Distance Measures and Smoothing Methodology for Imputing Features of Documents

Andrey FEUERVERGER, Peter HALL, Gelila TILAHUN, and Michael GERVERS

We suggest a new class of metrics for measuring distances between documents, generalizing the well-known resemblance distance. We then show how to combine distance measures with statistical smoothing to develop techniques for imputing missing features of documents. We treat in detail the case where these features are continuous variates, but we note that our methods can be adapted to settings where the features are ordered or unordered categorical variates (e.g., the names of potential authors of the documents). The results of applying our ideas to the dating of medieval manuscripts are briefly summarized.

Key Words: Bandwidth; Correspondence distance; Cross-validation; Dating; Kernel; Resemblance distance; Shingle.

1. INTRODUCTION

There is a large literature on measuring relationships among documents, much of it in the context of searching for and filtering pages of the World Wide Web. It includes distinctly statistical contributions; for example, work by Cutting, Karger, Pedersen, and Tukey (1992) on cluster-based approaches, and "distance"- or "association"-based techniques such as those discussed by Berry, Hendrickson, and Raghavan (1995); Broder, Glassman, Manasse, and Zweig (1997); Broder (1998); Berry and Browne (1999); and by authors of articles or chapters in the collections edited by Berry (2001, 2003) and Djeraba (2002). Section 2 shows that the distance measures suggested by Broder (1998) are part of a significantly larger class. Section 3 suggests ways in which such measures can be combined with statistical

Andrey Feuerverger is TKKK, Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario M5S 3G3, Canada (E-mail: andrey@utstat.toronto.edu). Peter Hall is TKKK, Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia. Gelila Tilahun is TKKKK, Department of Statistics, University of Toronto, 100 St George Street, Toronto, Ontario M5S 3G3, Canada. Michael Gervers is TKKK, Department of History, University of Toronto, 100 St George Street, Toronto, Ontario M5S 3G3, Canada.

Journal of Computational and Graphical Statistics, Volume 14, Number 2, Pages 1–9
DOI: 10.1198/106186005X47291

smoothing and used to impute missing attributes, such as dates, in documents.

The problem that motivated this work involved the dating, or "calendaring," of medieval manuscripts from the 11th to the 14th centuries. Training data gave dates for some of the documents, but many others were uncalendared. (We use the terms manuscript and document interchangeably, and each such instance can be regarded as a purposeful ordered sequence of words.) The methodology developed in Sections 2 and 3 allowed us to regress "date" on "manuscript" in an adaptive, nonparametric way, and thereby to estimate the date of virtually any given manuscript. Some of the results of this analysis will be briefly summarized at the end of Section 3.

More generally, the quantity that is regressed on "document" can be multivariate and continuous, ordered categorical, or unordered categorical. Examples, for which we admittedly do not have adequate data except in the medieval manuscript case, include the ordering of different drafts of a document, some of which would be dated and thus form the training dataset; assignation of authorship to documents, for which training data would consist of documents for which authorship was known unambiguously; allocation of "preferences" of readers for different writing styles, where the training data would comprise documents for which particular readers had known preferences; and imputation of opinions or marks which readers or examiners might have accorded documents had they read them. Section 3.2 discusses potential adaptations of our method, enabling us to solve problems such as these.

2. DISTANCE MEASURES FOR DOCUMENTS

2.1 SURVEY OF MEASURES OF ASSOCIATION AND DISTANCE

Linguistic and lexicographical contributions to association measures, often for recorded speech rather than written documents, include those developed by Reinert (1990) and Benzécri (1991). They involve dissecting the speech (or document) into "context units" of 10 to 20 words, not unlike the "shingles" introduced in Section 2.2. The presence or absence of individual words in the context units is then compared for different speakers (or writers), using a chi-squared goodness-of-fit statistic and correspondence analysis. Values of the statistic are used to compare word usage frequencies for different writers, or speakers, and these are then interpreted. This approach need not be particularly computer-intensive, after the initial chi-squared statistics have been produced. This reflects the fact that the objective is usually interpretation, rather than classification, in applications of these methods.

