

Re: SVD when calculated for a corpus of similar category?

Re: SVD when calculated for a corpus of similar category?

Source: <http://sci.tech-archive.net/Archive/sci.math.num-analysis/2006-11/msg00164.html>

- *From:* spellucci@xx (Peter Spellucci)
 - *Date:* Wed, 15 Nov 2006 14:59:42 +0000 (UTC)
-

In article <1163589649.820409.148250@xx>, "paluri" <santoshpaluri@xxxxxxxx> writes:

Peter Spellucci wrote:

In article <1163579376.066684.82740@xx>, "paluri" <santoshpaluri@xxxxxxxx> writes:
>Hello everybody,
>
>I have a doubt regarding SVD. Suppose i compute SVD for a huge corpus
>of similar category, and i have the decomposition as $[USV^T]$. Are the
>singular values in the diagonal matrix S arranged in descending order
>along the diagonal will be very near to each other? i mean, is the
>numerical difference between one singular value and the next in the
>diagonal will be negligible?. I would be thankful for the response...
>

??????? what please is a "huge corpus of similar category"?
you mean a set of nearby matrices ???
then the answer is yes:
let $\sigma(A,i)$ denote the singular values of A in descending order
and $\sigma(B,i)$ those of B. Then
 $|\sigma(A,i) - \sigma(B,i)| \leq \|A - B\|$ for all i
 $\|A - B\| = \sigma(A - B, 1)$
hence if A is near B, then all the singular values of A and B can be paired
corresponding to this order with this universal error bound
(follows from the Courant-Fischer-minimax characterization of eigenvalues)
hth
peter

By "huge corpus of similar category", i mean web pages downloaded from a similar category,
Actually i am creating a term by document matrix (rows indicating the terms, columns the documents and each element of the matrix indicating

Re: SVD when calculated for a corpus of similar category?

the frequency of each term in the corresponding document) of certain number of web pages and then i will apply SVD to that term by document matrix in order to calculate the similarity between the documents or web pages.

Now, what i am asking is , if i create the term by document matrix of pages or documents taken from the same category, i.e if they are already similar, then in the SVD of the Term by Document matrix which i create using these similar pages, does the singular values in the diagonal will be very near to each other, i.e. the numerical difference between one and next singular value in the diagonal will be very small..?

no.

your matrices will be integer matrices with entries not larger than the number of occurrences of a term in a document, hence not really large. the norm of that matrix will be at most number of elements times the largest element and not smaller than the largest entry. but if two matrices differ by one in an entry, at least some singular values will differ in the order of one, hence not really small
hth
peter

.