The Most Influential Paper Gerard Salton Never Wrote

David Dubin

Abstract

Gerard Salton is often credited with developing the vector space model (VSM) for information retrieval (IR). Citations to Salton give the impression that the VSM must have been articulated as an IR model sometime between 1970 and 1975. However, the VSM as it is understood today evolved over a longer time period than is usually acknowledged, and an articulation of the model and its assumptions did not appear in print until several years after those assumptions had been criticized and alternative models proposed. An often cited overview paper titled "A Vector Space Model for Information Retrieval" (alleged to have been published in 1975) does not exist, and citations to it represent a confusion of two 1975 articles, neither of which were overviews of the VSM as a model of information retrieval. Until the late 1970s, Salton did not present vector spaces as models of IR generally but rather as models of specific computations. Citations to the phantom paper reflect an apparently widely held misconception that the operational features and explanatory devices now associated with the VSM must have been introduced at the same time it was first proposed as an IR model.

Introduction

In a tribute written for the *Journal of the American Society for Information Science* (*JASIS*) (Crouch et al., 1996), Carolyn Crouch declares that Gerard Salton was more than just the leading authority in the field of information retrieval (IR). For thirty years, Crouch writes, "Gerry Salton *was* information retrieval" (p. 108) During times when the significance of computational IR research was in doubt, Salton defended and supported it "through the sheer force of his own personality and reputation" (Crouch et al., 1996, p.

David Dubin, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street, Urbana, IL 61820-6211
LIBRARY TRENDS, Vol. 52, No. 4, Spring 2004, pp. 748–764

108). Crouch's sentiments are echoed in the memoriam by Salton's other colleagues and former protégés, who reflect on his many contributions in research, teaching, writing, editing, and service to scholarly societies. They cite the textbooks he wrote, the SMART system developed under his leadership, the scholars that he mentored, and many other contributions. Donna Harman reminds the reader that Salton investigated "the use of the vector space model in clustering, relevance feedback, automatic linking, book indexing, passage retrieval, visualization, and many other areas" (Crouch et al., 1996, p. 108).

It is hardly surprising that Dr. Harman would cite Salton's pioneering research in the vector space model (VSM) for information retrieval: there are numerous citations crediting Salton with the original development of that IR model, as well as responses commenting on its advantages and limitations and proposing extensions or alternatives to it (Bollmann-Sdorra & Raghavan, 1993, 1998; Raghavan & Wong, 1986; Wong & Raghavan, 1984; Wong, Ziarko, & Wong, 1985; Wong, Ziarko, Raghavan, & Wong, 1986, 1987; McGill & Huitfeldt, 1979; Singhal, 2001; Howland & Park, 2004; Kobayashi & Aono, 2004). What is surprising, however, is that there is evidence that the VSM evolved over a much longer period of time than is usually acknowledged and that Salton did not publish an articulation of the model and its assumptions until several years after criticisms of those assumptions had been leveled and alternative models proposed (see section 7 below).

In giving credit to Salton for the vector model, a number of authors cite an overview paper titled "A Vector Space Model for Information Retrieval," which some show as published in the *JASIS* in 1975 and others as published in the *Communications of the Association for Computing Machinery* (CACM) in 1975. In fact, no such article was ever published, and citations to it usually represent a confusion of two 1975 articles (Salton, Wong, & Yang, 1975; Salton, Yang, & Yu, 1975), neither of which were overviews of the VSM as it is generally understood (see section 5 below). Some of Salton's own colleagues have been guilty of this mistake: both Cardie et al. and Singhal cite the *CACM* version, for example (Singhal, 2001; Cardie, Ng, Pierce, & Buckley, 2000). The paper is even cited in a few of the very last articles on which Salton is listed as a coauthor (Singhal, Salton, Mitra, & Buckley, 1996; Singhal & Salton, 1995). These papers were published close to or shortly after the time of his death, and so the errors cannot be blamed on Salton (remembered by his colleagues as a very careful and meticulous writer).

Another irony—one representing a more fitting tribute to Salton's legacy—is that locating papers containing the mistaken citation is very difficult using conventional citation databases such as the Web of Science. But discovery of the errors is greatly aided by search engines such as Google and CiteSeer—systems that employ techniques similar to those that Salton himself refined and recommended. The following papers were found in this way, and they cite one or the other versions of the bibliographic ghost: McCabe, Lee, Chowdhury, Grossman, & Frieder, 2000; Theophylactou & Lalmas, 1998; Arampatzis, van der Weide, Koster, van Bommel, 2000; Chen, 2001; Jiang & Littman, 2001; Nallapati, 2003. This leads us to the following questions: How did this mistake occur, and how was it perpetuated to the degree that it was? The answer seems to lie in a misconception widely held even by people who cite Salton's publications correctly: it is assumed that a description of the VSM must have been published sometime around 1975, even though it was not characterized as an IR model at that time.

