

Efficient Query Evaluation using Hybrid Index Organization

by

Zhou, Yingjie

Master Degree of Software Engineering

2011



**Faculty of Science and Technology
University of Macau**

University of Macau

Abstract

**EFFICIENT QUERY EVALUATION USING
HYBRID INDEX ORGANIZATION**

by

Zhou, Yingjie

Thesis Supervisor: Prof. Gong, Zhiguo

Co-Supervisor: Dr. U, Leong Hou

Master Degree Program of Software Engineering

Recently, a large amount of attention has been paid to efficient document retrieval from a gigantic data collection, such as in web search areas. For a keywords-search, the *top-K* computation scans the documents from multiple involved inverted lists and the computation should be stopped when there is no other unseen documents better than the *top-K* documents being seen so far, instead of traversing the whole lists. In this thesis, we give a comprehensive study and analyze the pros and cons of the state-of-the-art indexing structures and the *top-K* document retrieval techniques. Furthermore, we propose a new *top-K* evaluation framework from the following aspects.

1. *Reorganizing inverted index*– we study a new indexing structure such that more promising documents appear to the beginning of an inverted index.
2. *Refining execution strategies* – we study a new processing strategy such that a faster evaluation can be achieved.

The superior of our proposed techniques has been demonstrated by a thorough experimental evaluation which compares our proposed techniques to the state-of-the-art approaches in the final section. It turns out that, on average, the

response time of our methodologies is about **80%** less than that of the reviewed approaches.



TABLE OF CONTENTS

List of Figures	iii
List of Tables.....	v
LIST of Abbreviations	vii
PREFACE	ix
Chapter1 Introduction.....	1
1.1 <i>Top-K</i> Document Retrieval.....	1
1.2 Challenges and Solutions	2
1.3 Overview of Contents	3
Chapter2 Basic Concepts.....	5
2.1 Ranking Functions	5
Chapter3 Literature Review	7
3.1 Merge-sort on Document-sorted Index Organization	8
3.2 Evaluation Approaches on Score-sorted Index Organization	12
3.2.1 Threshold Algorithm	13
3.2.2 No Random Access Algorithm	15
3.2.3 Lattice-Based Rank Aggregation	18
3.2.4 Impact-based Pruning Method	23
3.3 Revisiting Globally Sorted Index	28
Chapter4 Hybrid Method.....	32
4.1 Problem analysis	32
4.2 our approaches	33
4.2.1 New Inverted Index Construction	33
➤ Equal Partition.....	33
➤ Deviation-based Partition	34
4.2.2 Retrieval Strategies.....	37
➤ Hybrid-I Method	37
➤ Hybrid-II Method	43
Chapter5 Experimental Results.....	47
5.1 Experimental Setup	47

5.2 Experimental Results	48
➤ Tuning Parameters of Partition Methods.....	48
➤ Partition Method Comparison.....	49
➤ Retrieval Approaches Comparison	50
Chapter6 Conclusions and Future Work.....	56
Bibliography.....	58