Alternative techniques adopt a higher-order approach to the problem, using (instead of literal word-matching) vector-space information retrieval methods. These employ relatively abstract conceptual indices instead of individual words, and singular value decompositions of large, sparse, term $\times$ text-object matrices, to estimate structure in word usage across documents. See, for example, Deerwester, Dumais, Landauer, and Harshman (1990); Berry, Dumais, and O'Brien (1995); Berry and Browne (1999, chaps. 3 and 4); and Simon and

Ding (2001). Berry, Hendrickson, and Raghavan (1996) suggested a related approach based on identification of document clusters, and gave a detailed account of the literature. Berry, Raghavan, and Zhang (2001) discussed preprocessing methods in the context of vector-space information retrieval. Husbands, Simon, and Ding (2001) gave details of the application of singular value decomposition for document retrieval. Further articles on similar topics can be found in the proceedings edited by Berry (2001, 2003).

Association measures connected with "angularity" are commonly used in document-based settings, and are sometimes referred to as "angle similarity" or "cosine" measures. Examples were given by Berry and Browne (2001). See, for example, their discussion of vector space models in chapter 3, and of query matching in section 4.2.2. Berry and Browne (2001, chap. 3) compared different ways of representing "frequencies" (e.g., of word usage) in documents, in terms of their application to vector space models for document association.

Properties of conventional distance functions (e.g., L^1, L^2 and supremum distance) were discussed by Djeraba (2002, sec. 4.5), against the background of applications to color-based image retrieval. In work of this nature there are sometimes important differences between metric distances and human perceptions of distance. See, for example, Santini and Jain (1999), who developed a similarity measure that exhibits features which match experimental findings in humans. See also Ashby and Perrin (1988).

Further techniques, starting from the work of Broder cited in Section 1, are discussed in the following.

2.2 CLASSES OF DISTANCE MEASURES BASED ON CORRESPONDENCES

Suppose a document $\mathcal{D}$ consists of n words in the order $w_1, \ldots, w_n$. To indicate that order is important we use sequence notation, $\mathcal{D} = (w_1, \ldots, w_n)$, to denote $\mathcal{D}$. Of course, the words are not necessarily distinct. A *shingle* S of order k is a consecutive sequence of k words, that is, $S = (w_{t+1}, w_{t+2}, \ldots, w_{t+k})$, where $0 \le t \le n-k$. Let $\mathcal{S}_k(\mathcal{D}) = \{S_{k1}, \ldots, S_{k,N(k)}\}$ be the set of distinct shingles of order k obtainable from $\mathcal{D}$. Then $1 \le N(k) \le n-k+1$.

Given two documents $\mathcal{D}_i$ and $\mathcal{D}_j$, $\mathcal{S}_k(\mathcal{D}_i) \cap \mathcal{S}_k(\mathcal{D}_j)$ denotes the set of distinct shingles of order k that are contained in both documents. The kth order *resemblance*, $\mathrm{res}_k(i,j)$, between $\mathcal{D}_i$ and $\mathcal{D}_j$ is the proportion of shingles, out of the set of all kth order shingles in $\mathcal{D}_i$ and $\mathcal{D}_j$, that are contained in both $\mathcal{D}_i$ and $\mathcal{D}_j$:

$$\mathrm{res}_k(i,j) = \frac{\|\mathcal{S}_k(\mathcal{D}_i) \cap \mathcal{S}_k(\mathcal{D}_j)\|}{\|\mathcal{S}_k(\mathcal{D}_i) \cup \mathcal{S}_k(\mathcal{D}_j)\|},$$

where $\|\mathcal{S}\|$ denotes the number of elements of a finite set $\mathcal{S}$. The kth order *resemblance distance* is $d_k(i,j) = 1 - \mathrm{res}_k(i,j)$. (e.g., Broder et al. 1997).

Although these concepts are important, their scope is limited. To place them into a broader context we introduce the notion of *correspondences*, which we define as follows. Let $n(i)$ and $n(j)$ denote the numbers of words in $\mathcal{D}_i$ and $\mathcal{D}_j$, respectively, and suppose $\mathcal{D}_i = (w_1, \ldots, w_{n(i)})$. Given that $\mathcal{S}_k(\mathcal{D}_i) \cup \mathcal{S}_k(\mathcal{D}_j)$ contains $N(i,j)$ elements, and that $1 \le \ell \le N(i,j)$, let $\nu_\ell(i,j)$ denote the number of times the ℓth element of $\mathcal{S}_k(\mathcal{D}_i) \cup \mathcal{S}_k(\mathcal{D}_j)$ is