Vector Spaces and Mathematical Models

We begin with a description of the VSM that Salton included in chapter 10 of his 1989 book on automatic text processing. That treatment includes the following characterization:

1. The VSM (like the Boolean and probabilistic models) represents information retrieval systems and procedures.
2. Global measures of similarity (such as the cosine measure) are computed between queries and documents.
3. Queries and documents are represented by term sets.
4. Both queries and documents can then be represented as ordered term vectors.
5. The components of the vectors are numbers representing either the importance of a term or simply the presence or absence of a term (1 or 0, respectively).

As mentioned above, the origins of these features are considerably earlier than the publications usually credited with the definition of the VSM. Salton himself did not publish a full articulation of the VSM as a retrieval model until this chapter, however, which appeared years after he was publicly credited with having invented the VSM.

The VSM is a mathematical model. Generalizing a definition by Rutherford Aris, Davis and Hersh (1981) define a mathematical model as a consistent mathematical structure designed to correspond to some physical, biological, social, psychological, or conceptual entity. They cite a number of uses for mathematical models, including:

1. predicting events in the physical world
2. guiding observation or experimentation
3. fostering conceptual understanding
4. assisting the "axiomatization of the physical situation" (Davis & Hersh, 1981, p. 78)
5. promoting progress in mathematics

So there are any number of ways in which the VSM might represent an advance for or contribution to IR research or systems design. Clarifying the particular role it plays as a model recommends a closer look at how vector representations are used to model other domains. The vector space is a very general and flexible abstraction, used to model many different domains and applications. When one makes the claim that a system or phenomenon is or can be modeled by a vector space, the first question one must consider is the level of abstraction at which that claim is being made:

> Algebraic—At the most abstract level, it can be a claim about addition and multiplication operations defined on a nonempty set of objects. Specifically, the claim that these operations satisfy all the algebraic axioms for a vector space (for example, addition commutes, multiplication distributes over addition, etc.). An example of a claim at this level is that the set of polynomials of degree no greater than *n* define a vector space (Lay, 1994).
>
> Measurement-theoretic—At another level, to say that something is represented by a vector space can be an empirical claim that two or more variables define a space. In that case, the substance of the claim is about ordinal and additive relations holding among the values of those variables for some known entities (that is, that the variables are *quantitative*) and also that distance between the entities is a function of the differences along each of the individual variables defining the space (Michell, 1990).
>
> Physical—Real vector spaces are often used to model physical forces such as gravity and relations such as velocity. For example, the direction and velocity of a boat may be represented by a vector, the speed and direction of the current is represented by a second vector, and the course and speed made good are shown to be the sum of those vectors (Fraleigh & Beauregard, 1987). Models such as these entail claims about the physical world.
>
> Data-centric—In multivariate analysis, vector spaces are used to model a set of observations. The data is typically represented as a matrix where items or cases are represented as rows and observations for a particular feature are represented as columns. Geometrically, the cases are understood to be plotted in the space of feature values, but no empirical claim about the features, the nature, or relations among the values need be advanced: in this case, the vector space is simply a way of presenting the values assigned to the observations. This representation typically precedes a transformation of the data, such as reexpressing them in a space of lower dimensionality in order to reveal latent structures or patterns (Green & Carroll, 1976). In that case, the operations performed using the data can be explained and understood as operations on vectors and matrices.

It is at this last data-centric level that one should understand the use of vector abstractions in most of Salton's IR publications: vector components represent raw or modified observations, and relations between vectors (such as the cosine of the angle between pairs of them) are devices for explaining computations or other design choices about how an IR system operates. As we shall see, the habit of describing data and computations in terms of operations on vectors eventually became so familiar that some later interpretations seem to lose sight of the role the vector model was intended to play.

Earliest Examples

The elements of what would come to be known as the VSM are evident in Salton's earliest publications on experimental IR and also the work of other authors (Switzer, 1965; Sammon, 1968). In a 1963 article in the *Journal of the Association for Computing Machinery* (*JACM*), Salton describes systems and methods for what at that time he calls "associative document retrieval techniques." Building on earlier work by people such as H. P. Luhn, Salton outlines the architecture for automated systems that extract words from machine-readable texts, select a subset of those words deemed significant enough to represent the document content, and compute measures of association between pairs of terms, pairs of documents, and between documents and queries.

Even in this early paper one finds frequencies of extracted words presented using matrix and vector notation and the cosine of angles between vectors recommended as a measure of association. The vector representation is employed to describe similarities computed using both extracted words and citation data. Furthermore, it is clear that vector representations are to be understood precisely at the data-centric level described above: the term-document matrix is called an *incidence matrix,* leaving no doubt that what the vector components model are observations. The similarity measures are at all points described as methods or operations on the data that *can be interpreted* as relations between vectors.

SMART was the system Salton developed over the course of his career as an IR researcher. More than just an IR system, SMART was the working expression of Salton's theories and the experimental environment in which those theories were evaluated and tested (Salton, 1971). The earliest papers describing the SMART system show that the same extraction and association procedures outlined in the *JACM* article are central to SMART's design and operation (Salton, 1965b; Salton & Lesk, 1965). In 1965 Salton published a paper in *IEEE Spectrum* titled "Progress in Automatic Information Retrieval" (1965a). That article discusses specific features of SMART and characterizes document representations and similarity computations in terms of vectors. In addition, relevance feedback experiments (conducted by J. J. Rocchio) are described in terms of query vector modifications. In all these examples, the vector spaces illustrate how computations such as similarity measures and relevance feedback are applied to the data; the vector spaces are models of computations executed by the system.

