http://mail-archives.apache.org/mod\_mbox/lucene-dev/200110.mbox/%3c20011016013336.90069.qmail@web20901.mail.yahoo.com%3e

From davekkw@yahoo.com Tue Oct 16 01:33:27 2001

Return-Path: <davekkw@yahoo.com>

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

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

Received: (qmail 79207 invoked from network); 16 Oct 2001 01:33:27 -0000

Received: from web20901.mail.yahoo.com (216.136.226.223)

by daedalus.apache.org with SMTP; 16 Oct 2001 01:33:27 -0000

Message-ID: <20011016013336.90069.qmail@web20901.mail.yahoo.com>

Received: from [165.21.122.230] by web20901.mail.yahoo.com via HTTP; Mon, 15 Oct 2001 18:33:36 PDT

Date: Mon, 15 Oct 2001 18:33:36 -0700 (PDT)

From: Dave Kor <davekkw@yahoo.com>

Subject: Re: TermVector retrieval implementation questions

To: lucene-dev@jakarta.apache.org

In-Reply-To: <3BCB257A.90807@earthlink.net>

MIME-Version: 1.0

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

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

--- Dmitry Serebrennikov <dmitrys@earthlink.net>

wrote:

> >

> >

> >That's something new. Unindexed fields such as

> keyword

> >fields won't have term ids? I hope you can clarify

> >further...

> >

> I believe keywords are indexed, just not tokenized.

> So the entire field

> is treated as a single term.

> This is typically used for storing fields like

> "price" or "id" or

> what-not that is more of a typical database-style

> one field - one value

> situation.

Okay... I seems that I have forgotten my lucene

terminology, I kept thinking indexed == tokenized.

That resulted in me saying that unindexed fields would

also be in tvs when I was actually referring to

untokenized fields. My apologies.

> >Hmm.. will there be a way we can convert/add

> >vectorization to the old segments? The users may

> want

> >some kind of migration path to the new format other

> >than reindexing the entire index.

> >

> Yes and no....

Like Doug, my schedule is a little too tight to

properly think it through right now so I'll reply

again on friday as promised earlier.

> Yes, I know. Me too. Interestingly enough, indexing

> seems to be

> completely IO-bound. I was watching CPU monitor last

> night as I was

> running some simple indexing and CPU never hit

> higher then 5%

> utilization. I didn't have a chance to compare this

> to a previous

> version yet. Does anyone know if this is expected

> behavior or is it

> because I managed to break something?

I once did a test of generating 5 million documents of

random lengths with randomly choosen words from a set

of 25000 elements (pre-loaded into memory).

Indexing took about 5+ hours and I too noticed that

most of the time, CPU usage was extremely low. Disk

activity pretty much took up the most time, especially

during segment merging which seem to happen

periodically at roughly one minute intervals.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Do You Yahoo!?

Make a great connection at Yahoo! Personals.

http://personals.yahoo.com