Spectral Analysis of Data*

Yossi Azar[†] Amos Fiat[†] Anna R. Karlin[‡] Frank McSherry[‡] Jared Saia[‡]

Every action must be due to one or other of seven causes: chance, nature, compulsion, habit, reasoning, anger, or appetite.
— Aristotle, *Rhetoric, Bk. II.*

No one wants advice — only corroboration.
— John Steinbeck, *The Winter of Our Discontent.*

ABSTRACT

Experimental evidence suggests that spectral techniques are valuable for a wide range of applications. A partial list of such applications include (i) semantic analysis of documents used to cluster documents into areas of interest, (ii) collaborative filtering — the reconstruction of missing data items, and (iii) determining the relative importance of documents based on citation/link structure. Intuitive arguments can explain some of the phenomena that has been observed but little theoretical study has been done. In this paper we present a model for framing data mining tasks and a unified approach to solving the resulting data mining problems using spectral analysis. These results give strong justification to the use of spectral techniques for latent semantic indexing, collaborative filtering, and web site ranking.

1. INTRODUCTION

Spectral techniques have proven, at least empirically, useful in a variety of data mining applications [4, 12, 11]. To apply these techniques, the data is typically represented as a set of vectors in a high-dimensional space. For example, if the data set is a corpus of documents, then each document can be represented as a a vector of terms $\vec{d}$, where the i-th component of the vector, d_i, is 1 if the i-th term occurs in the document and is 0 otherwise. With such a representation, the entire corpus can be viewed as a matrix, say A, each of whose columns represent a document.

The matrix representation can be used for other types of data sets where columns index objects in the data set, rows index attributes of those objects, and the $[i, j]$ entry of the matrix represents the value of the i-th attribute in the j-th object. Some examples of interest are where both rows and columns refer to web sites and the $[i, j]$ entry indicates that site i has a link to site j; another is that columns refer to individuals, rows refer to products, and the $[i, j]$ entry indicates something about how much individual j likes product i.

In this paper, we consider the application of spectral techniques to a variety of data mining tasks. We begin by presenting a general model that we believe captures many of the essential features of important data mining tasks. We then present a set of conditions under which data mining problems in this framework can be solved using spectral techniques, and use these results to theoretically justify the prior empirical success of these techniques for tasks such as object classification and web site ranking. We also use our theoretical framework as a foundation for developing new algorithms for collaborative filtering. Our data mining models allow both erroneous and missing data, and show how and when spectral techniques can overcome both.

The data mining model we introduce assumes that the data of interest can be represented as an object/attribute matrix. The model is depicted in Figure 1 which shows how three fundamental phenomena combine to govern the process by which a data set is created:

1. **A probabilistic model of data M:** We assume that there exists an underlying set of probability distributions that govern each object's attribute values (in the degenerate case, these values could be deterministically chosen). These probability distributions are captured by the probabilistic model M in the figure, where the random variable describing the ith attribute of the j-th object is denoted $M_{i,j}$. The actual value of this attribute is then obtained by sampling from the distribution $M_{i,j}$; we denote the resulting value m_{ij}. We assume that the M_{ij}'s are independent.

2. **An error process Z:** We assume that the data is noisy and error-ridden. The error process Z describes the manner by which the error is generated. We assume that the data value m_{ij} is corrupted by the addition of the error z_{ij}.

[*] Research supported in part by NSF EIA-9870740 and a grant from the US-Israel Binational Science Foundation (BSF).

[†] Dept. of Computer Science, Tel Aviv University, Tel-Aviv 69978, Israel

[‡] Dept. of Computer Science, University of Washington at Seattle

STOC'01 July 6-8, 2001, Hersonissos, Crete, Greece.
Copyright 2001 ACM 1-58113-349-9/01/0007 ...$5.00.

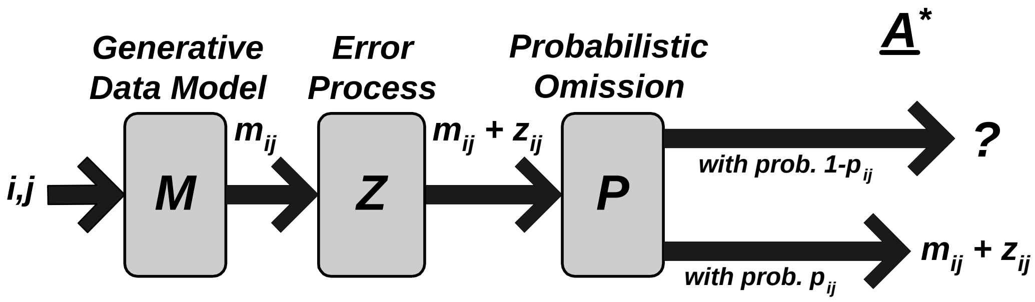

Figure 1: The Data Generation Model

3. **An omission process P:** Some of the data may not be available to the data miner. In our model, we assume that there is a probability distribution P governing the process by which data is omitted or made available. In particular, the value $m_{ij} + z_{ij}$ is available to the data miner with probability p_{ij}, and is omitted from the data set (which we represent by the presence of a "?") with probability $1 - p_{ij}$. We denote by A^* the resulting data set (which is then input to the data mining algorithm).

The goal of the data mining algorithm: given A^* as input (and no knowledge of M, P or Z), obtain meaningful information about M. In particular, we are interested in obtaining information about the matrix $\mathbf{E}(M)$, whose (i, j) entry is the expectation of the random variable M_{ij}.

