様式 7

論　文　内　容　要　旨

報告番号	甲　先　第　176　号	氏　名	Mahmoud Ibrahim Elhosiny Elmarhoumy
学位論文題目	Research on Automatic Classifiers Techniques for Text Classification テキスト分類のための自動分類手法に関する研究		

内容要旨

Text classification is the process of automatically classify texts that provides useful information for the user. Automatic Text Classification is very important task for the management information applications by assignment of automatic texts to one or more predefined categories, it could be used for a lot of management information applications such as an indexing mechanism for text retrieval and a component of an information filtering system. By 90s, the machine learning technique has been used to build the automatic text classifier by using the training set of document, this type of classifier has gained a high popularity due to its advantages: a high accuracy, and, most important, a considerable savings in term of expert labor power. To build the classifier model, traditional methods used set of pre-classified document from training set of document to train the classifier which will be ready to classify any test document. In the last few years the machine learning techniques has led to enhance the performance of classifier model by saving the time and straightforward portability to different domains. To enhance the automatic text classification task, we propose a novel approach for treating the problem of inductive bias incurred by the centroid classifier assumption. This approach is a trainable classifier, which takes into account tfidf as a text feature. The main goal of the proposed approach is to take advantage of the most similar training errors in the classification model for successively updating that model based on a certain threshold. The proposed approach is practical and flexible to implement. The complete performance of the proposed approach is measured at several threshold values on the Reuters-21578 text categorization collection. Experimental results show that the proposed approach can improve the performance of the centroid classifier better than traditional approaches (traditional centroid classifier, support vector machines, decision trees, Bayes nets, and N Bayes) by 1, 1.2, 4.1, 7.5, and 11%, respectively.

Moreover, we present a new hybrid Center Profile Vector (CPV) classification model based on the modified N-gram and centroid classifier models. The hybrid model exploits a new approach to calculate the term weight based on *tfsc, dfsc* in order to sort the terms in the model profile. Moreover, we present a new distance similarity method for the N-gram model that solves the problem of the difference in representation lengths among classes and documents. The hybrid (CPV) classification model provides promising classification rate compared with some other models such as N-gram Model, Centroid Classifier, Naive Bays and Support Vector Machine.

In this thesis, we investigate different approaches for automatic text classification. Firstly, we exploit the Naive Bays (NB) and Support Vector Machine (SVM) classifiers based on tfidf to weigh document's terms. Moreover, we exploit N-gram (traditional N-gram model with tfidf and Manhattan distance), DN-g (N-gram model with tfidf and new distance measure), TC (traditional centroid classifier), AC (average

centroid classifier) and NC (normalized centroid classifier), as baseline models as well. Secondly, we propose the modified N-gram model using the proposed tfsc,dfsc score to rank the profiles in N-gram. Finally we employ different term weighting schemes to establish the hybrid CPV model based on centroid classifiers and the modified N-gram to improve the classification performance. As the results show, the hybrid CPV classification model enhances the classification performance.

The proposed term weighting schemes and the proposed similarity distance measure have good effect on the classification performance for N-gram model. A new distance similarity measure is applied for the modified N-gram model to calculate the distance similarity. This new distance similarity measure solves the problem of the difference in profiles lengths among the classes and between the document and the class.

The proposed CPV model will gain the benefits of the centroid model and the modified N-gram model to achieve high classification rate as will be shown in the experimental result. Moreover, the proposed approach uses two weighting approaches at the same time, which makes the proposed PCV model able to gain the benefits of two weighting approaches simultaneously.

様式９

論文審査の結果の要旨

報告番号	甲先 乙先　第　176　号 工修	氏　　名	Mahmoud Ibrahim Elhosiny Elmarhoumy
審査委員	主査　寺田　賢治 副査　獅々堀　正幹 副査　任　福継		

学位論文題目

Research on Automatic Classifiers Techniques for Text Classification

（テキスト分類のための自動分類手法に関する研究）

審査結果の要旨

本論文では，自動テキスト分類を目的として，セントロイド分類過程において発生する誘導バイアス問題を解決するための新しい手法を研究する。本提案手法は，テキスト特徴として TFIDF を考慮するトレーニング可能な分類器によって実現する。提案手法の主な特長はしきい値に基づくモデルを連続的に更新するために，分類モデルにおいて類似度の最も高いトレーニング誤差を利用することである。これにより，提案手法を実践的かつ柔軟に実装が可能である。提案手法の性能評価は，ロイター−21578 テキスト分類コレクションを用いていくつかの閾値により実験を行った。この実験結果より，提案手法が従来手法（既存のセントロイド分類器、サポートベクタマシン、デシジョン・ツリー、ベイズ・ネット，および N ベイズ，各々1、1.2、4.1、7.5 および 11%)に比べ，セントロイド分類器がより良い性能改善が可能であることを確認した。

また，改良型 N-グラムおよびセントロイド分類器モデルに基づく新しいハイブリッド・センタ・プロファイル・ベクトル(CPV)分類モデルを提案した。ハイブリッドモデルにおいて、モデル・プロファイル中の用語をソートするために，*tfsc,dfsc* に基づく重み付け用語の計算を行う新しい手法を開発した。クラスとドキュメント間の表現長の差の問題を解決する N-グラム・モデルに新しい距離類似法を導入した。ハイブリッド CPV 分類モデルは，N-グラム・モデル、セントロイド分類器、ナイーブベイズ（NB），およびサポートベクタマシンの他のいくつかのモデルと比較して，保証された分類レートが得られる特長がある。

実験結果より，提案した CPV モデルが高い分類レートを達成し，さらにセントロイドモデルおよび改良型 N-グラム・モデルの優位性を示した。2 つの重み付け手法を同時に用いることによって，提案する CPV 分類モデルが，2 つの重み付け手法の利点を同時に獲得できることを確証した。

以上本研究は，当該分野の既存の問題を解決した貢献から価値のある研究であり，本論文は学位論文としての水準を満たし，博士（工学）の学位授与に値するものと判定する。