CS 395T **Large-Scale Data Mining** **Spring 2000**

Section 49460

Lecturer: Inderjit Dhillon Scribe: Neeraj Mittal

Lecture 04

Date: February 02, 2000

Keywords: *Singular Value Decomposition, Hubs and Authorities, Vector Space Model*

1 Singular Value Decomposition

Unlike the eigenvalue decomposition which exists only for square matrices, the singular value decomposition (SVD) exists for all matrices - it is a more fundamental decomposition. Let $\mathbf{A} \in \mathbf{R}^{m \times n}$, $m \geq n$, be a matrix. The *singular value decomposition* of $\mathbf{A}$ is the factorization $\mathbf{A} = \mathbf{U\Sigma V}^T$, where $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_n$ and $\mathbf{\Sigma}$ is a diagonal matrix, $\text{diag}(\sigma_1, \ldots, \sigma_n)$ with $\sigma_i \geq 0$, $0 < i \leq n$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are referred to as the *left* and *right singular vectors*, respectively, and the *singular values* of $\mathbf{A}$ are the diagonal elements of $\mathbf{\Sigma}$.

Using the singular value decomposition of $\mathbf{A}$, we have $\mathbf{AA}^T = \mathbf{U\Sigma V}^T\mathbf{V\Sigma}^T\mathbf{U}^T = \mathbf{U\Sigma I\Sigma U}^T = \mathbf{U\Sigma}^2\mathbf{U}^T$. Thus, the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{AA}^T$. Similarly, $\mathbf{A}^T\mathbf{A} = \mathbf{V\Sigma}^2\mathbf{V}^T$. Therefore, the columns of $\mathbf{V}$ are the eigenvectors of $\mathbf{A}^T\mathbf{A}$. The singular values of $\mathbf{A}$ are the non-negative square roots of the eigenvalues of $\mathbf{AA}^T$ or $\mathbf{A}^T\mathbf{A}$.

Let $\mathbf{u}_i$ and $\mathbf{v}_i$, $0 < i \leq n$, denote the i-th column of $\mathbf{U}$ and $\mathbf{V}$, respectively. Then,

$$
\begin{aligned}
\mathbf{A} &= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_n \end{bmatrix} \begin{bmatrix} \sigma_1\mathbf{v}_1^T \\ \sigma_2\mathbf{v}_2^T \\ \vdots \\ \sigma_n\mathbf{v}_n^T \end{bmatrix} \\
&= \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_n\mathbf{u}_n\mathbf{v}_n^T
\end{aligned}
$$

Let $\mathbf{A}_k = \sigma_1\mathbf{u}_1\mathbf{v}_1^T + \sigma_2\mathbf{u}_2\mathbf{v}_2^T + \cdots + \sigma_k\mathbf{u}_k\mathbf{v}_k^T$, $k \leq n$, be the k-truncated SVD. Observe that $\mathbf{A}_k$ is a matrix of rank k. Among all matrices of rank k, $\mathbf{A}_k$ serves as the "best" approximation to $\mathbf{A}$ in the following sense. The 2-norm of the approximation error $\|\mathbf{A} - \mathbf{A}_k\|_2 \leq \|\mathbf{A} - \mathbf{M}_k\|_2$, for any matrix $\mathbf{M}_k$ of rank k. This is a classical result in linear algebra, for a proof see Theorem 2.5.3 on page 72 of [GL96].

2 Hubs and Authorities

Let $\mathbf{h}_i$ and $\mathbf{a}_i$ denote the hub and authority vector, respectively, at the end of i-th iteration of the algorithm given by Kleinberg [Kle98] (see the notes of Lecture 2). Further, let $\mathbf{A}$ denote the adjacency matrix of the sub-graph of the link graph of the web induced by the set of relevant web-pages in response to a query. We prove that the hub and authority vectors converge to the principal left and right singular vectors, respectively, of $\mathbf{A}$. For convenience, we give the i-th iteration of the algorithm here.

$$\boxed{\begin{array}{lcl} \mathbf{a}_i & = & \mathbf{A}^T\mathbf{h}_{i-1} \\ \mathbf{h}_i & = & \mathbf{A}\mathbf{a}_i \\ \multicolumn{3}{l}{\text{normalize } \mathbf{a}_i \text{ and } \mathbf{h}_i} \end{array}}$$

i-th iteration of the algorithm.