1.1 Contributions of this Paper

Clearly, without any assumptions about M, Z and P, it is hopeless to achieve the data mining goal just laid out. The first contribution of this paper is to present a general set of conditions under which it is possible to efficiently retrieve meaningful information about $\mathbf{E}(M)$. In essence, our results show that if the underlying data model is sufficiently "structured", then the randomness of a probabilistic process, the addition of error and the fact that a significant fraction of the data may be missing will not prevent the data miner from recovering meaningful information about the "true" data.

More formally, we prove the following general theorem:

THEOREM 1. *Suppose that the availability matrix P is known to the data mining algorithm, and its entries are bounded away from 0. In addition, suppose that $\mathbf{E}(M)$ is a rank k matrix*[1] *with $\sigma_k = \omega(\sqrt{n})$, and the 2-norm of the error matrix Z is $o(\sigma_k)$, where σ_k is the k-th singular value of $\mathbf{E}(M)$. Then there is a polynomial time algorithm,* that takes as input only P and A^*, *that is guaranteed to reconstruct $1 - o(1)$ of the entries of $\mathbf{E}(M)$ to within an additive $o(1)$ error.*

Justification of Assumptions

The fundamental assumption being made in this paper is that $\mathbf{E}(M)$ is well approximated by a low rank matrix. A natural question to ask is whether such an assumption is justified in the context of data mining applications.

In fact, this question is fundamentally outside the scope of this paper, but intuitive arguments abound. In common to all the applications we consider one can argue philosophically that people, products, documents, terms in classical Greek, web sites, etc., are all inherently determined by or associated with a small number of fundamental properties, where each individual person, product, word, etc., can be described by a weighted vector of these base properties. Thus, the justification that $\mathbf{E}(M)$ be of low rank. The examples above of a document corpus, personal preferences and web links can all be placed in this framework. Of course, the real justification is empirical.

1.2 Applications

The second contribution of this paper is in showing that the data mining model we have described captures a number of bona fide important data mining problems, and in presenting a unified approach to their solution using Theorem 1. In this respect, our major results are the following:

Analysis of LSI as an information retrieval tool

Latent semantic indexing (LSI) is a successful technique for information retrieval (IR) from documents. It is empirically effective at overcoming *synonymy* (car vs. automobile) and *polysemy* (WWW spider vs. eight-legged spider). In an important first step towards providing a theoretical justification for the empirical success of LSI, Papadimitriou, Raghavan, Tamaki and Vempala [14] presented a probabilistic model describing the generation of a corpus of documents on a set of topics and showed that, for documents generated according to this model, the k-dimensional subspace produced by LSI yields, with high probability, sharply defined clusters among documents on each topic with respect to the cosine measure[2]. Two limitations of their model are that documents are assumed to be nearly "pure" (the subject of a single topic) and terms associated with different topics are assumed to be disjoint, and hence there is no polysemy in their model. An open problem from their paper is to extend their justification of LSI as an IR tool to more general document generation models, especially ones which incorporate polysemy.

We can describe this open problem as the *information retrieval problem*, which is a special case of our general data mining model: We assume that the presence of term i in document j (or the value of attribute i for object j) is a random variable with mean $\mathbf{E}(M_{ij})$. Thus, documents of similar se-

[1] Similar results hold if $\mathbf{E}(M)$ is well-approximated by a low rank, say rank k, matrix.

[2] which measures the distance between documents as the cosine of the angle between their corresponding vectors.

mantic composition are generated from similar probability distributions, however this similarity is hidden from the information retrieval algorithm by the probabilistic generative process. We further assume that the corpus is corrupted by an additive error process Z. The goal of the information retrieval algorithm is to learn meaningful data about the matrix $\mathbf{E}(M)$, such as the angles between the columns of $\mathbf{E}(M)$, given only the matrix A^* as input.

Papadimitriou *et. al.*'s result essentially shows that LSI is capable of computing meaningful information about $\mathbf{E}(M)$ when it is a particular type of low rank matrix, namely a slightly perturbed block matrix. We generalize this to show that LSI solves the information retrieval problem when M is an arbitrary matrix such that $\mathbf{E}(M)$ is well approximated by a low rank matrix. We also allow an additional error matrix Z — any error matrix of independent random values with mean 0 and constant deviation. Thus, our results prove that LSI works in wider variety of settings than those considered by [14], and in particular provides theoretical justification for the fact that LSI can overcome the problem of polysemy.

Collaborative Filtering

A fundamental problem in data mining, usually referred to as collaborative filtering (or recommendation systems) is to use partial information that has been collected about a group of users to make recommendations to individual users. (See e.g., [2, 8, 15, 13, 16, 9].) For instance, a movie recommendation system might recommend "Happiness" to someone who enjoyed "American Beauty" or "Alice in Wonderland" to someone who enjoyed "The Phantom Tollbooth". More generally, collaborative filtering can be viewed as the problem of taking an incomplete data set and attempting to determine properties of the absent data (perhaps complete reconstruction). To our knowledge, there has been very little prior theoretical work on collaborative filtering algorithms other than the work of Kumar, Raghavan, Rajagopalan and Tomkins who took an important first step of defining an analytic framework for evaluating collaborative filtering [13].

We model the collaborative filtering problem within the framework of our general data mining model as follows: We assume that the utility of product j for individual i is given by a random variable M_{ij} and which data is missing is determined by a probabilistic omission process P.

