Visualizing Class Structure of Multidimensional Data

Inderjit S. Dhillon, Dharmendra S. Modha & W. Scott Spangler
net.Mining, GBIS/NA
IBM Almaden Research Center
San Jose, CA 95120-6099
Emails: {dhillon,dmodha,spangles}@almaden.ibm.com

Abstract

We consider the problem of visualizing multidimensional data that has been categorized into classes. Our goal in visualizing is to quickly absorb inter- and intra-class relationships.

Towards this end, we introduce *class-preserving projections* of the multidimensional data onto two-dimensional planes which can then be displayed on a computer screen. These class-preserving projections maintain the high-dimensional class structure, and are closely related to Fisher's linear discriminants. By displaying sequences of such two-dimensional projections and by moving continuously from one projection to the next, we can create illusions of smooth motion through a multidimensional display. Such sequences are termed *class tours*.

We illustrate the proposed ideas by various computer simulations on the classical Iris plant dataset and a text corpus of book reviews.

1 Introduction

It is often desirable to classify data into meaningful categories. The *Yahoo!* hierarchy of the World-Wide Web is a prime example of the value of classification [Yahoo!, 1998]. Given such classes, we may want to quickly absorb inter- and intra-class relationships. Visualization is an effective tool for this task, and in this paper we propose a scheme for visually understanding the relationships between classes.

We will assume that the data is embedded in high-dimensional Euclidean space R^d. The data may naturally occur in this form, or *vector-space* models of the underlying data may be formed, e.g., text documents or images may be treated as vectors in a multidimensional feature space, see [Salton and McGill, 1983] and [Flickner et al., 1995]. We will also not worry about how the data is classified — the classification may be done manually as in the *Yahoo!* hierarchy, or may be obtained by a clustering method such as the *k-means* algorithm [Duda and Hart, 1973, Hartigan, 1975]. We assume that we know the representation of all the data points as vectors in R^d, and their respective class labels.

Our main aim in this paper is to visually understand the spatial relationships between various classes in order to answer questions such as:

1. how well-separated are different classes?
2. what classes are similar or dissimilar to each other?
3. how coherent is a given class?

Answers to these questions would enable the data analyst to gauge the quality of the classification or quality of the feature space, and additionally, infer inter-class relationships that may not be part of the given classification.

In a nutshell, we are interested in *visual discriminant analysis*.

2 Class-Preserving Projections

Our main tool for visualizing multidimensional data will be linear projections onto 2-dimensional planes which can then be displayed on a computer screen. The challenge is in the choice of these planes (and the associated projections). *We want to choose those projections that best preserve inter-class distances.*

We first consider the canonical case where the data is divided into three classes. Let $x_1, x_2, \ldots x_n$ be all the d-dimensional data points, and m_1, m_2, m_3 denote the class-means (or class-centroids). Let w_1 and w_2 be an orthonormal basis of the candidate 2-dimensional plane of projection. The point x_i gets projected to the pair $(w_1^T x_i, w_2^T x_i)$ and consequently, the means m_j get mapped to

$$(w_1^T m_j, w_2^T m_j), \quad j = 1, 2, 3.$$

One way to obtain good separation of the projected classes is to maximize the difference between the projected means. This may be achieved by choosing vectors $w_1, w_2 \in R^d$ such that the objective function

$$\begin{aligned} C(w_1, w_2) &= \sum_{i=1}^{2} \{ |w_i^T (m_2 - m_1)|^2 + |w_i^T (m_3 - m_1)|^2 \\ &\quad + |w_i^T (m_3 - m_2)|^2 \} \end{aligned}$$

is maximized. The above may be rewritten as

$$\begin{aligned} C(w_1, w_2) &= \sum_{i=1}^{2} \{ w_i^T \{ (m_2 - m_1)(m_2 - m_1)^T \\ &\quad + (m_3 - m_1)(m_3 - m_1)^T \\ &\quad + (m_3 - m_2)(m_3 - m_2)^T \} w_i \} \\ &= w_1^T S_B w_1 + w_2^T S_B w_2 \\ &= \mathrm{trace}(W^T S_B W), \end{aligned}$$

where

$$\begin{aligned} W &= [w_1, w_2], \;\; w_1^T w_2 = 0, \;\; w_i^T w_i = 1, \; i = 1, 2, \;\; \text{and} \\ S_B &= (m_2 - m_1)(m_2 - m_1)^T + (m_3 - m_1)(m_3 - m_1)^T \\ &\quad + (m_3 - m_2)(m_3 - m_2)^T. \qquad (1) \end{aligned}$$

The positive semi-definite matrix S_B can be interpreted as the *inter-class or between-class scatter matrix*. Note that S_B has rank ≤ 2, since $m_3 - m_2 \in \mathrm{span}\{m_2 - m_1, m_3 - m_1\}$.