included in the set of $n(i)-k+1$ sequences $(w_{t+1}, w_{t+2}, \ldots, w_{t+k})$, for $0 \le t \le n(i)-k$, of consecutive words in $\mathcal{S}_k(\mathcal{D}_i)$. In this notation, $N(i) = \sum_\ell \nu_\ell(i,j)$ and $N(j) = \sum_\ell \nu_\ell(j,i)$ are the numbers of (not necessarily distinct) shingles of order k in $\mathcal{D}_i$ and $\mathcal{D}_j$, respectively. (For simplicity we have suppressed the notation k.) Let $f(u,v)$ be a bivariate, nonnegative function. A kth order *correspondence* between $\mathcal{D}_i$ and $\mathcal{D}_j$ is defined to be

$$\mathrm{corr}_k(i,j) = \frac{\sum_{1 \le \ell \le N(i,j)} f\{\nu_\ell(i,j), \nu_\ell(j,i)\}}{F\left\{\nu_1(i,j), \ldots, \nu_{N(i,j)}(i,j), \nu_1(j,i), \ldots, \nu_{N(i,j)}(j,i)\right\}},$$

where the norming function F is chosen so that for all choices of $\mathcal{D}_i$ and $\mathcal{D}_j$, (a) $0 \le \mathrm{corr}_k(i,j) \le 1$ and (b) $\mathrm{corr}_k(i,j) = 1$ whenever $\mathcal{D}_i = \mathcal{D}_j$. The *correspondence distance* associated with the correspondence $\mathrm{corr}_k(i,j)$ is defined to be $d_k(i,j) = 1 - \mathrm{corr}_k(i,j)$.

2.3 EXAMPLES OF CORRESPONDENCE DISTANCES

We give four examples: (1) $f(u,v) = u^\alpha\, v^\alpha$ and

$$F(\vec{u}, \vec{v}) = F\left(u_1, \ldots, u_{N(i,j)}, v_1, \ldots, v_{N(i,j)}\right) \equiv \left(\sum_{\ell=1}^{N(i,j)} u_\ell^{2\alpha}\right)^{1/2} \left(\sum_{\ell=1}^{N(i,j)} v_\ell^{2\alpha}\right)^{1/2},$$

for $0 < \alpha < \infty$; (2) define f identically to (1) but put

$$F(\vec{u}, \vec{v}) = \sum_{\ell=1}^{N(i,j)} \left(u_\ell^{2\alpha} + v_\ell^{2\alpha} - u_\ell^\alpha\, v_\ell^\alpha\right);$$

(3) $f(u,v) = I(u > 0, v > 0)$ and $F(\vec{u}, \vec{v}) = \sum_\ell (u_\ell + v_\ell) = N(i,j)$; (4) $f(u,v) = \min(u,v)$ and $F(\vec{u}, \vec{v}) = \min(\sum_\ell u_\ell, \sum_\ell v_\ell) = \min\{N(i), N(j)\}$.

The correspondence distances of Types (1)–(4) all satisfy the normalization conditions (a) and (b) imposed at the end of the previous section. Special cases of Type (1) were treated by Quang, James, James, and Levina (1999) and Zhang and Korfhage (1999). However, it can be shown that Type (1) correspondence distances fail to satisfy the triangle inequality, and therefore do not define a metric. On the other hand, Type (2) correspondence distances are metrics; see Feuerverger, Hall, Tilahun, and Gervers (2004) for proofs. We suggest, therefore, that Type (2) correspondence distances be employed instead of Type (1). Although metric properties are not always critical to algorithm development, they are a considerable aid to intuition when interpreting the meaning of "distance" among documents. In particular, the triangle inequality is particularly important for interpreting index operations (Djeraba 2002, p. 70). Type (2) correspondence distances differ from Type (1) only in the method of normalization, and in particular capture the main geometric aspects of relationship that motivate Type (1) correspondences.

Type (3) correspondence distances may be interpreted as the limit, as $\alpha \downarrow 0$, of Type (2). Type (4) correspondence distances are related to L_1, or variational, distance. The definition of Type (2) correspondence distances can be substantially generalized. For example, the

metric property is enjoyed by the more general definition in which $f(u, v) = g(u)\, g(v)$ for a strictly monotone, nonnegative function g, and

$$F(\vec{u}, \vec{v}) = \sum_{\ell=1}^{N(i,j)} \left\{ g(u_\ell)^2 + g(v_\ell)^2 - g(u_\ell)\, g(v_\ell) \right\} .$$

In particular, the choice $g(u) = \log(u + 1)$ gives low weight to shingle-count frequencies, but unlike resemblance distance does not ignore the frequencies altogether.