Retrieval Models

In 1968 Salton published *Automatic Information Organization and Retrieval,* a book that presents a more developed treatment of the concepts introduced in the earlier IR papers and more details on the design and evaluation of the SMART system. Salton devotes chapter 6 entirely to retrieval models but, interestingly, that chapter contains none of the vector or matrix notation seen in the earlier papers. This is not to say that vector representations are absent from the book: as in the earlier writings, they appear in the context of explaining specific computations in the chapters on statistical operations (4) and the retrieval process (7). But for Salton a retrieval *model* was closer to the formal model later presented by Bookstein and Cooper (Bookstein & Cooper, 1976).[1] Retrieval models, according to this understanding, are more abstract than particular computations. The retrieval operation is understood as a mapping between the space of query words and the space of documents (that is, replacement of the former by the latter). Salton presents retrieval models in set-theoretic terms, though there is no reason why vectors could not be used to model retrieval at the same level of abstraction: John W. Sammon Jr. published an abstract model similar to Salton's using vectors rather than sets (Sammon, 1968). According to Salton, a retrieval model should explicate such issues as

- whether a particular set has a well-defined complement
- whether the request space is identical with the object space; that is, whether the set of possible query descriptions is the same as the set of possible document descriptions
- whether document and query identifiers are unstructured and independent of one another or whether relations between them are defined
- implications of order relations on queries and documents, such as whether a more specific query guarantees the retrieval of fewer documents and whether those will be a proper subset of a more general query
- whether the system contains a classification language (that is, a set of categories distinct from the document description language) and functions to map document and request descriptions into those categories
- whether elements of the description languages are all positive properties, or whether negation can be expressed independent of any other existing property

A retrieval model, according to Salton, represents documents, description features (such as index terms), queries, and the relationships within and across those sets. The vector spaces described in the 1968 book, however, are not models of documents, terms, or queries: they are models of numeric data and of computations with those data. The numbers represent the documents, terms, and queries within a system such as SMART. The vector space models are explanatory devices intended to help the reader understand how part of a system works; the retrieval model speaks to more general questions, such as those listed above.

Some of the retrieval modeling issues have since recurred in disputes that intimately couple them with those of vector representations (as explained below). But in 1968 Salton treated these modeling issues separately from those used to characterize similarity and relevance feedback computations.

The Term Discrimination Model

In 1974 and 1975 Salton published several important papers on a theory of indexing and a method for selecting words from documents and assigning numeric weights to them. The presentation of this model, called the "term discrimination value model" (TDV), would prove to be significant not only because an automatic indexing principle was expressed in this model but also because of its impact on the IR research community's perception of what became known as the VSM.

The term discrimination model proposes that document features (such as extracted words) most useful for indexing will be those that increase the average dissimilarity between pairs of documents. In the basic conception the computed similarity averaged over every document pair is compared with and without the inclusion of a feature under consideration. The features are then ranked by the difference between those averages, with the best having the most dramatic lowering of average similarity when they are included. The process of computing a discrimination value can be speeded by comparing each document to an artificial average or centroid document rather than computing similarities for every document pair.

It is not essential to the TDV indexing model that similarity computations be explained in terms of operations on vectors or that document features be weighted or ordered. But, not surprisingly, Salton explained the model geometrically using vectors as he had done in the earlier publications. The key publications on the TDV indexing model are a Cornell technical report (Salton, 1974) that was republished a year later as a monograph (Salton, 1975), an article in the January–February 1975 issue of the *JASIS* (Salton, Yang, & Yu, 1975), and an article in the November 1975 issue of *CACM* (Salton, Wong, & Yang, 1975).

The articles in *CACM* and *JASIS* (particularly the former) had the greatest impact on how the VSM came to be viewed. This is largely because of presentational choices that had little direct bearing on the thesis of either article. Most significantly, the CACM article is titled "A Vector Space Model for Automatic Indexing." One might consider this an unfortunate choice since (as discussed above) vector spaces are not essential to the TDV selection and weighting model. What both articles actually present is an "average document similarity model" for automatic indexing. Because Salton and his colleagues were computing document similarity the same way that they had been doing for years, they used the same mathematical models to explain how those computations were performed. Hence the vectors and vector operations. After over a decade of explaining their system design choices in this way, Salton and his colleagues seem to have grown comfortable with vector spaces as an economical explanatory tool. That may help account for why the vector space is foregrounded in the *CACM* article's title and in the opening paragraph, which begins "Consider a document space . . . "

In addition, both articles use the same illustration for their first figure: a three-dimensional coordinate system where index terms are depicted as orthogonal basis vectors and documents are plotted as vectors in the space of term weights. For purposes of advancing and explaining the thesis this illustration is correct, since it gives the reader a correct impression of how similarities were computed in the experiments conducted to evaluate TDV as an indexing strategy. But as we will see, the figure made a lasting impression on readers, and eventually more was read into this illustration than was warranted.

