<http://mail-archives.apache.org/mod_mbox/lucene-dev/200110.mbox/%3C4BC270C6AB8AD411AD0B00B0D0493DF0EE7C56@mail.grandcentral.com%3E>

From DCutting@grandcentral.com Wed Oct 10 17:29:32 2001

Return-Path: <DCutting@grandcentral.com>

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

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

Received: (qmail 55423 invoked from network); 10 Oct 2001 17:29:32 -0000

Received: from unknown (HELO mta.12.com) (65.198.8.41)

by daedalus.apache.org with SMTP; 10 Oct 2001 17:29:32 -0000

Received: (qmail 24705 invoked from network); 10 Oct 2001 17:26:58 -0000

Received: from unknown (HELO riker.grandcentral.com) (10.102.15.55)

by mta.12.com with SMTP; 10 Oct 2001 17:26:58 -0000

Received: by mail.grandcentral.com with Internet Mail Service (5.5.2653.19)

id <42Y1H39A>; Wed, 10 Oct 2001 10:19:03 -0700

Message-ID: <4BC270C6AB8AD411AD0B00B0D0493DF0EE7C56@mail.grandcentral.com>

From: Doug Cutting <DCutting@grandcentral.com>

To: "'lucene-dev@jakarta.apache.org'" <lucene-dev@jakarta.apache.org>

Subject: RE: Token retrieval question

Date: Wed, 10 Oct 2001 10:19:02 -0700

MIME-Version: 1.0

X-Mailer: Internet Mail Service (5.5.2653.19)

Content-Type: text/plain;

charset="iso-8859-1"

Content-Transfer-Encoding: quoted-printable

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

Right now, Lucene does not have good support for what you're doing. =

Lucene

as it stands is designed to support basic search, not other statistical =

text

processing. However there are two features that I would like to add to

Lucene that would help you.

1. Seekable TermDocs.

This would let you efficiently skip forward in a TermDocs to a =

particular

document number. This would enable some search optimizations. This

requires no API changes, as the TermDocs.skipTo() method already =

exists.

2. Stored Document Vectors

These would enable one to determine the set of terms in a document. =

This

would be useful for, e.g. document clustering.

This would add an IndexReader two methods:

public TermFreqVector getTermFreqVector(int docNumber);

public Term getTerm(int termNumber);

The TermFreqVector class would be defined something like:

public class TermFreqVector {

public int[] getTermNumbers();

public int[] getTermFrequencies();

}

The term number array would be sorted. The frequency of the term =

numbered

getTermNumbers()[i] is getTermFrequencies()[i].

Another class that would be useful is something like:

public class TermWeightVector {

public int[] getTermNumbers();

public float[] getTermWeights();

public void add(TermWeightVector other);

public float distance(TermWeightVector other);

}

Both of these are long-term changes, so it may be a while before they =

are

completed. That said, I would like to implement them, when I have =

time!

Doug

> -----Original Message-----

> From: Nestel, Frank [mailto:frank.nestel@coi.de]

> Sent: Wednesday, October 10, 2001 12:23 AM

> To: 'lucene-dev@jakarta.apache.org'

> Subject: Token retrieval question

>=20

>=20

>=20

> Hi,

>=20

> I've been reading the API and I couldn't figure out a

> nice and fast way to solve the following problem:

>=20

> I'd like to enumerate the tokens of a document (or=20

> document field). Do the internal datastructures

> of lucene allow such kind of traversal which is (as

> I understand) of course orthogonal to the access lucene=20

> is optimized for?=20

>=20

> More concrete I have s.th. like 20-50 tokens/words and one

> document and I'd like to ask the document if (and how often)

> it contains those particular tokens. The idea was to augment

> search results with (kind of I know) automatic query

> dependand keywords.

>=20

> The only way I see right now is to create 20-50 TermEnums

> and walk through them until I end up in my document or

> nowhere? Which is probably not feasible for a search result

> page with (say) 20 hits in a larger index.

>=20

> Any (more elegant) chance, I missed?

>=20

> Thank you,

> Frank

>=20

> --

> Dr. Frank Sven Nestel

> Principal Software Engineer

>=20

> COI GmbH Erlanger Stra=DFe 62, D-91074 Herzogenaurach

> Phone +49 (0) 9132 82 4611=20

> http://www.coi.de, mailto:Frank.Nestel@coi.de

> COI - Solutions for Documents

>=20