様式７

論文内容要旨

報告番号	甲 先 第 157 号	氏 名	Mohammad Golam Sohrab
学位論文題目	A Foundation of Class-Indexing and Class-Semantic-Indexing based Term Weighting Approaches for Automatic Text Classification テキスト自動分類のクラスインデックスとクラスセマンティックインデックスに基づく用語重み付けアプローチ		

内容要旨

Most of the previous studies related on different term weighting emphasize on the document-indexing-based and four fundamental information elements-based approaches to address automatic text classification (ATC). In this study, we introduce class-indexing-based term-weighting approaches and judge their effects in high-dimensional and comparatively low-dimensional vector space over the TF.IDF and five other different term weighting approaches that are considered as the baseline approaches.

First, we implement a class-indexing-based TF.IDF.ICF observational term weighting approach in which the inverse class frequency (ICF) is incorporated. In the experiment, we investigate the effects of TF.IDF.ICF over the Reuters-21578, 20 Newsgroups, and RCV1-v2 datasets as benchmark collections, which provide positive discrimination on rare terms in the vector space and biased against frequent terms in the text classification (TC) task. Therefore, we revised the ICF function and implemented a new inverse class space density frequency (ICSδF), and generated the TF.IDF.ICSδF method that provides a positive discrimination on infrequent and frequent terms. We present detailed evaluation of each category for the three datasets with term weighting approaches. The experimental results show that the proposed class-indexing-based TF.IDF.ICS$_{\delta}$F term weighting approach is promising over the compared well-known baseline term weighting approaches. In this approach we present a brief introduction of building essential knowledge on class-indexing and state-of-the-art of prototype class-indexing systems. We therefore introduce of proposed class-indexing based term weighting scheme with machine learning technique which is used in to extract weight from data automatically. A simple overview of the proposed class-indexing is incorporated with term index, document index, and class index. Therefore the combination of term and document index or the combination of term, document and class index, we can generate TF.IDF TF.ICF, TF.IDF.ICF, TF.ICSδF, and TF.IDF.ICSδF term weighting scheme respectively.

Secondly, we implement a combined term weighting schemes which are integrated with 10 different term weighting approaches and generate five (CTWS-Sum, CTWS-Avg, CTWS-MR, CTWS-GA, and CTWS-FFNN) new weighting methods to enhance automatic text classification. In the experiment, we investigate the effects of five different methods on conventional approaches and previous proposed approaches over the Reuters-21578, and 20 Newsgroups datasets. The experimental results show that the proposed CTWS-Sum, considering a very simple method is showing promising results on classification task.

Finally, we introduce a class-semantic-indexing, where semantic term weighting scheme is another criterion of weighting a term. In this approach, we integrate the dictionary based semantic indexing with corpus based class-indexing method. Measuring the semantic indexing of concepts is an intriguing problem in Natural Language Processing. Various approaches that attempt to approximate human judgment of semantic indexing, have been tried by researchers. Semantic analysis method utilizes lexical hierarchies in the language dictionary or co-occurrence patterns of words in a large corpus of texts to decide the semantic similarity or semantic association between unknown words. Therefore, A simple overview of the proposed class-semantic-indexing is discussed in this work, where the proposed class-semantic-indexing is incorporated with term, and document and class index with corpus-based and Wordnet-dictionary-based combinational approaches. The nodes of term, document and class index contain two fields: data and link. While provide the dataset as an input to assign scores for lexical terms, the data field of term index contains a term with two different weights. One is statistical corpus based term weight and another is Wordnet dictionary based term weight. To compute a certain term from a certain corpus from different hierarchy that is discussed in this work. In the Wordnet dictionary based approach, first we determine the top rank sense from a set of senses for a certain term. We therefore, compute the weight for that sense through document and class index; and combine the weight with class-indexing based term weight and get a new class-semantic-weight for classification task.

様式９

論文審査の結果の要旨

報告番号	[✓ 甲先] 乙先　第　１５７　号 工修	氏　　名	Mohammad Golam Sohrab
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

A Foundation of Class-Indexing and Class-Semantic-Indexing based Term Weighting Approaches for Automatic Text Classification
（テキスト自動分類のクラスインデックスとクラスセマンティックインデックスに基づく用語重み付けアプローチ）

審査結果の要旨

本論文では，テキスト自動分類のクラスインデックスとクラスセマンティックインデックスに基づく用語重み付けアプローチを提案し，自動分類の様々な手法の比較を行い，テキストマイニングにおける自動分類の高精度化に関する研究を行った。この研究では，従来のクラスインデックスの上に，クラスセマンティックインデックスを導入することにより，テキスト自動分類のパフォーマンスを改善することが期待されている。即ち，TF.IDF のみならず，TF.IDF.ICF を活かし，テキストのクラスのセマンティック特性を活かす提案である。

最初に，クラスセマンティックインデックスアプローチを提案した。これは逆のクラス頻度(ICF)が組み入れられるクラス指標付け基づいた TF.IDF.ICF の観測上の用語に重みを加えるアプローチである。Reuters-21578,20 のニュースグループ及び RCV1-v2 データセット上の TF.IDF.ICF の影響を考察し，提案したアプローチの有効性を検証した。

次に，クラスインデックスとクラスセマンティックインデックスに基づく用語重み付けアプローチを提案し，幾つかの統計的な予備実験を実施することにより，幾つかのパラメタ―を決定した。

最後に，提案したアプローチと従来の代表的なアプローチとの比較実験を行ったが，機械的な学習方法により，提案したテキスト自動分類手法が従来のものより優れたパフォーマンスができた。

提案された各手法に基づいて，実験システムを構築し，様々な評価実験を行ったが，提案された手法の有効性を確かめることができた。

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。