Once again, we assume that the matrix $\mathbf{E}(M)$ is well approximated by a low-rank matrix. For the collaborative filtering problem, this can be viewed as a psychological assumption on the simplistic nature of humankind. Under this assumption, we present an algorithm that, for any P whose entries are bounded away from 0, given a random subset of the entries of A^* (the instantiation of M, followed by discarding elements using P), can provably compute $\mathbf{E}(M[i,j])$ for a $1 - o(1)$ fraction of the missing entries of A^*, with $1 - o(1)$ accuracy.

Comparing these results to those of Kumar et al [13], we observe that their psychological assumptions about humanity are much more simplistic than ours[3], and they also require more a-priori information than we do. Our collaborative filtering algorithms handle any utility matrix with a good low rank approximation. No clustering or *a priori* knowledge of object similarity is required.

Theoretical support for Kleinberg's algorithm

Kleinberg's seminal work on web *hubs* and *authorities* has had a true impact on the real world [11].

We model the determination of hub and authority scores within the framework of our data mining model as follows. We assume that the matrix A^* is the result of a probabilistic process that determines whether a certain site will refer to another or not, based on the true importance (as an authority or a hub) of a site. The simplest version of the result assumes that there is a pair of vectors h and a, with entries between 0 and 1, such that h_i represents the true importance of web site i as a hub and a_i represents the true importance of web site i as an authority. The existence of a link from site i to site j is then a Bernoulli random variable with expectation $h_i a_j$.

It is an immediate consequence of our results that Kleinberg's definition of importance is robust in the sense that the important sites will remain important (almost) irrespective of the actual random choices made when the "real world" is constructed.

Paper Layout

Our paper will follow the following outline. Section 2 will give mathematical preliminaries. In Section 3, we will present results on the stability of the strong singular subspaces of a matrix A after perturbation by an additive error matrix E. In Section 4, we will specialize these stability results to the case where the entries of E are independent random variables with mean 0 and constant deviation. Finally, in Section 5, we use these stability results to solve the data mining problems discussed in the introduction.

2. PRELIMINARIES

We begin by reviewing some background material and then summarize our notation.

2.1 The Singular Value Decomposition

The **singular value decomposition** (SVD) of an m by n matrix A is a manner of rewriting the matrix as

$$A = UDV^T$$

where U and V are orthogonal $m \times m$ and $n \times n$ matrices and D is a diagonal matrix whose diagonal entries, σ_i, we call the singular values. These singular values are non-increasing and non-negative. The singular value decomposition is defined for all A and is unique up to certain degeneracies involving equal singular values. Observe that for v_i a column of V and u_i a column of U it is the case that $Av_i = \sigma_i u_i$ and $A^T u_i = \sigma_i v_i$. For this reason, we call (u_i, v_i) a **singular vector pair**. Singular vector pairs have an association with eigenvectors in that u_i and v_i are the eigenvectors corresponding to the i^{th} largest eigenvalues of the matrices AA^T and $A^T A$ respectively.

We define the m by n matrix A_k as

$$A_k = U_k D_k V_k^T$$

where U_k is the $m \times k$ matrix consisting of the first k columns of U, D_k is the $k \times k$ diagonal matrix consisting of the top k singular values, and V_k is the $n \times k$ matrix consisting of the first k columns of V. A very useful property of the SVD is that A_k is the best rank k approximation to A: of all

[3] It is not perfectly clear that this is a weakness of their model...

rank k matrices M, A_k minimizes the error $|A - M|_2$. In fact, it is the case that $|A - A_k| = \sigma_{k+1}$ which leads us to the conclusion that $\sigma_i - \sigma_{i+1}$ represents the importance of incorporating the i^{th} singular vector pair into our approximation.

See [7] for a more complete discussion of the SVD and its properties.

2.2 Symmetric versus non-symmetric matrices

In our proofs, we will be interested in applying theorems about symmetric matrices to non-symmetric matrices. It will be convenient for us to use the following well known relation [7] between the singular value decomposition of a non-symmetric matrix A and the symmetric (eigen) decomposition of the symmetric matrix

$$B = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}.$$

In particular, observe that if (u_i, v_i) is a singular vector pair of A, then both

$$\begin{bmatrix} v_i \\ u_i \end{bmatrix} \text{ and } \begin{bmatrix} v_i \\ -u_i \end{bmatrix}$$

are eigenvectors of the matrix B with eigenvalues σ_i and $-\sigma_i$ respectively. All other eigenvectors have eigenvalue zero. It is important to note that the top k eigenvectors of B will correspond to the top k singular vector pairs of A , and that their eigenvalues and singular values correspond exactly.

2.3 Summary of Notation

As above, let A be some original matrix, E a perturbation matrix, and $\widehat{A} = A + E$.

We use the matrix product UDV^T to denote the SVD of A, and the matrix product $\widehat{U}\widehat{D}\widehat{V}^T$ to denote the SVD of $\widehat{A}$. The diagonal elements of D (resp. $\widehat{D}$) are the singular values of A (resp. $\widehat{A}$) and are denoted, in non-increasing order, σ_i (resp. $\widehat{\sigma}_i$). Similarly, we will use $A_k = U_k D_k V_k^T$ (resp. $\widehat{A}_k = \widehat{U}_k \widehat{D}_k \widehat{V}_k^T$) to denote the best rank k approximation to A (resp. $\widehat{A}$).

Throughout the paper we will be using the 2-norm of matrices and vectors. The 2-norm of a vector v is of course defined as $|v|_2 = \sqrt{\sum_i v_i^2}$. The 2-norm of a matrix is

$$|M|_2 = \max_{|u|_2=1} |Mu|_2$$

We will drop the subscript 2 from all norms for clarity.