2.4 DOCUMENT PREPARATION

In the analysis of documents discussed earlier and in Section 3, and particularly in the practical application mentioned in Section 3.3, a first step is to remove all punctuation. Numbers are replaced by simply "#" before shingling, so different numbers are not distinguished. However, shingling distinguishes capitalized from noncapitalized words; for example, "regis" is regarded as different from "Regis."

3. SMOOTHING AMONG DISTANCE MEASURES

3.1 SMOOTHING METHODS FOR DOCUMENTS

Suppose we are using r measures of distance between document pairs. These r distances may represent measures based on r different shingle orders, or may involve different types of distance measures, for example. We denote the kth distance from $\mathcal{D}_i$ to $\mathcal{D}_j$ by $d_k(i, j)$, for $1 \le k \le r$. We shall smooth on the distances, assuming they capture the principal ways in which documents differ. Depending on the contexts of documents, for example, on how much the dataset has been refined before analysis, there may be significant information from other, ordered variables such as document length or the simple frequencies of certain key words. These too can be incorporated into our smoothing algorithm, by making obvious changes to the methodology discussed in the following.

Let K denote a nonnegative, nonincreasing function defined on the positive half-line. Suppose an attribute of interest, t_i say (e.g., the date), is missing from document $\mathcal{D}_i$; but that all documents $\mathcal{D}_j$, for $j \in \mathcal{J}$ say, have respective known attributes t_j. We suggest a nonparametric approach to imputing, or estimating, the attribute of $\mathcal{D}_i$. The kernel weight ascribed to $\mathcal{D}_j$, based on its nearness to $\mathcal{D}_i$, is denoted by

$$a(i, j) = a(i, j | h_1, \ldots, h_r) = \prod_{k=1}^{r} K\{d_k(i, j)/h_k\} , \tag{3.1}$$

where $h_1, \ldots, h_r$ denote bandwidths. We can estimate the unknown attribute, t_i, of $\mathcal{D}_i$ by kernel-based, local-constant regression on the distances. Specifically, $\hat{t}_i$ is the value of t

that minimizes $\sum_{j\in\mathcal{J}} (t_j - t)^2\, a(i,j)$, and so is given by

$$\hat{t}_i = \left\{ \sum_{j\in\mathcal{J}} t_j\, a(i,j) \right\} \Big/ \left\{ \sum_{j\in\mathcal{J}} a(i,j) \right\} . \tag{3.2}$$

Theoretical properties of this estimator are described by Feuerverger et al. (2004).

We suggest using cross-validation to select bandwidths, as follows. Let $\mathcal{K}$ denote the union, over $1 \le k \le r$, of the set of all indices $j \in \mathcal{J}$ such that $d_k(i,j)$ is among the m largest values of that quantity. Here m would be an appropriately small fraction of the total number of documents with known attributions. For each $j' \in \mathcal{K}$, and each vector $(h_1, \ldots, h_r)$ of bandwidths, we define

$$\hat{t}_{j'} = \hat{t}_{j'}(h_1, \ldots, h_r) \equiv \operatorname*{argmin}_t \sum_{j\in\mathcal{J},\, j\ne j'} (t_j - t)^2\, a(j', j | h_1, \ldots, h_r) .$$

Put

$$(\hat{h}_1, \ldots, \hat{h}_r) = \operatorname*{argmin}_{(h_1,\ldots,h_r)} \sum_{j'\in\mathcal{K}} \left\{ t_{j'} - \hat{t}_{j'}(h_1, \ldots, h_r) \right\}^2 . \tag{3.3}$$

Then $(\hat{h}_1, \ldots, \hat{h}_r)$ is our empirical choice of the bandwidth vector.

This procedure may be iterated, for example, by replacing $(h_1, \ldots, h_r)$ with $(\hat{h}_1, \ldots, \hat{h}_r)$, replacing $\mathcal{K}$ with the set $\mathcal{K}(i|\hat{h}_1, \ldots, \hat{h}_r)$ of indices $j \in \mathcal{J}$ such that $a(i, j|\hat{h}_1, \ldots, \hat{h}_r) \ne 0$, and cycling through the algorithm once more. In this case a kernel weight factor, depending on the bandwidth chosen at the previous step, can be incorporated into the series at (3.3). There is also a global form of cross-validation, in which the series above are taken over all choices of i as well as over their respective summands.

