http://mail-archives.apache.org/mod\_mbox/lucene-dev/200110.mbox/%3c3BD0944F.7060305@earthlink.net%3e

From lucene-dev-return-265-apmail-jakarta-lucene-dev-archive=jakarta.apache.org@jakarta.apache.org Fri Oct 19 20:59:51 2001

Return-Path: <lucene-dev-return-265-apmail-jakarta-lucene-dev-archive=jakarta.apache.org@jakarta.apache.org>

Delivered-To: apmail-jakarta-lucene-dev-archive@jakarta.apache.org

Received: (qmail 1335 invoked by uid 500); 19 Oct 2001 20:59:50 -0000

Mailing-List: contact lucene-dev-help@jakarta.apache.org; run by ezmlm

Precedence: bulk

Reply-To: lucene-dev@jakarta.apache.org

list-help: <mailto:lucene-dev-help@jakarta.apache.org>

list-unsubscribe: <mailto:lucene-dev-unsubscribe@jakarta.apache.org>

list-post: <mailto:lucene-dev@jakarta.apache.org>

Delivered-To: mailing list lucene-dev@jakarta.apache.org

Received: (qmail 1324 invoked from network); 19 Oct 2001 20:59:50 -0000

Message-ID: <3BD0944F.7060305@earthlink.net>

Date: Fri, 19 Oct 2001 14:59:59 -0600

From: Dmitry Serebrennikov <dmitrys@earthlink.net>

User-Agent: Mozilla/5.0 (Windows; U; Windows NT 5.0; en-US; rv:0.9.2) Gecko/20010726 Netscape6/6.1

X-Accept-Language: en-us

MIME-Version: 1.0

To: lucene-dev@jakarta.apache.org

Subject: Re: Getting word count

References: <4BC270C6AB8AD411AD0B00B0D0493DF0EE7C90@mail.grandcentral.com>

Content-Type: multipart/alternative;

boundary="------------050805010407080406030808"

X-Spam-Rating: daedalus.apache.org 1.6.2 0/1000/N

--------------050805010407080406030808

Content-Type: text/plain; charset=us-ascii; format=flowed

Content-Transfer-Encoding: 7bit

>

>

>>You cannot simply count the number of times the method

>>collect() is called on your collector because some queries

>>may result in

>>the same document being selected more than once and so you'd

>>end up with

>>a double-count. (Can anyone confirm that this is the case?)

>>

>

>It should not be the case. The collect() method should be called at most

>once per document.

>

>Doug

>

This is a good news! This would make counting that much more efficient.

My main concern was in the BooleanScorer, and I just verified that I was

worried needlessly - it maintains its own hashtable to avoid double

counting. On a related issue, are there any guarantees about the order

of document numbers in the calls to collect()?

--------------050805010407080406030808--