The Vector Space as an IR Model

The next significant evolutionary stage of how the VSM came to be perceived became evident in 1979. That year Salton published an article in the *Journal of Documentation* (JDoc) titled "Mathematics and Information Retrieval." This article was the first since the 1968 text to discuss issues of modeling in depth, and it is significant for two reasons:

1. This seems to be the first time Salton refers to the VSM as an IR model in print
2. Salton describes an *orthogonality assumption* for the first time in this article

Informally, one can understand the orthogonality issue as whether the vectors forming the basis of the space (that is, those representing variables under investigation) are at right angles to one another. Modeling variables as orthogonal basis vectors suggests that those variables either are or should be treated as statistically independent of one another. Salton's vector spaces (such as those in the 1975 TDV articles) model frequencies of extracted words with orthogonal basis vectors, which gives the false impression that words are assumed to occur independently of each other. As noted above, however, Salton's use of vector spaces is for modeling how an IR system performs particular computations. No empirical claim about word occurrences is implied: the equations and diagrams merely illustrate how the system was programmed to match documents and queries.

In "Mathematics and Information Retrieval" Salton uses the term "vector processing model" rather than vector space model, and this is the first suggestion that the VSM has shifted from being understood as a model for illustrating specific computations to being an IR model in its own right. This article recapitulates much of the set-theoretic modeling discussion in the 1968 text, but this time puts alongside it a section on "Retrieval as vector matching operations." The description of the vector representation and operations is similar to the earlier computational/operational illustrations but with some telling exceptions: Salton mentions an "underlying basis" out of which the vectors representing index terms are composed via linear combination. Precisely what this basis represents Salton declines to specify, but he states that to assume that this basis is orthogonal would be at odds with "actual fact" since "relationships may exist between individual vector attributes" (Salton, 1979, p. 8).

The significance of this shift in thinking is twofold: First, Salton's use of vector spaces has temporarily drifted from the operational, data-centric conception seen earlier to some other vague level of abstraction. Second, the question of correlation or orthogonality is explicitly linked to a modeling issue that Salton had identified in 1968: the existence of relations or dependencies among the document and query identifiers.

When Salton alludes to the mysterious "underlying basis," he may have in mind latent dimensions of the kind that can be uncovered through, for example, principal components analysis or factor analysis. Methods for representing documents in these empirically derived vector spaces had been proposed before (Switzer, 1965; Sammon, 1968) and since (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), but the techniques had not been used in Salton's research. Perhaps the "underlying basis" is supposed to represent psychological variables of the kind that can be studied by eliciting similarity judgments from experimental subjects. Several key studies investigating the suitability of vector spaces as psychological models of similarity were published in the years before and after "Mathematics and Information Retrieval" (Tversky, 1977; Tversky & Gati, 1978; Tversky & Gati, 1982). But such psychological models were never part of the SMART system. In any case, the basis cannot represent either the terms or the documents, since Salton claims that both term and document vectors are linear combinations of the basis vectors. The reader is left with the impression of entities that somehow have a real existence independent of the IR system design decisions and that the system models imperfectly. What could those be?

Salton may have imagined that the "underlying basis" represented both empirically derived and psychologically real dimensions. For example, an article by Koll, published the same year as Salton's *JDoc* article, describes a system (called WEIRD) in which a derived vector space is proposed as a solution to the problem of measuring conceptual similarity (Koll, 1979). Alternatively, Salton may have supposed the basis to represent concepts that are neither psychologically real nor derived from data but rather pure abstractions: A few years earlier Salton's future coauthor Michael McGill had published a paper relating SMART to an abstract but informal vector model proposed by Meincke and Atherton (McGill, 1976; Meincke & Atherton, 1979).

The other significance of the orthogonality/correlation issue in 1979 is that it is a special case of a retrieval modeling issue Salton had cited in 1968: relationships between elements of the various sets. The year 1979 saw the first coupling of this abstract modeling issue with vector representations that had been discussed separately in the 1968 book. Furthermore, the earliest characterizations of Salton's VSM as an IR model appeared that year in separate publications by Salton, McGill, and Koll (Salton, 1979; McGill & Huitfeldt, 1979; Koll, 1979).

Koll identifies the basis in Salton's vector model as the index term vectors. Within a few years, Salton would come to agree that it is the index term vectors (not some other basis) that are assumed to be orthogonal in his VSM. But that position is equally problematic: if the basis vectors represent index terms then those vectors are not *assumed* to be orthogonal, they simply *are* orthogonal, because all that the vectors represent is the way that term frequency data are used in the system's computations.