Cross-validation may also be used to estimate the mean squared error, $s(i)^2$ say, of the estimator $\hat{t}_i$, as follows. On this occasion we take $\mathcal{K}$ to be the set $\mathcal{K}(i|h_1, \ldots, h_r)$, where $(h_1, \ldots, h_r)$ is now the bandwidth vector used to calculate $\hat{t}_i$. Compute $\hat{t}_j$ for each $j \in \mathcal{K}$, again using the bandwidths $(h_1, \ldots, h_r)$, and take

$$\hat{s}(i)^2 = \left\{ \sum_{j\in\mathcal{K}} \left(t_j - \hat{t}_j\right)^2 a(i, j | h_1, \ldots, h_r) \right\} \Big/ \left\{ \sum_{j\in\mathcal{K}} a(i, j | h_1, \ldots, h_r) \right\}$$

to be our estimator of $s(i)^2$.

3.2 APPLICATIONS TO GENERAL ATTRIBUTES

Applications to multivariate continuous attributes, for example pairs (date, length) for undated document fragments, require only minor modifications of the univariate case discussed earlier. In particular, if the attributes have m components then, in place of the weight at (3.2), we use one based on a product over components as well as over correspondence order (the latter is k in (3.1)).

The case where attributes are categorical is more challenging. For ordered categorical variates, such as "experience" or "ability" of author, the different categories may be arranged consecutively on a line, perhaps with their relative distances apart adjusted to reflect prior notions of closeness. Regression of the attribute variables on "document" would be implemented in the continuum, and each imputed point on the line would be interpreted as the discrete attribute to which it was closest.

Unordered categorical variates, for example m different authors or m different essay markers, could be placed at the m vertices of a simplex in $(m-1)$-dimensional Euclidean space. If necessary the edge lengths of the simplex could be distorted to reflect prior beliefs about relative distances between attribute pairs. For example, distances between authors might be assigned after examining various measured distances between pairs of manuscripts having known authorship. Of course, the simplex geometry places restrictions on possible distortions. Much as in the ordered categorical case, regression of attributes would be undertaken in the continuum and then rendered discrete by shrinking to nearest simplex vertices.

Finally, we remark that our smoothing methods may, in principle, be applied to contexts such as dating archaeological sites and artifacts. To do so, we must first associate with them vectors of appropriate attributes and measures specific to such entities, and also specify suitable distances among such vectors.

3.3 SUMMARY OF APPLICATION TO CALENDARING PROBLEM

We applied our method to data on manuscripts written between the 11th and 15th centuries. These manuscripts were charters concerning property holdings or transfers in the county of Essex, England. Many were taken from entries in the *Hospitaller Cartulary* of 1442, and are essentially land deeds involving the Order of the Hospital of St. John of Jerusalem. These manuscripts are part of a database assembled by University of Toronto historian Michael Gervers. Further details of the context, the data, and the results were given by Gervers (1982, 1996) and Feuerverger et al. (2004). In summary, there were 3,353 dated manuscripts in the training sample, and some 5,000 undated manuscripts. The dates ranged from 1089 to 1466, and followed an approximately bell-shaped distribution.

By means of random sampling the training sample was divided into three disjoint groups. The first, consisting of 3,034 documents, served as a training set; the second, of 167 documents, was used for validation; and the third, of 152 documents, was set aside to serve as a test set. The decision to set aside a validation subset, rather than use leave-one-out validation, was made to reduce computational labor.

We report here only the results for shingles of size 2. (In practice, the shingle sizes used would be determined on the basis of the predictive ability.) In this setting the cross-validation method proposed in Section 3.1 suggests taking $(m, h) = (5, 6.7 \times 10^{-3})$, which gave a mean absolute error of approximately 11 years. The exact result was 11.1 years in the case of resemblance distance, and depended only a little on the choice of distance measure. This may be compared to a mean absolute error of 36.6 years which is obtained if the

mean year of the training documents (which was 1245.8) is used to estimate the date for each document in the training set. We note that the mean absolute error was also quite robust against variation in the value of m, as well as in the bandwidth values used. For this particular dataset, shingles of size 2 were found to be somewhat more informative than shingles of size 1 or 3, and little additional accuracy was gained in combining shingle size 2 with these other shingle sizes. Very slight bias effects were noted for dates at the edge of the range, due to the fact that for such manuscripts close matches cannot exist. For the sake of brevity we do not discuss results of applications to the 5,000 uncalendared manuscripts, for which errors cannot be accurately described because the true dates are unknown.