Ignoring the normalization and analyzing the first two iterations, we have,

$$\left.\begin{array}{lcl} \mathbf{a}_1 & = & \mathbf{A}^T\mathbf{h}_0 \\ \mathbf{h}_1 & = & \mathbf{A}\mathbf{a}_1 \end{array}\right\} \quad \Rightarrow \quad \mathbf{h}_1 = (\mathbf{A}\mathbf{A}^T)\,\mathbf{h}_0$$

and

$$\left.\begin{array}{lcl} \mathbf{a}_2 & = & \mathbf{A}^T\mathbf{h}_1 \\ \mathbf{h}_2 & = & \mathbf{A}\mathbf{a}_2 \end{array}\right\} \quad \Rightarrow \quad \mathbf{h}_2 = (\mathbf{A}\mathbf{A}^T)\,\mathbf{h}_1 = (\mathbf{A}\mathbf{A}^T)^2\,\mathbf{h}_0$$

In general, with normalization,

$$\begin{array}{lcl} \mathbf{h}_k & = & \dfrac{(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{h}_0}{\|(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{h}_0\|}, \text{ and} \\ \mathbf{a}_k & = & \dfrac{(\mathbf{A}^T\mathbf{A})^{k-1}\,\mathbf{a}_1}{\|(\mathbf{A}^T\mathbf{A})^{k-1}\,\mathbf{a}_1\|}. \end{array}$$

Note that both $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ are symmetric and semi-positive definite, that is, all eigenvalues are non-negative. We first prove some properties of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$.

Theorem 1 *Let $\mathbf{A} \in \mathbf{R}^{m\times n}$ be a matrix. Then $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ have identical non-zero eigenvalues. Further, if $\mathbf{q}$ is an eigenvector of $\mathbf{A}\mathbf{A}^T$ then $\mathbf{A}^T\mathbf{q}$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$.*

Proof: Let λ be a non-zero eigenvalue of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{q}$ be the eigenvector corresponding to λ. Then $(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,\mathbf{q}$. Premultiplying both sides by $\mathbf{A}^T$, we have $\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = (\mathbf{A}^T\mathbf{A})\cdot(\mathbf{A}^T\mathbf{q}) = \lambda\cdot(\mathbf{A}^T\mathbf{q})$. Therefore λ is an eigenvalue of $\mathbf{A}^T\mathbf{A}$ with $\mathbf{A}^T\mathbf{q}$ as the corresponding eigenvector. ∎

Theorem 2 *Let $\mathbf{A} \in \mathbf{R}^{m\times n}$ be a matrix. Then $\mathbf{A}\mathbf{A}^T$ is a semi-positive definite matrix.*

Proof: Let λ be an eigenvalue of $\mathbf{A}\mathbf{A}^T$ and $\mathbf{q}$ be the eigenvector corresponding to λ. Then $(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,\mathbf{q}$. Premultiplying both sides by $\mathbf{q}^T$, we have $\mathbf{q}^T\mathbf{A}\mathbf{A}^T\mathbf{q} = \lambda\,\mathbf{q}^T\mathbf{q}$ which implies that $\lambda = \dfrac{\mathbf{q}^T\mathbf{A}\mathbf{A}^T\mathbf{q}}{\mathbf{q}^T\mathbf{q}} = \dfrac{\mathbf{z}^T\mathbf{z}}{\mathbf{q}^T\mathbf{q}}$, where $\mathbf{z} = \mathbf{A}^T\mathbf{q}$. Observe that $\mathbf{q}^T\mathbf{q} > 0$ and $\mathbf{z}^T\mathbf{z} \geq 0$. Therefore $\lambda \geq 0$ implying that all eigenvalues of $\mathbf{A}\mathbf{A}^T$ are non-negative. ∎

We now prove that the eigenvalues of $(\mathbf{A}\mathbf{A}^T)^k$, $k \geq 1$, are related to the eigenvalues of $\mathbf{A}\mathbf{A}^T$. In particular, if λ is an eigenvalue of $\mathbf{A}\mathbf{A}^T$ then λ^k is an eigenvalue of $(\mathbf{A}\mathbf{A}^T)^k$. Moreover, $\mathbf{A}\mathbf{A}^T$ and $(\mathbf{A}\mathbf{A}^T)^k$ have same eigenvectors.

Theorem 3 *Let $\mathbf{A} \in \mathbf{R}^{m \times n}$ be a matrix. Further, let λ and $\mathbf{q}$ be an eigenvalue and eigenvector, respectively, of $\mathbf{A}\mathbf{A}^T$. Then λ^k and $\mathbf{q}$ are the eigenvalue and eigenvector, respectively, of $(\mathbf{A}\mathbf{A}^T)^k$, $k \geq 1$.*

Proof: We know that $(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,\mathbf{q}$. Premultiplying both sides by $\mathbf{A}\mathbf{A}^T$, we have $(\mathbf{A}\mathbf{A}^T)^2\,\mathbf{q} = \lambda\,(\mathbf{A}\mathbf{A}^T)\,\mathbf{q} = \lambda\,(\lambda\,\mathbf{q}) = \lambda^2\,\mathbf{q}$. Thus, by mathematical induction, $(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{q} = \lambda^k\,\mathbf{q}$, $k \geq 1$. ∎