Finally, for any matrix M, we will use $M^{(i)}$ to refer to its i-th column. We denote by a_i the projection of the $A^{(i)}$ onto the first k columns of U (where the dimension k is implicit), i.e., $a_i = U_k^T A^{(i)}$.

3. THE STABILITY OF SINGULAR SUBSPACES

The principal work of the paper concerns the potential to retrieve strong singular subspaces of a matrix A after some perturbation E is applied. Analyses in a similar spirit, but less general, have been conducted in [14, 10].

Central to our results is the following result of Stewart's [7] describing the stability of eigenvectors of symmetric matrices after a symmetric error is applied.

Theorem 2 (Stewart). *[17] Let B and $B+F$ be symmetric $n \times n$ matrices with eigenvectors*

$$Q = \begin{bmatrix} Q_1 \; Q_2 \end{bmatrix} \text{ and } \widehat{Q} = \begin{bmatrix} \widehat{Q}_1 \; \widehat{Q}_2 \end{bmatrix}$$

where both Q_1 and $\widehat{Q}_1$ are $n \times k$ matrices. Let λ_i (resp. $\widehat{\lambda}_i$) be the eigenvalue associated with the i-th column of Q (resp. $\widehat{Q}$). If $|F|$ is $o(\widehat{\lambda}_k - \lambda_{k+1})$, then

$$\widehat{Q}_1 = Q_1 R + F_Q$$

where R is an orthogonal matrix and $|F_Q|$ is $O\left(\frac{|F|}{|\widehat{\lambda}_k - \lambda_{k+1}|}\right)$.

We use Stewart's theorem to obtain the following corollary.

Corollary 3. *Let A and $\widehat{A} = A + E$ be $m \times n$ real matrices where*

$$A = UDV^T \text{ and } \widehat{A} = \widehat{U}\widehat{D}\widehat{V}^T$$

are the SVDs of A and $\widehat{A}$ Let $\delta_k = \widehat{\sigma}_k - \sigma_{k+1}$. If $|E|$ is $o(\delta_k)$ we can write the first k columns of $\widehat{U}$ and $\widehat{V}$ as

$$\begin{aligned} \widehat{U}_k &= U_k R + E_U \\ \widehat{V}_k &= V_k R + E_V \end{aligned}$$

where U_k and V_k are the first k columns of U and V respectively, R is an orthonormal matrix and the norms of E_U and E_V are $O(|E|/\delta_k)$.

Proof. We will apply Theorem 2, letting

$$B = \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix} \text{ and } F = \begin{bmatrix} 0 & E^T \\ E & 0 \end{bmatrix}$$

From Section 2.2, we have that

$$\begin{bmatrix} v_i \\ u_i \end{bmatrix} \text{ and } \begin{bmatrix} \widehat{v}_i \\ \widehat{u}_i \end{bmatrix}$$

are eigenvectors of B and $B + F$ with corresponding eigenvalues σ_i and $\widehat{\sigma}_i$. Since the norm of F is equal to the norm of E, and $\widehat{\lambda}_k - \lambda_{k+1} = \widehat{\sigma}_k - \sigma_{k+1}$, Theorem 2 allows us to conclude that

$$\begin{bmatrix} \widehat{V}_k \\ \widehat{U}_k \end{bmatrix} = \begin{bmatrix} V_k \\ U_k \end{bmatrix} R + \begin{bmatrix} E_V \\ E_U \end{bmatrix}$$

which we decompose into the conclusion of the corollary. □

We are now able to present the main result of this section, which gives the precise assumptions needed to preserve the angles between rows or columns of a matrix projected into its top k singular vectors after it is perturbed by an additive error matrix. We will show later that there are many applications where all of these assumptions are met.

Theorem 4. *Let A and $\widehat{A} = A+E$ be $m \times n$ real matrices. Assume that $|E| \in o(\widehat{\sigma}_k - \sigma_{k+1})$. We define*

$$a_i = U_k^T A^{(i)} \qquad \widehat{a}_i = \widehat{U}_k^T \widehat{A}^{(i)} \qquad \text{and} \qquad e_i = \widehat{U}_k^T E^{(i)}$$

If it is the case that

$$|a_i| \in \theta(|A^{(i)}|) \qquad \textit{and} \qquad |e_i| \in o(|a_i|)$$

then it is the case that

$$|a_i - R\widehat{a}_i| \in o(|a_i|) \qquad \textit{and} \qquad |A_k^{(i)} - \widehat{A}_k^{(i)}| \in o(|A_k^{(i)}|)$$

PROOF. We will apply Corollary 3 and conclude that $E_U^T = \widehat{U}_k^T - R^T U_k^T$ has norm $o(1)$. Recall that we assume that $|a_i| \in \theta(|A^{(i)}|)$ and $|e_i| \in o(|a_i|)$.

$$\begin{aligned}
|a_i - R\widehat{a}_i| &= |a_i - R\widehat{U}_k^T \widehat{A}^{(i)}| \\
&= |a_i - R\widehat{U}_k^T A^{(i)} + R\widehat{U}_k^T E^{(i)}| \\
&= |a_i - (U_k^T + RE_U^T)A^{(i)} + Re_i| \\
&= |-RE_U^T A^{(i)} + Re_i| \\
&\leq |E_U^T A^{(i)}| + |e_i| \\
&\in o(|a_i|)
\end{aligned}$$

