http://mail-archives.apache.org/mod\_mbox/lucene-dev/200110.mbox/%3c4BC270C6AB8AD411AD0B00B0D0493DF0EE7C8A@mail.grandcentral.com%3e

From lucene-dev-return-253-apmail-jakarta-lucene-dev-archive=jakarta.apache.org@jakarta.apache.org Thu Oct 18 20:59:19 2001

Return-Path: <lucene-dev-return-253-apmail-jakarta-lucene-dev-archive=jakarta.apache.org@jakarta.apache.org>

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

Received: (qmail 73142 invoked by uid 500); 18 Oct 2001 20:59:19 -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 73126 invoked from network); 18 Oct 2001 20:59:19 -0000

Message-ID: <4BC270C6AB8AD411AD0B00B0D0493DF0EE7C8A@mail.grandcentral.com>

From: Doug Cutting <DCutting@grandcentral.com>

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

Subject: RE: boosting

Date: Thu, 18 Oct 2001 13:48:18 -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

I think document boosting would be a great feature to add to Lucene. This

is the sort of mechanism that, e.g., Google uses to boost documents from

highly referenced sites.

Lucene currently has a factor per document field that is multiplied into

scores: the norm. This is used to normalize for field length. The norm is

calculated as

1/sqrt(numTerms)

where numTerms is the number of terms in a particular field of a particular

document. Thus a term that occurs once in a four term field (norm=1/2)

scores twice as highly as a term that occurs once in an 16 term field

(norm=1/4).

A document boost could be pre-multiplied into the norm for each field in the

document, and stored in the index, as usual. The search algorithm could

stay the same, and would not need to do any additional calculations. So far

so good.

The problem is that norms are stored in the index as a single byte, by

multiplying their raw value by 255. Thus it is impossible to boost a field

with only a single term, since they already have the maximum value that can

be stored as a norm, 255. To get around this, a new representation for

norms is required.

One approach would be to store the norms as 255\*sqrt(norm), so that the

maximum unboosted value would be 16. Then, to when a byte is read use

byte^2/255 to convert it back to the norm value. This would reduce the

scale of unboosted norms to from one to sixteen, instead of from one to 255,

but probably wouldn't actually make scoring quality much worse. Then

document boosts could be by up to a factor of 16.

Does that sound like a good way to do this?

Doug

-----Original Message-----

From: soshima@business.com [mailto:soshima@business.com]

Sent: Thursday, October 18, 2001 11:49 AM

To: lucene-dev@jakarta.apache.org

Subject: boosting

Lucene has term query boosting for fields. But does anyone know how to do

individual Document boosting? So basically I want to put a numerical value

into a document and depending on its weight have it be more/less revelant

than other documents...Its basically mixing revelancy search with numerical

sorting.....Anyways I just want to boost docs. Thanks.

-scott