情 報 学 基 礎 25－3
（1992. 5. 12）

疑似シソーラスの自動構築への機械学習アプローチ

ビジェイ V．ラガバン

南西ルイジアナ大学高等計算機研究センター

ユーザからのフィードバックを利用してシソーラスを自動的に生成する処理は疑似シソーラス構築と呼ばれている。疑似シソーラス構築処理を定式化し、疑似シソーラス生成の厳密な公式を導出した。疑似シソーラスの構築が機械学習の問題として扱えるように疑似シソーラス生成式を一次構造に変換する方法を考案した。パーセプトロン基準関数に基づく疑似シソーラス構築のための学習アルゴリズムを開発した。学習によって生成された疑似シソーラスは、学習データの集合の中で適合する文書・検索式の対に正の取扱い値を、適合しない文書・検索式の対には負の取扱い値を付与するという意味において、最適であることが保証される。疑似シソーラスを用いた一次実験の結果では、検索精度で10％から47％の改善が確認された。

A MACHINE LEARNING APPROACH TO AUTOMATIC PSEUDO-THESAURUS CONSTRUCTION

Vijay V. Ragnavan

The Center for Advanced Computer Studies
University of Southwestern Louisiana

P.O. Box 44330, Lafayette, LA 70504-4330, USA

In information retrieval it is common to use a term thesaurus as a language normalization mechanism. In recent years there have been some attempts to generate a thesaurus automatically. The process of generating a thesaurus automatically by means of user feedback is referred to as pseudo-thesaurus construction. The problem of pseudo-thesaurus construction is formally specified and rigorous mathematical equation for generating a pseodo-thesaurus is derived. A way to transform the pseudo-thesaurus equation into a suitable linear structure in order that it can be studied as a problem in machine learning is advanced. A learning algorithm that enables pseudo-thesaurus construction from user feedback based on the perceptron criterion function is developed. The method of obtaining the pseudo-thesaurus is tightly coupled with the way in which the resultant thesaurus is to be incorporated into the retrieval process. The pseudo-thesaurus generated is guaranteed to be optimal in the sense that it always assigns a positive "treatment value" to a relevant(document, query)pair and a negative treatment value to a nonrelevant (document, query) pair in a set of training pairs. The preliminary experimental results show that this method, which produces improvements in retrieval performance ranging from 10% to 47% for the collection tested, is promising and warrants further investigation.