Let λ_i and $\mathbf{q}_i$, $0 < i \leq n$, be the eigenvalues and the corresponding eigenvectors, respectively, of $\mathbf{A}\mathbf{A}^T$. Without loss of generality, assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. We can write $(\mathbf{A}\mathbf{A}^T)^k\,\mathbf{h}_0$ as $\lambda_1^k \mathbf{q}_1 \mathbf{q}_1^T \mathbf{h}_0 + \lambda_2^k \mathbf{q}_2 \mathbf{q}_2^T \mathbf{h}_0 + \cdots + \lambda_n^k \mathbf{q}_n \mathbf{q}_n^T \mathbf{h}_0$. If $\lambda_1 > \lambda_2$ and $\mathbf{q}_1^T \mathbf{h}_0 \neq 0$ then the hub vector converges to $\mathbf{q}_1$. It is possible to prove that the principal eigenvector of $\mathbf{A}\mathbf{A}^T$ is a non-negative vector [HJ85, p.503]. Since $\mathbf{h}_0$ is the vector of all 1's, this implies that $\mathbf{q}_1^T \mathbf{h}_0 \neq 0$. Note that $\mathbf{q}_1$ is the principal left singular vector of $\mathbf{A}$. Similarly, the authority vector converges to $\mathbf{A}^T \mathbf{q}_1$ - the principal right singular vector of $\mathbf{A}$.

3 Vector Space Model

We discuss the vector space model to represent a set of unstructured text documents. The basic idea is to represent each document as a vector of certain weighted word frequencies. Before we can construct the representation, the documents must be parsed and preprocessed as follows:

1. Ignoring case, extract all unique words from the entire set of documents.
2. Eliminate non-content-bearing *stopwords* such as "a", "and", "the", etc. Standard lists of stopwords exist for the English language.
3. For each document, count the number of occurrences of each word.
4. Using heuristic, eliminate non-content-bearing "high-frequency" or "low-frequency" words. Eliminating such words and the stopwords removes little or no information, while speeding up the computation.
5. After above elimination, suppose w unique words remain. Assign a unique identifier between 1 and w to each remaining word, and a unique identifier between 1 and d to each document.

The above preprocessing scheme yields the number of occurrences of word j in document i, say f_{ji}, and the number of documents which contain the word j, say d_j. Using these counts, we now create d document vectors in $\mathbf{R}^w$, namely $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ as follows. For $0 < j \leq w$, set the j-th component of document vector $\mathbf{x}_i$, $0 < i \leq d$, to the product of three terms,

$$x_{ji} = t_{ji} \cdot g_j \cdot s_i$$

where t_{ji} is the *term weighting component* and is based on f_{ji}, g_j is the *global weighting component* and is based on d_j, and s_i is the *normalization component* for $\mathbf{x}_i$. Intuitively, t_{ji} captures the relative importance of a word in a document, while g_j captures the overall importance of a word in the entire set of documents. The objective of such weighting schemes is to enhance discrimination between various documents vectors and to enhance retrieval effectiveness.

There are many schemes for selecting the term, global, and normalization components. For example, t_{ji} could be set to 1 if the word j occurs in the document i and to 0 otherwise. A drawback of this scheme is that it gives every word that appears in a document equal relevance.

Alternatively, f_{ji} could be used for t_{ji}. This scheme gives more credit to words that appear more frequently, but often *too much* credit. For instance, a word that appears ten times in a document is not usually ten times more important that a word that only appears once. Logarithms can be used to deemphasize the frequency. For example, t_{ji} could be set to $\log(f_{ji}+1)$.

Global weights are used to place emphasis on terms that are discriminating, and they are based on the dispersion of a particular word throughout the documents. For example, g_j could be set to $\log(\frac{d}{d_j})$. This formula is known as the *inverse document frequency* (IDF) and gives larger weight to a word that occurs in fewer documents. The use of global weighting can, in theory, eliminate the need for stopword removal since stopwords would have very small weights. In practice, it is easier to remove the stopwords in the preprocessing phase so that there are fewer words to handle.

Once we have computed the term and global weights, it is often useful to normalize the columns in the final matrix. If we do not, short documents may not be recognized as relevant. To normalize, s_i can be set to $(\sum_{j=1}^{w}(t_{ji}\,g_j)^2)^{-\frac{1}{2}}$. Note that this particular normalization causes all the document vectors to lie on the unit sphere. Kolda [Kol97] presents 5, 5, and 2 schemes, respectively, for the term, global, and normalization components.

Finally, a given query can be formulated as a query vector $\mathbf{q}$, and the documents that contain the query words may be returned by the vector $\mathbf{X}^T\mathbf{q}$, where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d]$. Note that the query vector entries may also be scaled by the term, global and normalization components. A weighting scheme can, therefore, be specified by a six letter combination that indicates term, global and normalization components for the document matrix and the term, global and normalization components for the query [Kol97], for example, `lfn.txx` or `lgn.lgx`. The SMART system developed at Cornell uses the above vector space model for query retrieval.

References

[GL96] Gene H. Golub and Charles F. Van Loan. *Matrix computations.* Johns Hopkins Studies in the Mathematical Sciences. The Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[HJ85] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

[Kle98] Jon Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms*, 1998. Extended version in Journal of the ACM 46(1999). Also appears as IBM Research Report RJ 10076, May 1997.

[Kol97] T. G. Kolda. *Limited-Memory Matrix Methods with Applications.* PhD thesis, The Applied Mathematics Program, University of Maryland, College Park, Mayland, 1997.