When a commentator on the VSM says that term basis vectors are assumed to be orthogonal, this is a misstating of the actual fact that dependencies among words in natural language are ignored. Approaches such as WEIRD and Latent Semantic Indexing *do* compute and use information about these dependencies, and although SMART's similarity computations never worked that way, there is ample evidence in the writings of Salton and his colleagues that they understood word/term dependencies and conducted many experiments to employ term associations in retrieval (Salton, 1963; Lesk, 1969; Salton, Buckley, & Yu, 1983).

It is a subtle error of language or description to claim that the VSM assumes term vectors are orthogonal. And it is no coincidence that this error first appears when the VSM was first characterized as a retrieval model instead of a computation model. If term vector orthogonality is a simplifying assumption, then that implies the existence of correlated terms independent of their operational definition in the computational design choices. But, as with the "underlying basis" of 1979, it is not clear what those entities could be. Evidently, the familiarity of vector space illustrations has led to a confounding of objective facts (that term dependencies and word associations exist) with implications for how those facts might be modeled (as correlations between vectors in a vector space). In 1968 Salton had included the character of relationships among members of the descriptor set as a retrieval modeling issue. By 1979, discussion of those relationships had become inseparable from discussion of similarity computations. That confusion continued to shape reactions to Salton's contributions over the subsequent years.

Early Reactions to the Vector Model

Responses and counterproposals to the vector model from 1979 onward are each interesting on their own terms and for their own reasons (Koll, 1979; Wong et al., 1986, 1985; Bollmann-Sdorra & Raghavan, 1998). But understanding how they shaped the understanding of the VSM itself requires attention to two issues:

1. Respondents did not realize how recently the VSM came to be characterized as a retrieval model. Looking back at the earlier illustrative vector models of similarity and relevance feedback computations, they assumed the VSM went back at least as far as the 1975 TDV papers discussed above.
2. The IR modeling issues are no longer distinct from the computational modeling issues, as they were in 1968.

The most significant early response to Salton was Wong and Raghavan's "Vector Space Model for Information Retrieval: A Reevaluation" (1984). This paper pointed to inconsistencies in earlier proposals for defining vector correlation. It is the first in a series that would propose a different method for using word co-occurrence data to define an orthogonal basis for a vector space; Wong and Raghavan called it the Generalized Vector Space Model (GVSM) (Wong et al., 1985; Raghavan & Wong, 1986; Wong et al., 1987).

Beyond these contributions it is interesting to look at how Wong and Raghavan interpreted Salton's earlier writings and to see the impact of this interpretation on how we conceive of the VSM today. Reviewing the 1960s and 1970s publications, Wong and Raghavan suggest that Salton's vectors are informal, notational devices and not intended as a logical tool. They accuse Salton of ignoring issues such as whether the algebraic axioms defining a vector space are even satisfied. According to Wong and Raghavan (1984), that amounts to "casual flirtings" (p. 170) with the concept of vector spaces and should not be taken seriously. These criticisms are understandable in light of how they are interpreting the earlier publications.

As stated earlier, in the pre-1979 writings, vectors are used for modeling term frequency observations and for explaining similarity and relevance feedback computations. Salton's vector spaces are rigorous and formally correct, but the vector models themselves are illustrative (not merely notational). The axioms defining a vector space are satisfied simply because at the algebraic level the vector space in question is the familiar Euclidean space of real numbers. The orthogonality of the basis follows from definition, since what a vector space represents is nothing more than how computations are performed by a system such as SMART.

Wong and Raghavan are looking back with the assumption that the VSM has been an IR model all along. From that perspective, they reasonably ask whether the VSM implies a vector space in the formal sense. But in reality the formality of the vector space was never in doubt, only what was meant by an IR model.

Wong and Raghavan's GVSM is a perfectly reasonable proposal for using word co-occurrence data in an IR system. But they *present* it as a formal model for vector correlation and orthogonality in IR. The issues of dependencies and patterns in textual data take a back seat to questions of how linear dependence, projection, and correlation are defined. What began as an illustrative formalism came to significantly shape the way theoretical questions were expressed and the language in which solutions were proposed.

A later response to Salton shows this intertwining of models even more clearly. In 1993 Bollmann-Sdorra and Raghavan published "On the Delusiveness of Adopting a Common Space for Modeling IR Objects: Are Queries Documents?" Recall that, like the issues bearing on orthogonality, this is another retrieval model issue that Salton had identified in 1968: whether the request space is identical with the object space. Bollman-Sdorra and Raghavan address this important issue again but *entirely* within the framework of vector computations. Each of their claims is supported by examples that show how particular parameter combinations (similarity measures, preference orders, etc.) lead to unexpected or counterintuitive results. The fact that these examples are all contrived does not invalidate their arguments, but it does mean that the question of whether queries are documents is being addressed without ever advancing a claim about actual documents or actual queries—only via hypothetical examples of document and query representations.

Finally, consider what it means to say that an IR system is based on the vector space model. On the one hand, it may mean that specific data processing procedures are executed in the same manner as (or similarly to) computations in the SMART system (term weighting, similarity measures, relevance feedback, and so on). On the other hand, it may mean only that the computations can be explained or illustrated using vector spaces, whether or not they are anything like SMART's procedures. Either way one is foregrounding models of numeric or binary data that are in turn models of index terms, documents, queries, and user profiles.