It is clear that the search for the maximizing w_1 and w_2 can be restricted to the column (or row) space of S_B. But as we noted above, this space is at most of dimension 2. Thus, in general, *the optimal w_1 and w_2 must form an orthonormal basis spanning the plane determined by the vectors $m_2 - m_1$ and $m_3 - m_1$* (in the degenerate case when S_B is of rank one, i.e., when m_1, m_2 and m_3 are collinear, w_1 should be in the direction of $m_2 - m_1$ while w_2 can be chosen to be any unit vector orthogonal to w_1).

Geometrically, the plane spanned by the optimal w_1 and w_2 is parallel to the plane containing the three class-means m_1, m_2 and m_3. It should be noted that under this projection, the distances between the projected means are *exactly equal* to the corresponding distances in the original d-dimensional space. Thus, we can say that inter-class distances are preserved, and we call such a projection a *class-preserving projection*.

We illustrate such a class-preserving projection on the famous Iris plant dataset in Figure 1. This dataset is 4-dimensional and contains 50 members in each of the three classes of plants : *Iris Setosa*, *Iris Versicolour* and *Iris Virginica*. Note that the three classes are well separated in this figure. For comparison, in Figure 2 we have

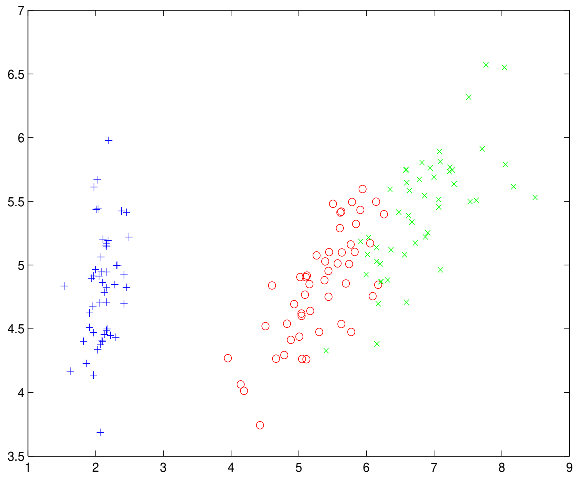

Figure 1: A *Class-Preserving Projection* of the Iris data

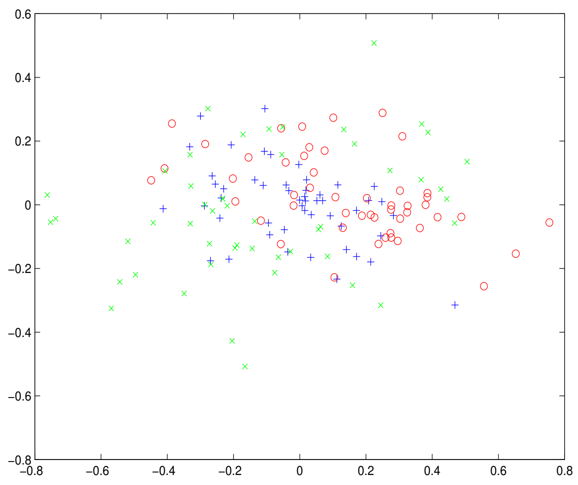

Figure 2: A poor projection of the Iris data

shown the projection onto a poorly chosen plane where the distinction between the classes is lost.

