Diss. ETH Nr. 8446

Vektorielles Rechnen im Information Retrieval

Abhandlung
zur Erlangung des Titels eines
Doktors der Technischen Wissenschaften
der
EIDGENÖSSISCHEN TECHNISCHEN HOCHSCHULE

vorgelegt von
Jürg Grossmann
Dipl. Math. ETH
geboren am 19. Juli 1958
von Zürich

Angenommen auf Antrag von
Prof. Dr. C.A. Zehnder, Referent
Prof. Dr. K. Bauknecht, Korreferent

Zusammenfassung

Die Speicherung und Wiedergewinnung grosser Mengen von unformatierter Information, wie sie für die Anwendungen des Information Retrievals typisch sind, stellt sehr hohe Ansprüche an die Kapazität von Speichermedien und an die Leistungsfähigkeit von Rechnern und Softwaresystemen. Leistungssteigerungen erreicht man durch die Parallelisierung von Verarbeitungsschritten auf verschiedenen Ebenen.

In dieser Arbeit werden zuerst Methoden und Modelle des Information Retrievals diskutiert. Besondere Aufmerksamkeit gilt dabei dem Systemaufbau und dem Kontroll- und Datenfluss dieser Modelle, was für die Beurteilung der Parallelisierungmöglichkeiten wichtig ist.

Die Betrachtung verschiedener paralleler Rechnerarchitekturen und unterschiedlicher Ansätze für parallele Verarbeitung führt dann zu einer Darstellung der Möglichkeiten, die sich für die Parallelisierung der Operationen in Retrieval-Systemen ergeben. Es zeigen sich dabei zwei Stossrichtungen. Auf der einen Seite werden grosse Anstrengungen unternommen, Geräte zu entwickeln, welche den besonderen Anforderungen von Retrieval-Operationen angepasst sind und die Parallelität dieser Operationen ausnützen. Diese Ansätze haben den Nachteil, dass die Entwicklung der Komponenten sehr teuer ist und ihr Einsatzbereich wegen der Spezialisierung sehr beschränkt ist. Die andere Stossrichtung geht dahin, bestehende parallele Architekturen, welche für Allzweckrechner konzipiert sind, für die Zwecke des Information Retrievals einzusetzen. Bei dieser Lösung ist offen, wie die parallele Architektur am besten für Information Retrieval zu nutzen ist.

Für sogenannte Vektorrechner, einer Klasse von Rechnern mit paralleler Verarbeitung auf der Ebene der Instruktionen, werden dann konkrete Möglichkeiten der Anpassung verschiedener Algorithmen gezeigt. Zuerst wird auf die Architektur und die Funktionsweise dieser Rechner eingegangen, um daraus die Voraussetzungen für eine erfolgreiche Vektorisierung, das heisst Formulierung und Anpassung von Algorithmen für Vektorrechner, abzuleiten. Diese allgemeinen Aussagen werden dann übertragen auf Algorithmen aus drei verschiedenen Teilgebieten des Information Retrievals, nämlich aus dem eigentlichen Retrieval, aus der automatischen Indexierung und aus der automatischen Klassifizierung. Die Überlegungen lassen erwarten, dass die untersuchten Algorithmen für eine Verarbeitung auf Vektorrechnern geeignet sind.

Für die Überprüfung dieser Aussage wurde ein Retrieval-System implementiert, welches als Testbett für die verschiedenen Algorithmen und ihren Varianten diente. Die Experimente mit Sammlungen von 1000 bis 10'000 Dokumenten bestätigen die gemachten Aussagen und belegen, dass nach entsprechender Anpassung der Algorithmen Vektorverarbeitung für diese Anwendungen lohnenswert sein kann.

Abstract

The storage and retrieval of large quantities of unformatted data are typical for information retrieval applications. They make high demands not only on the capacity of storage media but also on the efficiency of hardware and software systems. One way to increase the efficiency is the parallelisation of independent operations.

This thesis starts out with a discussion of several methods and models of information retrieval. Special attention is payed to the overall system architectures and the data and control flow of these models since they are important for feasibility studies on parallelisation.

A discussion of different approaches to parallel processing in general concludes that there are two major possibilities to employ parallel processing in information retrieval: the first is to construct special equipment which is tuned to the special requirements of retrieval operations and their inherent parallelism. The problem with this solution is its expensivness and the small range of applications the newly developed equipment may be used for. The other possibility is to implement information retrieval applications on existing general purpose machines with parallel architectures. The problem with this solution is how the parallel architecture can best be exploited for information retrieval applications.

The thesis proceeds by showing some practical ways of doing this on a special class of parallel computers, the vector computers: first the architecture and the principles of operation of these machines are described. From there a set of guidelines for a sucessfull vectorisation are deduced, i.e. implementing new and modifying old algorithms for vector computers. These guidelines are then applied to three types of algorithms for information retrieval. The algorithms are taken from three areas of information retrieval, namely document retrieval, automatic indexing and clustering. The discussion shows that the algorithms are very well suited for processing on vector processors.

In order to assert this theoretical statement, a retrieval system was implemented that served as a test bed for the different algorithms. The experiments confirmed the applicability of vector processing to this kind of application. However, good results are only obtained when vector processors do not only vectorize simple arithmetic operation, but also complex ones like masked operations for if-statements or indexed array access.