The second result is of a similar nature, but has one particular advantage over the relation between a_i and $\widehat{a}_i$. Namely, we are actually able to compute $\widehat{A}_k^{(i)}$ from $\widehat{A}$, whereas computation of $R\widehat{a}_i$ requires knowledge of the matrix R, which is unfortunately unavailable for the purposes of our data mining applications.

$$\begin{aligned}
|A_k^{(i)} - \widehat{A}_k^{(i)}| &= |U_k a_i - \widehat{U}_k \widehat{a}_i| \\
&= |U_k a_i - (U_k R + E_U)\widehat{a}_i| \\
&= |U_k a_i - U_k R\widehat{a}_i - E_U \widehat{a}_i| \\
&\leq |U_k(a_i - R\widehat{a}_i)| + |E_U \widehat{a}_i| \\
&\leq |a_i - R\widehat{a}_i| + |E_U||\widehat{a}_i| \\
&\in o(|a_i|) \\
&\in o(|A_k^{(i)}|)
\end{aligned}$$

□

Theorem 4 generalizes results from [14, 10] from the case where A_k is a block diagonal matrix consisting of k blocks to the case where A_k is an arbitrary rank k matrix.

Although the theorem is presented in terms of columns, it applies equally well in the the context of rows of A and $\widehat{A}$.

We will be particularly interested in angles between vectors. Let $\angle(x, y)$ denote the angle between vectors x and y. A simple extension of spatial proximity to angular proximity can be used to obtain the following corollary.

COROLLARY 5. *Let a_i and a_j be projected columns of A.*[4] *If both a_i and a_j satisfy the conditions in Theorem 4, then it is the case that*

1. $|\angle(a_i, a_j) - \angle(\widehat{a}_i, \widehat{a}_j)| \in o(1)$
2. $|\angle(A_k^{(i)}, A_k^{(j)}) - \angle(\widehat{A}_k^{(i)}, \widehat{A}_k^{(j)})| \in o(1)$

4. STABILITY UNDER RANDOM PERTURBATION

For the data mining applications we shall study, we will need to specialize Theorem 4 to the case where the error E introduced is a random matrix whose entries are independent random variables with mean zero and constant deviation. Doing so yields the following corollary to Theorem 4.

COROLLARY 6. *Let A be a matrix with*

$$\sigma_k - \sigma_{k+1} \in \omega(\sqrt{m+n}).$$

Let $\widehat{A} = A + E$, where E is a matrix whose entries are independent random variables with mean zero and constant deviation. Let $a_i = U_k^T A^{(i)}$ be the projection of the i-th column of A onto A's top k singular vectors. Let $e_i = \widehat{U}_k^T E^{(i)}$. We say that this column is good *if $|a_i|$ is $\theta(|A^{(i)}|)$ and $|e_i|$ is $o(|a_i|)$. Then, with high probability, for all good columns i,*

1. $|a_i - R\widehat{a}_i| \in o(|a_i|)$
2. $|A_k^{(i)} - \widehat{A}_k^{(i)}| \in o(|A_k^{(i)}|)$.
3. $|A_k^{(i)}[\ell] - \widehat{A}_k^{(i)}[\ell]| \in o(1)$ *for all but $o(m)$ values of ℓ.*

Moreover, for any pair of columns i, j, such that both satisfy the previous conditions it follows that:

1. $|\angle(a_i, a_j) - \angle(\widehat{a}_i, \widehat{a}_j)| \in o(1)$
2. $|\angle(A_k^{(i)}, A_k^{(j)}) - \angle(\widehat{A}_k^{(i)}, \widehat{A}_k^{(j)})| \in o(1)$.

The same results hold, mutatis mutandis, for rows of the matrices A and $\widehat{A}$.

PROOF. This corollary's proof lies in observing that the norm of a random matrix whose entries are independent random variables with mean zero and constant deviation is almost certainly $\theta(\sqrt{m+n})$. This follows from a result of Boppana[5] [5] showing that such a symmetric random matrix of dimension n has norm $O(\sqrt{n})$ with high probability. We apply this observation to the matrix

$$\begin{bmatrix} 0 & E^T \\ E & 0 \end{bmatrix}$$

which has the same norm as E. Therefore, we get a bound of $O(\sqrt{m+n})$ on $|E|$. Since, in addition, $\widehat{\sigma_k} \geq \sigma_k - |E|$, we can conclude that $|E| \in o(\widehat{\sigma_k} - \sigma_{k+1})$, and thus Theorem 4 applies.

From the fact that the angles between good columns change by nominal amounts, we can conclude that the fraction of entries in these columns whose error is $\Omega(1)$ is at most $o(1)$. □

Remarks

1. Corollary 6 applies only to to columns which are "good". What does this really mean? The condition that a_i is $\theta(|A^{(i)}|)$ means that the i-th column of the matrix is well represented by the top k singular vectors. For example, if A were rank k every column would satisfy this condition. This condition is here to avoid complications with vectors who are not well described by

[4] In fact, one or both could be projected rows of A.

[5] who in turn extended a result of Furedi and Komlos [6]

the structure of A. Note that these vectors would be poorly approximated in A_k even without random error. The second condition, that $e_i \in o(|a_i|)$, is almost always true for those $|a_i| \in \omega(1)$. (See Lemma 10 in the Appendix for a precise version of this statement.)