Projection schemes similar to ours have previously been used in classical linear discriminant analysis. In particular, our class-preserving projections are closely related to Fisher's linear discriminant and its generalizations. Fisher's linear discriminant differs slightly from our optimal since it also accounts for the scatter within each class. We have chosen to ignore such *within-class scatter* for computational reasons as we plan to deal with very large and high-dimensional datasets. For more details on these projection schemes, the reader is referred to [Fisher, 1936, Duda and Hart, 1973, Bryan, 1951, Kullback, 1959]. We emphasize that our class-preserving projections preserve distances between the three class-means, i.e., the *multidimensional scaling* *error* is zero for these class-means. For the more general problem of preserving inter-point distances between all the data points after projection, see [Kruskal, 1964, Kruskal, 1977, Duda and Hart, 1973].

2.1 Discriminating More than 3 Classes

In the above discussion, we arrived at a certain 2-dimensional projection that "best" preserves the relationship between 3 classes. But in most practical situations, we will encounter a greater number of classes. The *Yahoo!* hierarchy, for instance, segments web pages into thousands of categories. What projections will enable us to examine the interplay between these multitude of classes?

Clearly, one option is to examine the k classes three at a time by taking projections identical to those described in Section 2. There are a total of $\binom{k}{3}$ such 2-dimensional projections, each determined by three class-means. We show one such projection on an interesting text dataset, which we first describe.

The raw data is in the form of 1412 book reviews, each of which is a short text document containing about 25-100 words. For our treatment, we transform this data into a *vector-space model*, where each document is expressed as a numerical vector (of words) [Salton and McGill, 1983]. This is accomplished by the following steps:

1. convert all words to lower-case;
2. for each document, count the number of occurrences of each word;
3. discard (a) common words, such as "and", "the", (b) uncommon words, such as "sioux", and (c) one and two letter words, such as "be" and "to";
4. assign a unique integer i, $1 \leq i \leq d$, to each word, where d is the total number of remaining words;
5. represent each document by a d-dimensional vector x, whose i-th component x_i is the number of occurrences of word i in the document.

In addition, we also normalize each document vector to have unit Euclidean norm in order to remove the effects of the length of a document. Many other such scaling strategies have been studied; for more details, see [Salton and McGill, 1983] and [Salton and Buckley, 1988]. For this text corpus of book reviews, we used a total of $d = 867$ words.

After forming the vector-space model, we divided the book reviews into seven classes by invoking the k-means algorithm. We then formed the seven vectors $m_j - m$, where m is the mean of the entire dataset, and named class j by the words corresponding to the three largest values in $m_j - m$. These words are listed in the legend at the upper-right corner of Figure 3.

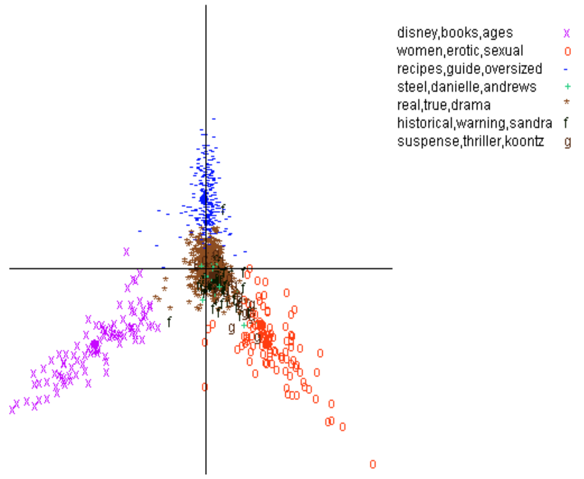

Figure 3: Projection discriminates 3 classes

Figure 3 shows the class-preserving projection that discriminates between the three classes named "disney, books, ages", "women, erotic, sexual" and "recipes, guide, oversized", respectively. The "recipes" class appears to be more coherent, while all the three classes seem equally distant from each other. Note that the proximity of some "erotic" reviews to "recipe" reviews is explained by books such as "Delicious Sex" by Gael Greene that "show how to spice up your love life".

In Figure 3, the three discriminated classes are well-separated while the four other classes appear intermingled near the origin. This is not surprising since the projections were expressly chosen to preserve the differences only between three classes. In general, we would like to differentiate more than three classes in the same view.