The Vector Space Model Defined at Last

Salton's 1983 book with Michael McGill, *Introduction to Modern Information Retrieval,* does not include an in-depth discussion of modeling issues, apart from a short section on them in the chapter on future directions in information retrieval. As a result, the book does not lay out assumptions and parameters of the VSM in detail and, indeed, refers to the VSM as an information retrieval model only in passing (p. 422). This brief allusion to the "vector space model" may mark a terminological shift, since earlier papers, as mentioned above, had used the phrase "vector processing model."

For the most part, the 1983 text uses vector spaces only to explain and illustrate the computations performed by the SMART system, just as in the work published before 1979. In section 2 of chapter 4, however, the authors state that SMART is based on a model in which "Each term included in a given document or query vector is assumed to be unrelated (orthogonal) to the other terms, and all the terms are considered equally important (except for distinctions inherent in the assignment of weights to the individual terms)" (p. 130).

Salton and McGill go on to explain that the orthogonality assumption is only a "first-order approximation to the true situation" (1983, p. 130) since words do not occur independently in texts. They justify the assumption with the argument that taking term dependencies into account adds complexity and (based on experimental evidence) seems to have little practical impact on retrieval success.

This discussion is noteworthy for two reasons. First, the orthogonality assumption is described as applying to the term vectors (rather than some unspecified basis as in the 1979 article). Secondly, it is another telling example of the retrieval/computational model confusion. On the one hand, the authors correctly express a retrieval model issue, that is, the decision to treat words as unrelated. They acknowledge that dependencies known to exist between words in texts are not represented, measured, or used by the system. Salton and McGill understand the impact that this might have on retrieval results and explain why they choose to dismiss that concern.

On the other hand, the authors describe this decision in terms of a vector orthogonality assumption. As explained earlier, term vector orthogonality is not an assumption but rather a fact resulting from definition. Indeed, it is not even accurate to describe the retrieval model as depending on an *assumption* of term independence; the SMART system makes no probabilistic inference that could be falsified but merely computes document/query similarity in particular ways.[2] This characterization of SMART is another unfortunate consequence of seeing vector spaces as an IR model. As mentioned earlier, it invited Wong and Raghavan to question Salton's theoretical rigor the following year.

Salton's 1989 book, *Automatic Text Processing,* includes the author's first full description of the VSM as an IR model . Ironically, much of the characterization is adapted directly from Wong and Raghavan's earlier criticism of what they interpreted Salton to have meant. The illustration of the document space in chapter 10 is an exact copy of figure 1 in Wong and Raghavan's 1984 paper (and their 1986 follow-up) and depicts the term vectors at oblique angles to one another rather than at right angles as in the 1975 TDV papers. Based on Wong and Raghavan's criticism, Salton corrects an earlier (1979) error on the use of term and document correlations to define an orthogonal basis and follows their example in calling for additional information to define the correlations. Citing Raghavan and Wong, Salton repeats the 1983 mischaracterization that term vector orthogonality implies an assumption of term independence.

Epilogue: The Paper Salton Never Wrote

As one would expect, published references to the vector model are usually much briefer than the detailed responses, extensions, and alternative proposals discussed above. An author may state, for example, that his or her experimental system realizes or is based on the VSM. Or the VSM may simply be included in a list of other models or formalisms.

It is ironic that in these references the most popular citations for the VSM seem to be the two TDV papers, the 1983 text, and the 1971 collection of SMART system articles. These choices are understandable: the *CACM* article was suggestively titled, and both it and the *JASIS* article included the same evocative illustration for figure 1. The 1971 text concerns SMART, the design of which largely defined the loose bundle of operational assumptions and expectations that people associate with systems based on the VSM. The 1983 book by Salton and McGill included descriptions that made it clear that the abstract and computational modeling issues that had been kept distinct in 1968 were by then inextricably intertwined.

Those four publications, however, are far less significant in terms of the VSM's evolution than the 1979 *JDoc* article (which first presented it as an IR model in its own right), the 1984 and 1986 criticisms by Raghavan and Wong, and the 1989 chapter which finally expressed in detail how the VSM was supposed to be interpreted. If most casual references to the VSM ignore these milestones, that probably reflects a misconception that the operational features and explanatory devices now associated with the VSM must have been introduced at the same time it was first proposed as an IR model.

The strongest evidence for such a misconception is seen in the error of changing the name of one of the TDV articles to "A Vector Space Model for Information Retrieval" by authors who are citing the VSM, not the TDV term selection and weighting theory. As stated in the introduction, even some members of the Cornell SMART research group have made this mistake, and that has resulted in Dr. Salton appearing as coauthor on work citing a paper he never wrote. But the real evolution of the VSM (as people conceived it) is even more fascinating than citation errors for which Dr. Salton bears none of the blame. What began as a growing comfort in using vector spaces to explain computations led to the use of language that suggested the VSM was a retrieval model in its own right. When Salton and his colleagues were challenged on the implications of taking that language seriously, they joined their critics in reinterpreting their earlier writings.

