ABSTRACT

A NEW CLUSTERING SCHEME
AND
ITS USE IN AN INFORMATION RETRIEVAL SYSTEM
INCORPORATING THE SUPPORT OF A DATABASE MACHINE

CAN, Fazlı
Ph.D. in Computer Engineering
Supervisor: Assoc.Prof.Dr. Esen A. ÖZKARAHAN
Nov. 1984, 288 pages

The need for immediate and accurate access to the current literature at one side and the information explosion on the other side have caused the development of information retrieval systems. in this work, information retrieval problem is studied and new concepts and methodologies are proposed for its solution.

The new proposals are cover coefficient and cluster seed power concepts and the methodologies for estimating the number of clusters within a collection and the number of members within a cluster. These concepts and methodologies are used in a new single-pass clustering algorithm. A multi-pass clustering algorithm is introduced to show the validity of the cover coefficient concept for clustering purposes. in the thesis, the complexity analysis of the algorithms, a new centroid generation policy in connection with the new cover coefficient concept are presented. An algorithm for the maintenance of the clusters in expanding document collection environments and its complexity analysis are also presented.

The similarity and stability concepts for clustering algorithms are introduced, then the clustering algorithms are analyzed by a set of experiments with respect to these concepts. For the purpose of the experiments, a document collection of 167 articles from the ACM-TODS publications has been constructed. The characteristics of the collection, the findings of the experiments and some observed basic relationships are illustrated in detail.

In the thesis, an information system model which integrates the information retrieval and database management systems is proposed. Unlike the previous studies aimed at this purpose, which more or less reduce one system into the other, the proposed model aims to accomplish this integration by a synthesis of the techniques and methodologies of both systems. For this purpose, a database machine, the Relational Associative Processor (RAP), is enhanced with the new text retrieval instructions. Context sensitive free text retrieval operations are implemented by using the new instructions. in the model, a clustering subsystem and a conceptual data model are used for information retrieval purposes. The performance of the database machine in text retrieval operations and a comparative performance evaluation of the single-pass and the multi-pass clustering algorithms in information retrieval are presented.

Additional concepts/methodologies that utilize cover coefficient concepts are also introduced in the thesis.

Keywords: information (text,document) retrieval systems, cover coefficient, cluster seed power,clustering algorithms, complexity analysis of algorithms, similarity and stability of clustering algorithms, centroid generation, database management systems, text retrieval computers, RAP database machine, term discrimination value, document significance value.