More precisely, we want to obtain a 2-dimensional projection that best discriminates the q classes with means $m_1, m_2, \ldots m_q$, each containing $n_1, n_2, \ldots n_q$ data points, respectively. Taking an approach similar to that in Section 2, we can formulate this problem as the search for orthonormal $w_1, w_2 \in R^d$ that maximizes

$$C(w_1, w_2) = \mathrm{trace}(W^T S_B W), \tag{2}$$

where

$$W = [w_1, w_2], \ \ w_1^T w_2 = 0, \ \ w_i^T w_i = 1, \ i = 1, 2, \ \text{and}$$

$$S_B = \sum_{i=2}^{q} \sum_{j=1}^{i-1} n_i n_j (m_i - m_j)(m_i - m_j)^T. \tag{3}$$

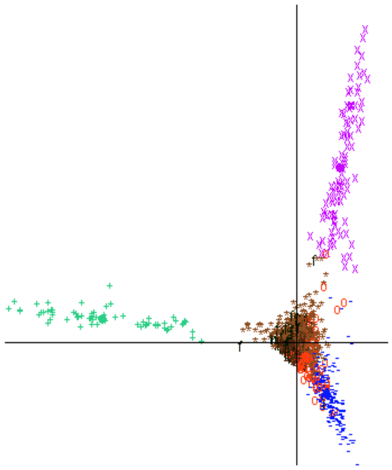

Figure 4: Projection discriminates 4 classes

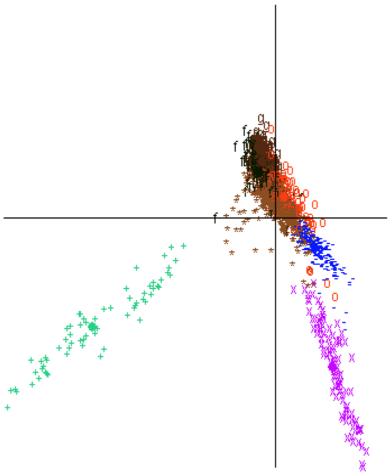

Figure 5: Projection discriminates 7 classes

Note that the positive semi-definite matrix S_B has rank $\leq q-1$ since the vectors $m_i - m_j$, $j \neq 1$ are linearly dependent on the $q-1$ vectors $m_i - m_1$, $i = 2, \ldots, q$. It is well known that *the vectors w_1 and w_2 that maximize the objective function in (2) are the eigenvectors (or principal components) corresponding to the two largest eigenvalues of S_B.*

In Figure 4, we show such a projection that preserves inter-class differences between four of the seven classes. The "Danielle Steel" cluster is seen to be very coherent and well-separated from the "disney", "recipes" and "erotic" clusters. Note again that the three remaining clusters form a "blob" near the origin.

The reader might have noticed the extra factor $n_i n_j$ in (3) that was not present in (1). In general, when $q > 3$, the distances between the q means will not be *exactly* preserved after projection. By weighting each term in (3) by the factor $n_i n_j$, we are placing greater emphasis on the distances between the means of the *larger* classes.

The matrix S_B as given above in (3) is the sum of $\binom{q}{2}$ rank-one matrices. We can show that S_B can be expressed more compactly as the sum of q rank-one matrices. In particular,

$$S_B = n^{(q)} \sum_{i=1}^{q} n_i (m_i - m^{(q)})(m_i - m^{(q)})^T, \quad (4)$$

where

$$\begin{aligned} n^{(q)} &= n_1 + n_2 + \cdots + n_q, \quad \text{and} \\ m^{(q)} &= \frac{1}{n^{(q)}} (n_1 m_1 + n_2 m_2 + \cdots n_q m_q) \end{aligned}$$

is the mean of all the data points in the q classes under consideration. This alternate expression for S_B is clearly better for computational purposes. The interested reader can look at the Appendix for a proof of the equivalence of formulae (3) and (4).

Thus, given k classes we have a mechanism for viewing the inter-class relations between q of them. There are a total of

$$\sum_{q=3}^{k} \binom{k}{q} = 2^k - \frac{k(k+1)}{2} - 1$$

such informative class projections. However, as q gets larger, the distinction between classes starts getting blurred in these views. Figure 5 gives a crowded view that attempts to discriminate all 7 classes. Such crowding is inevitable since our linear projections are limited to 2-dimensional planes. Ideally, we would like to "see" the data projected onto the entire $(q-1)$-dimensional subspace spanned by the columns of the matrix S_B of (3). However, we are limited to 2-dimensional computer displays. In Section 4, we propose a way to overcome this limitation. But before we do so, we look at an enhancement to each individual projection.