Acknowledgments

I wish to thank Boyd Rayward for his encouragement and patience, Hans-Hermann Bock for sending me a copy of one of Dr. Salton's less-known papers, and Pauline Cochrane, Linda Smith, Allen Renear, Carole Palmer, and the members of the GSLIS Research Writing Group for their many suggestions. Errors and mischaracterizations are my own responsibility.

Notes

1. Bollmann and Raghavan distinguish between IR theories/models (which concern documents, texts, and users as empirical entities) and IRS theories/models (which concern them as formal entities in a system) (Bollmann & Raghavan, 1991). The present article focuses on a different contrast: theories and models of documents, texts, and users (whether empirical or formal) vs. models of computations that are executed on representations of those entities.
2. The VSM can, however, be explained or interpreted within the framework of a general probabilistic model, as Norbert Fuhr (2001) has shown.

References

Arampatzis, A., van der Weide, T., Koster, C., & van Bommel, P. (2000). *An evaluation of linguistically-motivated indexing schemes.* Paper presented at the BCS-IRSG 2000 Colloquium on IR Research, April 5–7, Sidney Sussex College, Cambridge, England.

Bollmann, P., & Raghavan, V. V. (1991). The axiomatic approach for theory development in IR. In *Working notes of NSF workshop on future directions in text analysis, retrieval, and understanding* (pp. 16–22). Chicago: National Science Foundation.

Bollmann-Sdorra, P., & Raghavan, V. V. (1993). On the delusiveness of adopting a common space for modeling IR objects: Are queries documents? *Journal of the American Society for Information Science, 44*(10), 579–587.

Bollmann-Sdorra, P., & Raghavan, V. V. (1998). On the necessity of term dependence in a query space for weighted retrieval. *Journal of the American Society for Information Science, 49*(13), 1161–1168.

Bookstein, A., & Cooper, W. (1976). A general mathematical model for information retrieval systems. *Library Quarterly, 46*(2), 153–167.

Cardie, C., Ng, V., Pierce, D., & Buckley, C. (2000). Examining the role of statistical and linguistic knowledge sources in a general knowledge question-answering system. In Sergei Nirenburg (Ed.), *Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000)* (pp. 180–187). San Francisco: Morgan Kaufmann.

Chen, H. (2001). *Looking for better Chinese indexes: A corpus-based approach to base NP detection and indexing.* Unpublished doctoral dissertation, Guangdong University of Foreign Studies, Guangzhou, China.

Crouch, C., McGill, M., Lesk, M., Sparck-Jones, K., Fox, E. A., Harman, D., & Kraft, D. H. (1996). In memoriam: Gerard Salton, March 8, 1927–August 28, 1995. *Journal of the American Society for Information Science, 47*(2), 108–115.

Davis, P. J., & Hersh, R. (1981). *The mathematical experience.* Boston: Birkhäuser.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Fraleigh, J. B., & Beauregard, R. A. (1987). *Linear algebra.* Reading, MA: Addison-Wesley.

Fuhr, N. (2001). Models in information retrieval. In M. Agosti, F. Crestani, & G. Pasi (Eds.), *Lectures on information retrieval: Third European Summer-School, ESSIR 2000 Varenna, Italy, September 11–15, 2000* (pp. 21–50). Berlin: Springer-Verlag.

Green, P. E., & Carroll, J. D. (1976). *Mathematical tools for applied multivariate analysis.* New York: Academic Press.

Howland, P., & Park, H. (2004). Cluster-preserving dimension reduction methods for efficient classification of text data. In M. W. Berry (Ed.), *Survey of text mining: Clustering, classification, and retrieval* (pp. 3–23). New York: Springer-Verlag.

Jiang, F., & Littman, M. L. (2001). Approximate dimension reduction at NTCIR. In J. Adachi & N. Kando (Eds.), *Proceedings of the Second NTCIR Workshop on Research in Chinese and Japanese Text Retrieval and Text Summarization.* Tokyo: National Institute of Informatics.

Kobayashi, M., & Aono, M. (2004). Vector space models for search and cluster mining. In M. W. Berry (Ed.), *Survey of text mining: Clustering, classification, and retrieval* (pp. 103–122). New York: Springer-Verlag.

Koll, M. (1979). WEIRD: An approach to concept-based information retrieval. *ACM SIGIR Forum, 13*(4), 32–50.

Lay, D. C. (1994). *Linear algebra and its applications.* Reading, MA: Addison-Wesley.

Lesk, M. E. (1969). Word-word association in document retrieval systems. *American Documentation, 20*(1), 27–38.

McCabe, M. C., Lee, J., Chowdhury, A., Grossman, D., & Frieder, O. (2000). On the design and evaluation of a multi-dimensional approach to information retrieval [poster session]. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 363–365). New York: ACM Press.

McGill, M. J. (1976). Knowledge and information spaces: Implications for retrieval systems. *Journal of the American Society for Information Science, 27*(4), 205–210.