[Received July 2002. Revised July 2004.]

REFERENCES

Ashby, F. G., and Perrin, N. A. (1988), "Toward a Unified Theory of Similarity and Recognition," *Psychological Review*, 95, 124–150.

Benzécri, J. P. (1981), *Pratique de l' Analyse des Données; Linguistique et Lexicologie*, Paris: Dunod.

Berry, M. W. (2001), *Computational Information Retrieval*, Philadelphia: SIAM.

——— (2003), *Survey of Text Mining: Clustering, Classification, and Retrieval*, New York: Springer.

Berry, M. W., and Browne, M. (1999), *Understanding Search Engines: Mathematical Modeling and Text Retrieval*, Philadelphia: SIAM.

Berry, M. W., Dumais, S. T., and O'Brien, G. W. (1995), "Using Linear Algebra for Intelligent Information Retrieval," *SIAM Review*, 37, 573–595.

Berry, M. W., Hendrickson, B., and Raghavan, P. (1996), "Sparse Matrix Reordering Schemes for Browsing Hypertext," in *The Mathematics of Numerical Analysis*, eds. J. Renegar, M. Shub, and S. Smale, Providence: American Mathematical Society, pp. 99–123.

Berry, M. W., Raghavan, P., and Zhang, X. (2001), "Symbolic Preprocessing Techniques for Information Retrieval using Vector Space Models," in *Computational Information Retrieval*, ed. M.W. Berry, Philadelphia: SIAM, pp. 75–86.

Broder, A. Z. (1998), "On the Resemblance and Containment of Documents," in *1997 International Conference on Compression and Complexity of Sequences (SEQUENCES '97)*, Los Alamitos, CA: IEEE Computer Society, pp. 21–29.

Broder, A. Z., Charikar. M., Frieze, A. M., and Mitzenmacher, M. (2000), "Min-wise Independent Permutations," *Journal of Computer Systems and Science*, 60, 630–659.

Broder, A. Z., Glassman, S. C., Manasse, M. S., and Zweig, G. (1997), "Syntactic Clustering of the Web," SRC Technical Note No. 1997-015, Digital Equipment Corporation; in *Proceedings of the Sixth International World Wide Web Conference*, pp. 391–404.

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992), "Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections," in *Proceedings Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, June 21–24 1992, eds. N.J. Belkin, P. Ingwersen, and A. M. Pejtersen, New York: Association for Computing Machinery, pp. 318–329.

Deerwester, S., Dumais, S. T., Landauer, T., and Harshman, R. (1990), "Indexing by Latent Semantic Analysis," *Journal of the American Society of Information Science*, 41, 391–407.

Djeraba, C. (2002), *Multimedia Mining—A Highway to Intelligent Multimedia Documents*, Boston: Kluwer.

Feuerverger, A., Hall, P., Tilahun, G., and Gervers, M. (2004), "Measuring Distance, and Smoothing, Among Medieval Manuscripts and Pages of the World Wide Web," available online at `http://TKKKK`.

Gervers, M. (1982, 1996), *The Cartulary of the Knights of St. John of Jerusalem in England* (parts 1& 2), London: Oxford University Press.

Husbands, P., Simon, H., and Ding, C. H. Q. (2001), "On the Use of Singular Value Decomposition for Text Retrieval," in *Computational Information Retrieval*, ed. M.W. Berry, Philadelphia: SIAM, pp. 145–156 .

Quang, P. X., James, B., James, K. L., and Levina, L. (1999), "Document Similarity Measure for the Vector Space Model in Information Retrieval," NASAG Problem 99-5.

Reinert, M. (1990), "Une Méthodologie d'analyse des Données Textuelles et une Application: Aurélia de G. de Nerval," *Bulletin Méthod. Sociol.*, 26, 24–54.

Santini, S., and Jain, R. (1999), "Similarity Measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21, 871–883.

Simon, H., and Ding, C. H. Q. (2001), "On the Use of the Singular Value Decomposition for Text Retrieval," in *Computational Information Retrieval*, ed. M.W. Berry, Philadelphia: SIAM, pp. 145–156.

Zhang, J., and Korfhage, R. R. (1999), "A Distance and Angle Similarity Measure Method," *Journal of the American Society for Information Science*, 50, 772–779.