frequency. However, it is not a distinctive feature to classify PPI relevant or nonrelevant articles. If such irrelevant features are assigned higher scores by a feature selection method, the performance obtained by those features would be degraded. On the contrary, these features are assigned lower values by our proposed methods, because their context dissimilarity between the PPI relevant and nonrelevant articles depresses their scores. The second category features are shared by the context similarity-based methods, such as the terms "activate" in Table 3 and "activity" in Table 4. Their evaluation scores are raised by the context similarity within the PPI relevant articles, which is important for the classification purpose.

In order to further study the proposed methods on common and special selected features, the top 1000 features are selected on both data sets, respectively. We perform experiments on the pairs of one context similarity-based method and one frequency-based feature selection method. First, the common features selected for each pair by both feature selection methods are fed into the $Model_{SVM\_poly}$. Then the performance of this $Model_{SVM\_poly}$ based on the common features is compared with the performance achieved based on all the top 1000 features selected by the context similarity-based method and the frequency-based method, respectively. Our purpose is to reveal which kind of feature selection methods can increase the performance more with their special selected features. The results are listed in Tables 5 and 6 on the $Data_{BCII}$ and $Data_{BCIII}$, respectively. It can be seen that the increments of context similarity-based methods are higher than the frequency-based methods, so the special features selected through context similarity-based methods can bring more distinctive information for the classifier on both data sets.

Dimension Reduction Rate. In addition to $F1$ measure, dimension reduction rate is another important aspect of feature selection. Therefore, a dimension reduction is also studied during the experiments. To compute a dimension reduction rate together with the $F1$ measure, a scoring scheme from Gunal and Edizkan [31] is defined as follows:

$$\text{Score} = \frac{1}{k}\sum_{i=1}^{k}\frac{\dim_N}{\dim_i}F1_i, \tag{13}$$

Table 5: The comparison of common and special selected features on the $Data_{BCII}$. C denotes the $F1$ measure of the $Model_{SVM\text{-}poly}$ based on the common features selected by the frequency-based method and the context similarity-based method. The integer in parentheses is the number of the common features; CS denotes the $F1$ measure obtained by the context similarity-based method based on the top 1000 features. The number in parentheses is the increments compared with C; F denotes the $F1$ measure obtained by the frequency-based method based on the top 1000 features. The number in parentheses is the increments compared with C.

		GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
GI	C	71.62(714)	71.49(734)	71.40(702)	71.47(730)
	CS	76.64(+5.02)	76.57(+5.08)	76.81(+5.41)	76.60(+5.13)
	F	71.76(+0.14)	71.76(+0.27)	71.76(+0.36)	71.76(+0.29)
DF	C	72.31(625)	71.23(644)	71.06(613)	71.21(643)
	CS	76.64(+4.33)	76.57(+5.34)	76.81(+5.75)	76.60(+5.39)
	F	72.48(+0.17)	72.48(+1.25)	74.48(+3.42)	72.48(+1.27)
CDM	C	73.54(534)	73.17(552)	72.96(537)	73.16(552)
	CS	76.64(+3.10)	76.57(+3.40)	76.81(+3.85)	76.60(+3.44)
	F	74.92(+1.38)	74.92(+1.75)	74.92(+1.96)	74.92(+1.76)
Acc2	C	72.87(635)	73.40(650)	73.88(643)	73.40(650)
	CS	76.64(+3.77)	76.57(+3.17)	76.81(+2.93)	76.60(+3.20)
	F	74.04(+1.17)	74.04(+0.64)	74.04(+0.16)	74.04(+0.64)
TFIDF	C	72.90(668)	73.49(740)	73.40(692)	73.47(693)
	CS	76.64(+3.74)	76.57(+3.08)	76.81(+3.41)	76.60(+3.13)
	F	74.87(+1.97)	74.87(+1.38)	74.87(+1.47)	74.87(+1.40)
$GINI_{NTF}$	C	73.67(720)	73.77(754)	73.49(710)	73.60(730)
	CS	76.64(+2.97)	76.57(+2.80)	76.81(+3.32)	76.60(+3.00)
	F	75.10(+1.43)	75.10(+1.33)	75.10(+1.61)	75.10(+1.50)

Table 6: The comparison of common and special selected features on the $Data_{BCIII}$. C denotes the $F1$ measure of the $Model_{SVM\text{-}poly}$ based on the common features selected by the frequency-based method and the context similarity-based method. The integer in parentheses is the number of the common features; CS denotes the $F1$ measure obtained by the context similarity-based method based on the top 1000 features. The number in parentheses is the increments compared with C; F denotes the $F1$ measure obtained by the frequency-based method based on the top 1000 features. The number in parentheses is the increments compared with C.

		GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
GI	C	55.22(740)	55.04(781)	55.54(764)	56.37(745)
	CS	58.78(+3.56)	58.76(+3.72)	59.09(+3.55)	58.57(+2.00)
	F	56.38(+1.16)	56.38(+1.34)	56.38(+0.84)	56.38(+0.01)
DF	C	57.02(579)	56.85(593)	56.70(658)	57.20(609)
	CS	58.78(+1.76)	58.76(+1.91)	59.09(+2.39)	58.57(+1.37)
	F	57.61(+0.59)	57.61(+0.76)	57.61(+0.91)	57.61(+0.41)
CDM	C	57.22(544)	57.11(545)	57.50(550)	57.18(560)
	CS	58.78(+1.56)	58.76(+1.65)	59.09(+1.59)	58.57(+1.39)
	F	57.09(−0.13)	57.09(−0.02)	57.09(−0.41)	57.09(−0.09)
Acc2	C	56.17(656)	56.13(678)	57.00(671)	56.51(673)
	CS	58.78(+2.61)	58.76(+2.63)	59.09(+2.09)	58.57(+2.06)
	F	57.09(+0.92)	57.09(+0.96)	57.09(+0.09)	57.09(+0.58)
TFIDF	C	57.20(668)	57.14(701)	57.05(692)	57.09(690)
	CS	58.78(+1.58)	58.76(+1.62)	59.09(+2.04)	58.57(+1.48)
	F	57.80(+0.60)	57.80(+0.66)	57.80(+0.75)	57.80(+0.71)
$GINI_{NTF}$	C	57.35(720)	57.19(754)	57.17(715)	57.30(698)
	CS	58.78(+1.43)	58.76(+1.57)	59.09(+1.92)	58.57(+1.27)
	F	57.96(+0.61)	57.96(+0.77)	57.96(+0.79)	57.96(+0.66)

Table 7: Rate scores of dimension reduction on the $Data_{BCII}$ and $Data_{BCIII}$, respectively.

	GI	DF	CDM	Acc2	TFIDF	$GINI_{NTF}$	GI_{cs}	DF_{cs}	CDM_{cs}	$Acc2_{cs}$
$Data_{BCII}$	4640	4642	4664	4679	4693	4700	4729	4734	4738	4731
$Data_{BCIII}$	2684	2667	2693	2695	2705	2712	2727	2723	2729	2728

where k is the number of trails, $\dim_N$ is the maximum feature size, $\dim_i$ is the feature size at the ith trail, and $F1_i$ is the $F1$ measure of the ith trail. Here, $\dim_i$ is a set of sequences, 100, 300, 500, 700, 900, 1100, 1300, 1500, 1700, and 1900, and so k is 10. The results of dimension reduction analysis using the described scoring scheme are presented in Table 7. It is apparent from this table that the context similarity-based feature selection methods provide better performance than the frequency-based methods.

4. Conclusions

In this paper, novel context similarity-based feature selection methods were introduced for text classification in the biological domain to classify protein interaction articles. They assign importance scores to features based on their similarity measure of context information within certain text categories. Using two different data sets, the performance of the proposed methods was investigated and compared against four document frequency-based and two term frequency-based methods. The effectiveness of the proposed methods was demonstrated and analyzed on the $F1$ measure, the profile of selected features, and dimension reduction rate for the IAC tasks. Since IAC is a binary text classification task in biological domain, we also want to know the performance of the proposed methods when they are extended to multiclass problems. Hence, an adaptation of the context similarity-based selection method to multiclassification problems remains an interesting future task.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors want to thank the anonymous reviewers for their helpful comments and suggestions. This work is supported by the National Natural Science Foundation of China (nos. 61202135 and 61402231), the Natural Science Foundation of Jiangsu Province (nos. BK2012472 and BK2011692), the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (no. 12KJD520005), and the Qing Lan Project.

References

[1] S. Wang, D. Li, X. Song, Y. Wei, and H. Li, "A feature selection method based on improved fisher's discriminant ratio for text sentiment classification," *Expert Systems with Applications*, vol. 38, no. 7, pp. 8696–8702, 2011.