McGill, M. J., & Huitfeldt, J. (1979). Experimental techniques for information retrieval. *Annual Review of Information Science and Technology, 14,* 93–127.

Meincke, P. P., & Atherton, P. (1979). Knowledge space: A conceptual basis for the organization of knowledge. In A. Neelameghan (Ed.), *Ordering systems for global information networks: Proceedings of the Third International Study Conference on Classification Research held at Bombay, India during 6–11 January 1975* (pp. 158–164). Manipal, India: Manipal Power Press. (FID Publication Number 553)

Michell, J. (1990). *An introduction to the logic of psychological measurement.* Hillsdale, NJ: Lawrence Erlbaum.

Nallapati, R. (2003). Semantic language models for topic detection and tracking. In C. Boulis, E. Beck, & V. Lavrenko (Eds.), *Proceedings of the HLT-NAACL 2003 student research workshop* (pp. 1–6). Edmonton, Canada.

Raghavan, V. V., & Wong, S. K. (1986). A critical analysis of the vector space model for information retrieval. *Journal of the American Society for Information Science, 37*(5), 279–287.

Salton, G. (1963). Associative document retrieval techniques using bibliographic information. *Journal of the Association for Computing Machinery, 10*(4), 440–457.

Salton, G. (1965a, August). Progress in automatic information retrieval. *IEEE Spectrum, 2*(8), 90–103.

Salton, G. (1965b). An evaluation program for associative indexing. In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation: Symposium proceedings* (pp. 201–210). Washington, DC: National Bureau of Standards. (NBS Miscellaneous Publication No. 269)

Salton, G. (1968). *Automatic information organization and retrieval.* New York: McGraw-Hill Book Company.

Salton, G. (Ed.). (1971). *The SMART retrieval system: Experiments in automatic document processing.* Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1974). *A theory of indexing.* Ithaca, NY: Cornell University. (Tech. Rep. No. CU-CSD-74–203)

Salton, G. (1975). *A theory of indexing.* Philadelphia: Society for Industrial and Applied Mathematics. (Regional Conference Series in Applied Mathematics No. 18)

Salton, G. (1979). Mathematics and information retrieval. *Journal of Documentation, 35*(1), 1–29.

Salton, G. (1989). *Automatic text processing: The transformation, analysis and retrieval of information by computer.* Reading, MA: Addison-Wesley.

Salton, G., Buckley, C., & Yu, C. T. (1983). An evaluation of term dependence models in information retrieval. In G. Salton & H. Schneider (Eds.), *Research and development in information retrieval: Proceedings, Berlin, May 18–20, 1982* (pp. 151–173). Berlin: Springer-Verlag.

Salton, G., & Lesk, M. E. (1965). The SMART automatic document retrieval system—An illustration. *Communications of the ACM, 8*(6), 391–398.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval.* New York: McGraw-Hill.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Salton, G., Yang, C., & Yu, C. (1975). A theory of term importance in automatic text analysis. *Journal of the American Society for Information Science, 26*(1), 33–44.

Sammon, J. W. (1968). *Some mathematics of information storage and retrieval.* New York: Griffis Air Force Base, Rome Air Development Center, Air Force Systems Command. (Tech. Rep. No. RADC-TR-68–178)

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin, 24*(4), 35–43.

Singhal, A., & Salton, G. (1995). Automatic text browsing using vector space model. In *Proceedings of the Fifth Annual IEEE Dual-Use Technologies and Applications Conference* (pp. 318–324). Washington, DC: IEEE Computer Society Press.

Singhal, A., Salton, G., Mitra, M., & Buckley, C. (1996). Document length normalization. *Information Processing and Management, 32*(5), 619–633.

Switzer, P. (1965). Vector images in document retrieval. In M. E. Stevens, V. E. Giuliano, & L. B. Heilprin (Eds.), *Statistical association methods for mechanized documentation* (pp. 163–171). Washington, DC: National Bureau of Standards. (NBS Miscellaneous Publication 269)

Theophylactou, M., & Lalmas, M. (1998). *A Dempster-Shafer belief model for document retrieval using noun phrases.* Paper presented at the BCS-IRSG 20th Annual Colloquium on IR Research: Discovering New Worlds of IR, March, Grenoble.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*(4), 327–352.

Tversky, A., & Gati, I. (1978). Studies of similarity. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization.* Hillsdale, NJ: Erlbaum.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review, 89*(2), 123–154.

Wong, S. K. M., & Raghavan, V. V. (1984). Vector space model of information retrieval: A reevaluation. In *Proceedings of the 7th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 167–185). Cambridge: Cambridge University Press.

Wong, S. K. M., Ziarko, W., Raghavan, V. V., & Wong, P. C. N. (1986). On extending the vector space model for Boolean query processing. In *Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 175–185). New York: ACM Press.

Wong, S. K. M., Ziarko, W., Raghavan, V. V., & Wong, P. C. N. (1987). On modeling of information retrieval concepts in vector spaces. *ACM Transactions on Database Systems, 12*(2), 299–321.

Wong, S. K. M., Ziarko, W., & Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 18–25). New York: ACM Press.