http://mail-archives.apache.org/mod\_mbox/lucene-dev/200110.mbox/%3C4BC270C6AB8AD411AD0B00B0D0493DF0EE7C40@mail.grandcentral.com%3E

From DCutting@grandcentral.com Mon Oct 8 17:36:21 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 23053 invoked from network); 8 Oct 2001 17:36:21 -0000

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

by daedalus.apache.org with SMTP; 8 Oct 2001 17:36:21 -0000

Received: (qmail 14225 invoked from network); 8 Oct 2001 17:33:45 -0000

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

by mta.12.com with SMTP; 8 Oct 2001 17:33:45 -0000

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

id <42Y1H2JS>; Mon, 8 Oct 2001 10:25:55 -0700

Message-ID: <4BC270C6AB8AD411AD0B00B0D0493DF0EE7C40@mail.grandcentral.com>

From: Doug Cutting <DCutting@grandcentral.com>

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

Subject: RE: context and hit positions with Lucene

Date: Mon, 8 Oct 2001 10:25:46 -0700

MIME-Version: 1.0

X-Mailer: Internet Mail Service (5.5.2653.19)

Content-Type: text/plain;

charset="iso-8859-1"

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

> From: Lee Mallabone [mailto:lee@grantadesign.com]

>

> > The

> > index does not store the byte-position of words in the

> original document.

>

> Does that rule out the potential to implement proximity

> operators? I need to

> implement NEAR (and then SAME for paragraph searches), but

> I'm a novice in

> terms of search engine implementations. Am I likely to be out

> of my depth

> attempting that right now with Lucene?

Lucene does not directly support paragraph-based searching.

Lucene does support proximity searches, e.g., exact phrases, and within-N

words (slop). Please see the documentation for PhraseQuery, especially the

setSlop(int) method:

http://jakarta.apache.org/lucene/api/org/apache/lucene/search/PhraseQuery.ht

ml

Phrase slop is thus essentially WITHIN. The queryParser class does not yet

have a syntax to specify slop.

> > Perhaps we should add a utility method such as:

> >

> > public static Set getHitTokens(Set queryTerms, Reader

> text, Analyzer a)

> This looks good, but what about the (future) case where you

> have complex

> (possibly nested) proximity searches and only want to

> highlight the relevant

> tokens when they appear near each other?

As you point out, the method I suggest would highlight isolated occurrences

of terms from query phrases in hit documents, even when they do not occur in

phrases. (Note that for the document to be a hit, they will somewhere also

occur together in a phrase, and possibly quite frequently for a high-scoring

hit.) Google and most other search engines implement term highlighting this

way, and I think it is acceptable. One could of course write a

TokenStream-based query evaluator that correctly interpreted phrasal

restrictions when highlighting. Personally, I do not think it is worth the

effort, so I am not volunteering to do it myself.

Doug