3 Class-Similarity Graphs

Two-dimensional linear projections are a continuous transformation of the data; hence, two points which are close in R^d will remain close in each of these projections. However, two points which are close in a 2-dimensional projection need not be close in R^d. This is an inevitable loss of information in projecting high-dimensional data to lower dimensions. To mitigate this information loss, we propose the addition of *class-similarity graphs* to our 2-dimensional plots.

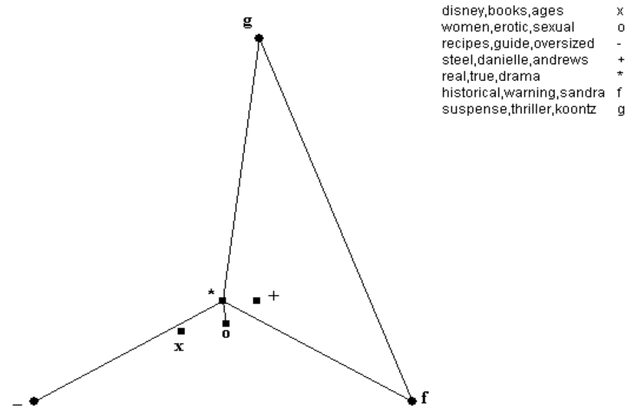

Figure 6: A class-similarity graph

A class-similarity graph is described as follows [Duda and Hart, 1973, pp. 238]. The vertices of this graph are the class centroids $m_1, m_2, \ldots, m_k$, and there is an edge between the means (or vertices) m_i and m_j if

$$d_2(m_i, m_j) \leq \tau,$$

where d_2 denotes the Euclidean distance, while τ is a user-controlled threshold parameter. If τ is very large, then all centroids will be connected. On the other hand, if τ is very small, then no centroids will be connected. It is thus intuitively clear that the choice of this threshold parameter is very important in revealing similarities (or dissimilarities) between the class-means. To obtain "natural" connections between the centroids, τ will have to be greater than typical distances between related classes but less than typical distances between unrelated ones.

We can display the class-similarity graph by overlaying it on our 2-dimensional class-preserving projections. Figure 6 shows such a graph on the projection that discriminates between the three classes named "suspense", "recipes" and "historical". Note that to show the similarity graph clearly in this figure, we have shown only the seven centroids and removed the individual data points. In the above figure, the "suspense", "historical" and "drama" centroids form a clique, while the "erotic" and "disney" centroids are disconnected even though they lie next to each other in this projection. Thus, the class similarity graph adds another *information dimension* to linear projections.

4 Class Tours

Thus far we have limited ourselves to static 2-dimensional photographs of the dataset, each of which conveys some limited class relationships. However, projection onto the entire $(k-1)$-dimensional subspace spanned by the vectors $m_i - m_1$, $i = 2, \ldots, k$, *exactly* preserves the distances between all the k class-means and thus contains "all" the inter-class information. Ideally, we would like a mechanism for viewing this *inter-class differentiating* or *class-preserving* linear subspace.

In order to simulate such multidimensional displays, [Asimov, 1985, Asimov and Buja, 1985] proposed the use of motion graphics. Specifically, they introduced the concept of *tours* which are sequences of 2-dimensional projections interspersed with a number of intermediate projections. These intermediate projections are obtained by interpolation, and thus tours create an illusion of continuous smooth motion through a multidimensional display. Building on the idea of tours, we propose *class tours* which are sequences of our class-preserving 2-dimensional projections, and are an effective tool to "view" the $(k-1)$-dimensional class-preserving subspace. Metaphorically speaking, a class tour constructs a dynamic, global "movie" from a number of static, local photographs.

The basic idea behind class tours is simple: choose a target 2-dimensional projection from the (subset of) nearly 2^k class-preserving projections, move smoothly from the current projection to this target, and continue. The main question of interest is in the choice of the orthonormal basis used for viewing each 2-dimensional projection so that the motion appears smooth. [Asimov, 1985, Asimov and Buja, 1985] have proposed using geodesic interpolation paths between the current and the target planes. These geodesic paths are simply a rotation in the (at most) 4-dimensional linear subspace containing both the current and the target 2-d planes. Various smoothness properties of such geodesic paths are explored in great detail in [Buja et al., 1998]. For a description of various implementation details, see [Hurley and Buja, 1990, section 2.2.1].