[2] T. S. Guzella and W. M. Caminhas, "A review of machine learning approaches to Spam filtering," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10206–10222, 2009.

[3] B. Zhou, Y. Y. Yao, and J. Lou, "A three-way decision approach to email spam filtering," in *Proceedings of the 23rd Canadian Conference on Artificial Intelligence (Canadain AI '10)*, vol. 6085 of *Lecture Notes in Computer Science*, pp. 28–39, 2010.

[4] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, vol. 8, no. 1, pp. 78–88, 2011.

[5] S. A. Özel, "A web page classification system based on a genetic algorithm using tagged-terms as features," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3407–3415, 2011.

[6] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale Bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[7] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, pp. 412–420, 1997.

[8] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[9] N. Azam and J. T. Yao, "Incorporating game theory in feature selection for text categorization," in *Proceedings of the 13th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC '11)*, vol. 6743 of *Lecture Notes in Computer Sience*, pp. 215–222, Springer, 2011.

[10] H. Liang, J. Wang, and Y. Yao, "User-oriented feature selection for machine learning," *The Computer Journal*, vol. 50, no. 4, pp. 421–434, 2007.

[11] S. Piramuthu, "The protein-protein interaction tasks of biocreative III: evaluating feature selection methods for learning in data mining applications," *European Journal of Operational Research*, vol. 156, no. 2, pp. 483–494, 2004.

[12] Y. Y. Yao and Y. Zhao, "Attribute reduction in decision-theoretic rough set models," *Information Sciences*, vol. 178, no. 17, pp. 3356–3373, 2008.

[13] Y. Y. Yao, Y. Zhao, and J. Wang, "On reduct construction algorithms," *Transactions on Computational Science*, vol. 2, pp. 100–117, 2008.

[14] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[15] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 1–5, 2007.

[16] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432–5435, 2009.

[17] N. Azam and J. T. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," *Expert Systems with Applications*, vol. 39, no. 5, pp. 4760–4768, 2012.

[18] J. Yang, Y. Liu, X. Zhu, Z. Liu, and X. Zhang, "A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," *Information Processing and Management*, vol. 48, no. 4, pp. 741–754, 2012.

[19] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[20] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[21] Y.-Q. Wei, P.-Y. Liu, and Z.-F. Zhu, "A feature selection method based on improved TFIDF," in *Proceedings of the 3rd International Conference on Pervasive Computing and Applications (ICPCA '08)*, pp. 94–97, Alexandria, Egypt, October 2008.

[22] W. E. Winkler, "tring comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage," in *Proceedings of the Section on Survey Research Methods (American Statistical Association)*, vol. 359, pp. 354–359, 1990.

[23] M. A. Jaro, "Advances in record linkage methodology as applied to the 1985 census of Tampa Florida," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 414–420, 1989.

[24] M. A. Jaro, "Probabilistic linkage of large public health data files," *Statistics in Medicine*, vol. 14, no. 5–7, pp. 491–498, 1995.

[25] V. N. Vapnik, *Statistical Learning Theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control, John Wiley & Sons, New York, NY, USA, 1998.

[26] T. Xia and Y. Du, "Improve VSM text classification by title vector based document representation method," in *Proceedings of the 6th International Conference on Computer Science and Education (ICCSE '11)*, pp. 210–213, August 2011.

[27] M. Antunes, C. Silva, B. Ribeiro, and M. Correia, "A hybrid ais-svm ensemble approach for text classification," in *Proceedings of the 10th International Conference on Adaptive and Natural Computing Algorithms*, pp. 342–352, 2011.

[28] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[29] M. Krallinger and A. Valencia, "Evaluating the detection and ranking of protein interaction relevant articles: the biocreative challenge interaction article sub-task (ias)," in *Proceedings of the 2nd BioCreative Challenge EvaluationWorkshop*, pp. 29–39, 2007.

[30] M. Krallinger, M. Vazquez, F. Leitner et al., "The protein-protein interaction tasks of bioCreative III: classification/ranking of articles and linking bio-ontology concepts to full text," *BMC Bioinformatics*, vol. 12, supplement 8, article S3, 2011.

[31] S. Gunal and R. Edizkan, "Subspace based feature selection for pattern recognition," *Information Sciences*, vol. 178, no. 19, pp. 3716–3726, 2008.