2. When we apply corollary 6 we actually assume that $\widehat{\sigma}_k - \widehat{\sigma}_{k+1} \in \omega(\sqrt{m+n})$. Our purpose in so doing is to present the theorems in a form that allows the data miner to verify that the preconditions of the theorem hold. $\widehat{\sigma}_k - \widehat{\sigma}_{k+1} \in \omega(\sqrt{m+n})$ follows by observing that for any i, $|\sigma_i - \widehat{\sigma}_i| \leq |E|$ and $|E| \in O(\sqrt{m+n})$. Thus, $\sigma_k - \sigma_{k+1} \in \omega(\sqrt{m+n})$, if and only if $\widehat{\sigma}_k - \widehat{\sigma}_{k+1} \in \omega(\sqrt{m+n})$.

5. DATA MINING

We next show how Corollary 6 can be used to solve some of the data mining questions described in the introduction.

5.1 Information Retrieval

We begin by considering the information retrieval problem discussed in Section 1.2. In this context take $A = \mathbf{E}(M)$, a matrix whose $[i, j]$ entry is the expectation of M_{ij}, the random variable used to generate the (i, j)th entry of the pure model matrix. Let $\widehat{A}$ be the matrix whose $[i, j]th$ coordinate contains a sample of the random variable, namely m_{ij}.

Our first goal is to determine information about the matrix A, given the matrix $\widehat{A}$. This typically is of the form of information extraction (actual entries in A) or similarity (the angles between rows or columns of A). Corollary 6 implores us to take the following approach to this problem:[6]

1. Determine the largest k such that
$$\widehat{\sigma}_{k+1} - \widehat{\sigma}_k = \omega(\sqrt{m+n}).$$
2. Compute $\widehat{A}_k$ the optimal rank k approximation to $\widehat{A}$.
3. For any desired information about A, use the answer obtained by considering $\widehat{A}_k$ instead.

The confidence we have in the information output by this process is given in the following theorem.

THEOREM 7. *Let M, A and $\widehat{A}$ be defined as above, and let the notion of a good column or row be defined as in Corollary 6. Assume that the random variables in M have constant standard deviation and the separation $\widehat{\sigma}_k - \widehat{\sigma}_{k+1}$ of the matrix $\widehat{A}$ is $\omega(\sqrt{m+n})$. Then with probability $1 - o(1)$, for all good columns j,*
$$|A_k^{(j)}[\ell] - \widehat{A}_k^{(j)}[\ell]| \in o(1)$$
for all but $o(m)$ values of ℓ. Moreover, for all good columns i and j:

- $|A_k^{(j)} - \widehat{A}_k^{(j)}| \in o(|A_k^{(j)}|)$
- $|\angle(A_k^{(i)}, A_k^{(j)}) - \angle(\widehat{A}_k^{(i)}, \widehat{A}_k^{(j)})| \in o(1)$

Analagous statements hold for good rows.

PROOF. Consider the error between the matrices $\widehat{A}$ and A. Any particular entry in this error matrix $E = \widehat{A} - A$ is a random variable with the distribution
$$E[i, j] = m_{i,j} - \mathbf{E}(M_{i,j}).$$
Therefore, entries in this matrix are independent random variables with mean zero. From the second remark after Corollary 6, we can assume that $\sigma_k - \sigma_{k+1} = \omega(\sqrt{m+n})$. Additionally, since the deviations of $E[i, j]$ are bounded, we can apply Corollary 6 to conclude that the angles between good rows and columns and most entries in good rows and columns change by nominal amounts. □

It is straightforward to see that if $\widehat{A}$ is further corrupted by the addition of an error matrix whose entries have mean 0 and constant deviation, the same conclusions hold.

5.1.1 Discussion of Latent Semantic Indexing

The implications of Theorem 7 with respect to the use of LSI for information retrieval in documents should be fairly clear. Though the result is more general, for illustration purposes, consider the special case where the presence of term i in document j is an independent Bernoulli random variable with expectation $A[i, j]$. Documents of similar semantic composition will be generated from very similar probability distributions (i.e., corresponding document vectors in A will be nearly identical). Notice however, that even if the probability distributions for two columns are identical, the resulting documents obtained from the random rounding of those probabilities (columns in $\widehat{A}$) can be significantly different. Theorem 7 says that the similarity of the two documents in terms of the underlying generative model will be recovered in the transformation to $\widehat{A}_k$. Similarly, two documents with polysemous terms, say a document on the topic of the world-wide-web and a document on spider webs, will be well separated in $\widehat{A}_k$, despite the high probability of the word "web" appearing in each, if the underlying generative models for each are well separated.

Thus, $\widehat{A}_k$, the k-dimensional subspace produced by LSI when applied to the probabilistically generated matrix $\widehat{A}$, yields, with high probability, sharply defined clusters among documents of similar composition in terms of the underlying model A (with respect to the cosine measure). This of course assumes that A is well approximated by a rank k matrix itself. A rich rank k document generation model could be defined, for example, by assuming that there are k semantic categories or topics from which the documents are constructed, and letting A be the product of two matrices T and D, where T is an $m \times k$ matrix whose (i, ℓ)-th entry is the probability that a document on topic ℓ contains term i, and D is a $k \times n$ matrix whose (ℓ, j)-th entry is the fraction of document j on topic ℓ.

Theorem 7 thus helps explain the effectiveness of LSI as a technique for information retrieval and, in particular, for dealing with polysemy and synonymy.

5.1.2 Discussion of Kleinberg's Link Analysis

We described in Section 1.2 a generative model for the link structure of the web, defined by a pair of vectors h and a, such that there is an link from site i to site j ($\widehat{A}_{ij} = 1$) with probability $h_i a_j$. In this case the matrix $\mathbf{E}(M)$ is a rank one matrix. Theorem 7 tells us that computing the top left and right singular vectors of $\hat{A}$ will allow us to recover almost

[6] We have described all the algorithms in this paper using asymptotic notation. These can be converted to well-defined algorithms by replacing the asymptotic notation with appropriate (small) absolute constants.

all the entries in h and a, and hence the true importance of web sites.