5 Conclusions

In this paper, we have proposed the use of *class-preserving projections* for *visual discriminant analysis*. These projections satisfy a certain optimality criterion that attempts to preserve distances between the class-means. Our projections are similar to Fisher's linear discriminants that are commonly used in classical discriminant analysis. We also use *class tours* which enable us to view sequences of our class-preserving projections, interspersed with interpolating projections that create an illusion of smooth motion through a multidimensional subspace. Class tours allow the touring of all the data points in an informative *inter-class differentiating* linear subspace. Thus, our visualization scheme allows us to fully capture the inter-class structure of complex, multi-dimensional data.

The ideas presented in this paper have been implemented in Java. In fact, Figures 3 through 6 have been generated using our implementation.

Appendix

Proposition. *The matrices S_B given in equations (3) and (4) are identical.*

Proof. Expanding each term in equation (3) and using symmetry in i and j, we get

$$\begin{aligned}
S_B &= \sum_{i=1}^{q} \sum_{j=1, j\neq i}^{q} n_i n_j (m_i m_i^T - m_i m_j^T) \\
&= \sum_{i=1}^{q} \{ n_i(n_1 + n_2 + \cdots + n_q) m_i m_i^T - n_i^2 m_i m_i^T \} \\
&\quad - \sum_{i=1}^{q} \sum_{j=1, j\neq i}^{q} n_i n_j m_i m_j^T, \\
&= n^{(q)} \sum_{i=1}^{q} n_i m_i m_i^T - \sum_{i=1}^{q} \sum_{j=1}^{q} n_i n_j m_i m_j^T, \\
&= n^{(q)} \left(\sum_{i=1}^{q} n_i m_i m_i^T - n^{(q)} m^{(q)} m^{(q)T} \right), \\
&= n^{(q)} \sum_{i=1}^{q} n_i (m_i - m^{(q)})(m_i - m^{(q)})^T. \quad \square
\end{aligned}$$

Acknowledgements

The authors are grateful to Mark Plutowski for providing the text corpus of book reviews.

References

[Asimov, 1985] Asimov, D. (1985). The grand tour: A tool for viewing multidimensional data. *SIAM J. Sci. Stat. Comput.*, 6(1):128–143.

[Asimov and Buja, 1985] Asimov, D. and Buja, A. (1985). Grand tour methods: an outline. In *Proceedings of the 17th Symposium on Interface of Computer Science and Statistics*.

[Bryan, 1951] Bryan, J. G. (1951). The generalized discriminant function: mathematical foundation and computational routine. *Harvard Educ. Rev.*, 21:90–95.

[Buja et al., 1998] Buja, A., Cook, D., Asimov, D., and Hurley, C. (1998). Dynamic projections in high-dimensional visualization: Theory and computational methods. *Journal of Computational and Graphical Statistics*. to appear.

[Duda and Hart, 1973] Duda, R. O. and Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley.

[Fisher, 1936] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, 7(Part II):179–188. also in *Contributions to Mathematical Statistics* (John Wiley, New York, 1950).

[Flickner et al., 1995] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by image and video content: the QBIC system. *IEEE Computer*, 28(9):23–32.

[Hartigan, 1975] Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.

[Hurley and Buja, 1990] Hurley, C. and Buja, A. (1990). Analyzing high-dimensional data with motion graphics. *SIAM J. Sci. Stat. Comput.*, 11(6):1193–1211.

[Kruskal, 1964] Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129.

[Kruskal, 1977] Kruskal, J. B. (1977). Multidimensional scaling and other methods for discovering structure. In Enslein, K., Ralston, A., and Wilf, H. S., editors, *Statistical Methods for Digital Computers*, pages 296–339. Wiley.

[Kullback, 1959] Kullback, S. (1959). *Information Theory and Statistics*. John Wiley, New York.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 4(5):513–523.

[Salton and McGill, 1983] Salton, G. and McGill, M. J. (1983). *Introduction to Modern Retrieval*. McGraw-Hill Book Company.

[Yahoo!, 1998] Yahoo! (1998). http://www.yahoo.com.