

Accepted Manuscript

An efficient method to evaluate intersections on big data sets

Yangjun Chen, Weixin Shen

PII: S0304-3975(16)30352-8
DOI: <http://dx.doi.org/10.1016/j.tcs.2016.07.018>
Reference: TCS 10866

To appear in: *Theoretical Computer Science*

Received date: 25 July 2015
Revised date: 14 July 2016
Accepted date: 15 July 2016

Please cite this article in press as: Y. Chen, W. Shen, An efficient method to evaluate intersections on big data sets, *Theoret. Comput. Sci.* (2016), <http://dx.doi.org/10.1016/j.tcs.2016.07.018>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.



An Efficient Method to Evaluate Intersections on Big Data Sets

YANGJUN CHEN and WEIXIN SHEN

The University of Winnipeg

Set intersections are important in computer science. Especially, intersection of inverted lists is a fundamental operation in information retrieval for text databases and Web search engines. In this paper, we discuss an efficient and effective way to implement this operation in the context of very big data sets. The main idea behind it is to do binary search over sorted interval sequences, each of which corresponds to an inverted list and is constructed by establishing a trie over the sequences of set identifiers as well as a kind of tree encoding, by which each node in the trie is assigned an interval. In many cases, an interval sequence is much shorter than its corresponding inverted list. In particular, the lowest common ancestors of intervals in a trie can be utilized to control a binary search to skip over useless interval containment checks, which enables us to reach an optimal off-line algorithm to do the task, and is theoretically better than any traditional on-line methods (at cost of more space). Experiments have been conducted, showing that the trade-off of space for time is worthwhile.

Categories and Subject Descriptors: F.2.2 [Analysis of algorithms and Problem Complexity]: Non-numerical Algorithms and Problems *Pattern matching; computation on discrete structures*

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Set intersection, inverted files, interval sequences, search engines.

1. INTRODUCTION

In mathematics, the *intersection* $A \cap B$ of two sets A and B is the set that contains all elements of A that also belong to B . In practice, however, the problem is typically related to a collection of sets $S = \{S_1, S_2, \dots, S_M\}$ and we are often asked to evaluate the intersection over a sub-collection of S :

$$S_{i_1} \cap S_{i_2} \cap \dots, S_{i_m}$$

for some $m \leq M$.

This is a key operation in information retrieval, especially for Web search engines and text databases, by which each S_i ($i \in \{1, \dots, M\}$) is a subset of document identifiers containing a certain word, called an *inverted list*. Then, to find all the documents containing a set of words $w_1, \dots, w_k$, a set intersection like the above over all the inverted lists associated with these words needs to be conducted.

This work is supported by NSERC Canada. This is a modification and extension of two papers respectively published in *Int. Conf. on Advances in Big Data Analytics, IEEE*, July 21-24, 2014, USA [19]; and the *11th Int. Conf. on Foundations of Computer Science, IEEE*, July 26-30, 2015, USA [20].

Author's addresses: Dept. of Applied Computer Science, University of Winnipeg, 515 Portage Ave. Winnipeg, Manitoba, Canada R3B 2E9.

Permission to make digital or hardcopies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credits permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 2015 ACM 1539-9087/2010/03-ART39 \$15.00

Download English Version:

<https://daneshyari.com/en/article/4952469>

Download Persian Version:

<https://daneshyari.com/article/4952469>

[Daneshyari.com](https://daneshyari.com)