5.2 Collaborative Filtering

We next consider the problem of mining an incomplete data set, as for example we would need to do for the collaborative filtering problem.

A Model for Collaborative Filtering

We model the collaborative filtering problem as follows.

- Let A represent a complete data set (an $m \times n$ matrix).
- Omit entry $A[i, j]$ with probability p_{ij}, (where omissions are independent of one another). We denote the resulting symbolic matrix A^*, so we have:

$$A^*[i, j] = \begin{cases} A[i, j] & \text{w.p. } p_{ij} \\ ? & \text{w.p. } 1 - p_{ij} \end{cases}$$

  where "?" is a placeholder indicating omitted data.
- Retrieve meaningful information about A, such as the values of the ?s.

A Collaborative Filtering Algorithm

We propose the following algorithm, called CF, for recovering the ?'s assuming that the omission probabilities p_{ij} are known. We will later describe a technique for estimating the p_{ij} values.

1. Define the matrix $\widehat{A}$ as follows:

$$\widehat{A}[i, j] = \begin{cases} A[i, j]/p_{ij} & \text{if } A^*[i, j] \neq \ ? \\ 0 & \text{if } A^*[i, j] = \ ? \end{cases}$$

2. Compute the SVD of $\widehat{A}$.
3. For each $[i, j]$ entry of A^* that is ?, output $\widehat{A}_k[i, j]$ as the estimate of $A[i, j]$.

Analysis of the Algorithm CF

We next show that the algorithm CF succeeds in reconstructing most of the missing entries of the matrix A.

THEOREM 8. *Let A, A^* and $\widehat{A}$ be defined as above, and let the notion of a good column or row be defined as in Corollary 6. Then if*

- *all $p_{ij} \in \Omega(1)$, and*
- *the separation $\widehat{\sigma}_k - \widehat{\sigma}_{k+1}$ of the matrix $\widehat{A}$ is $\omega(\sqrt{m+n})$,*

then with probability $1 - o(1)$, in any good column (or row), $1 - o(1)$ of the entries of A_k are reconstructed to within an additive error of $o(1)$.

PROOF. Once again, consider the error between the matrix $\widehat{A}$ and our original matrix A. Note that any particular entry in $E = \widehat{A} - A$ is a random variable with the distribution

$$E[i, j] = \begin{cases} \frac{A[i,j]}{p_{ij}} - A[i, j] & \text{w.p. } p_{ij} \\ -A[i, j] & \text{w.p. } 1 - p_{ij} \end{cases}$$

Thus, the matrix E is composed of independent random variables with mean zero and standard deviation $({p_{ij}}^{-1} - 1)^{1/2} A[i, j]$. Therefore, provided the p_{ij} are $\Omega(1)$ and the $A[i, j]$ are $O(1)$, Corollary 6 applies, yielding the theorem. □

Estimating The Omission Probabilities

The algorithm just presented assumed that the probability of retention of a particular parcel of data was well known to the algorithm. Such information is of course unlikely to be available in practice, as the only observation of the probabilities lies in the sampled values.

However, if we believe that the probabilities of retention themselves exhibit structure,[7] not dissimilar to the assumption on the nature of A, we will be able to recover a very good approximation to the matrix of probabilities, using the techniques of Section 5.1 as follows:

- Let $\widehat{P}$ be the matrix obtained by taking the matrix A^* and replacing all ?s with 0 and all present data with a 1.
- Compute the largest k such that

$$\sigma'_{k+1} - \sigma'_k = \omega(\sqrt{m+n}),$$

  where σ'_i is the i-th singular value of $\widehat{P}$.
- Compute $\widehat{P}_k$.
- Use $\widehat{P}_k[i, j]$ as an approximation to the omission probability p_{ij}.

Since $\widehat{P}$ is a random rounding of the matrix $P = \{p_{ij}\}$, we can retrieve "good" approximations (in the sense of Theorem 7) to the p_{ij}, using this technique.

It is important to note, however, that our proof of the performance of CF relies on having the precise omission probabilities. At this time, we do not know whether the same performance is guaranteed if CF takes as input only a very good approximation to the omission probabilities. Resolving this question is a key open problem.

5.3 The General Data Mining Model

It should be fairly obvious by now that very similar algorithms and theorems to those presented about the specific applications such as information retrieval and collaborative filtering can be provided for the general data mining model – the spectral techniques are robust in the presence of the combination of probabilistic data set generation, noise introduction and data omission. We leave the details to the full paper, but feel they should be fairly obvious from the discussion to this point. One key caveat, as just mentioned, is that we do not yet fully understand the effect of using the estimated omission probabilities in the matrix P (as just described) in the collaborative filtering step.

Perhaps the key limitation of the general data mining model is the fact that the random variables that form the

[7] This might be the case, if there are only a small number of possible reasons that an item is omitted from A. For example, if omission occurs because customer j is entirely unaware and has no opinion on product i, then a low rank P can be easily justified. If "awareness" is a function of advertising on TV and radio, and every person has TV/Radio listening habits and every product has a TV/Radio advertising budget then we would expect P to be of rank 2. The (i, j) entry of P would reflect the product of j's TV habits times i's TV budget plus the product of j's Radio habits times i's radio budget. Alternatively, the probabilities in P exhibit the right sort of structure if they are proportional to the values in A because, for example, people are more likely to buy things they like, or go to movies they like.

entries of M are assumed to be generated independently. Removing this restriction is an important direction for future research.

6. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proc. of the 1993 ACM SIGMOD Conference*, pp. 207-216, 1993.

[2] C. Basu, H. Hirsh and W. Cohen. Recommendation as Classification: Using Social and Content-Based Information in Recommendation. In *Proceedings of AAAI*, 1998.

[3] M. Berry, Z. Drmac, and E. Jessup. Matrices, Vector Spaces and Information Retrieval. In *SIAM Review Volume 41, Number 2*, pp. 335-362, 1999.

[4] M. Berry, S. Dumais, and G. O'Brien. Using linear algebra for intelligent information retrieval. In *SIAM Review 37(4)*, pp. 573–595, 1995.

[5] R.B. Boppana. Eigenvalues and Graph Bisection: An Average-Case Analysis. In *Proc. of 28th Annual FOCS*, pp. 280–285, 1987.

[6] Z. Furedi and J. Komlos. The eigenvalues of random symmetric matrices. Combinatorica 1:3, pp. 233-241, 1981.

[7] G.H. Golub and C. F. Van Loan. Matrix Computations, third Edition, the John Hopkins University Press, 1996.

[8] W. Hill, L. Stead, M. Rosenstein and G. Furnas Recommending and Evaluating Choices in A Virtual Community of Use. In *Proceedings of the CHI-95 Conference.*

[9] Jester `shadow.ieor.berkeley.edu/humor`

[10] R. Kannan, S. Vempala, A. Vetta, On Clusterings — Good, Bad and Spectral, Proceedings of 41st Annual IEEE Symposium on Foundations on Computer Science, 2000.

[11] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, pp. 668-677, 1998

[12] F. Korn, A. Labrinidis, Y. Kotidis, and C. Faloutsos. Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. In *VLDB* New York, NY, 1998.

[13] S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Recommendation Systems: A Probabilistic Analysis. In *Foundations of Computer Science*, pp. 664-673, 1998.

Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[14] C. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala. Latent Semantic Indexing: A Probabilistic Analysis. In *Proceedings of ACM Symposium on Principles of Database Systems*, 1997.

[15] U. Shardanand and P. Maes Social Information Filtering: Algorithms for Automating "Word of Mouth". . In *Proceedings of the CHI-95 Conference.*

[16] Sleeper `www.pmetrics.com/sleeper`

[17] G.W. Stewart. Matrix Algorithms, Volume 1: Basic Decompositions. Society for Industrial and Applied Mathematics, 1998.

7. APPENDIX

We will need the following lemma.

LEMMA 9. *Let $\{e_i\}$ be a set of n random vectors of length m where each coordinate of each vector is an independent random variable with mean 0 and bounded in absolute value by some constant c. Let U be an m by k matrix whose columns are orthonormal, where k is a constant. If f_i is the projection of e_i onto U, i.e., $f_i^T = e_i^T U$, then*

- *Each $|f_i|$ is $O(1)$ with probability $1 - o(1)$.*
- *All $|f_i|$ are $O(\sqrt{\log(n)})$ with probability $1 - o(1)$.*

PROOF. Let $U^{(\ell)}$ be the ℓth column of U. Define $X_k = \sum_{1 \le j \le k} e_i[j] U^{(\ell)}[j]$, so that $X_m = e_i^T U^{(\ell)}$. Then the sequence X_k is clearly a martingale, and $|X_k - X_{k-1}| \le |cU^\ell[k]|$. Applying Azuma's inequality (see [MR95], p. 92) we get that

$$Pr(|X_m| \ge \lambda) \le 2 \exp\left(\frac{-\lambda^2}{2c^2|U^{(\ell)}|^2}\right) = 2 \exp\left(\frac{-\lambda^2}{2c^2}\right).$$

For $\lambda \in \omega(1)$, we obtain the first claim. Letting $\lambda = C\sqrt{\log(n)}$, for an appropriately large C and applying a union bound, we obtain the second claim.

□

LEMMA 10. *Under the conditions of Corollary 6, with probability $1 - o(1)$, it is the case that*

- $|e_i| \in o(\sqrt{m+n})$
- $|e_i| \in O(1)$ *for all but $o(n)$ columns*

PROOF. Recall that

$$\begin{aligned} |e_i| &= |\widehat{U}_k^T E^{(i)}| \\ &\le |R^T U_k^T E^{(i)}| + |E_U^T E^{(i)}| \end{aligned}$$

Lemma 9 indicates that the norm of the first term is certainly $o(\sqrt{m+n})$. The second term is the product of a $o(1)$ norm matrix with a $O(\sqrt{m+n})$ norm vector, yielding a $o(\sqrt{m+n})$ vector. Together these give us the first assertion.

The second assertion is slightly more complicated. Again, Lemma 9 indicates that the first term is $O(1)$ for almost all columns. Concerning the second term, note the Frobenius norm of $E_U E$. Recall that

$$|E_U E|_F^2 = \sum_i \sigma_i^2$$

$E_U E$ is rank k, and thus has at most k singular values. Each of these singular values is $o(\sqrt{m+n})$, bounding $|E_U E|_F^2 \in o(m+n)$. As it is also the case that for any matrix M

$$\sum_i |M^{(i)}|^2 = |M|_F^2,$$

we can infer that the sum of squared length of columns in $E_U E$ is $o(m+n)$. Therefore the number of columns with length $\Omega(1)$ is $o(m+n)$ □