36. M. E. Wysession, L. Bartko, J. B. Wilson, *ibid.*, p. 13667.
37. T. Seno and S. Maruyama, *Tectonophysics* **160**, 23 (1989).
38. C. Kincaid and P. Olson, *J. Geophys. Res.* **92**, 13832 (1987).
39. T. A. Cross and R. H. Pilger, *Nature* **274**, 653 (1978).
40. R. D. Jarrard, *Rev. Geophys.* **24**, 217 (1986).
41. This work is based in part on one chapter of S.Z.'s thesis (University of Michigan, 1994). (S.Z. thanks G. Hulbert, H. Pollack, L. Ruff, and K. Satake). Funded by the David and Lucile Packard Foundation and NSF grant EAR-9496185. S.Z. is supported by a Texaco Fellowship at the California Institute of Technology. Some computations were performed at the Pittsburgh Supercomputer Center. Contribution 5451 of the Division of Geological and Planetary Sciences of the California Institute of Technology.

Gauging Similarity with *n*-Grams: Language-Independent Categorization of Text

Marc Damashek

A language-independent means of gauging topical similarity in unrestricted text is described. The method combines information derived from *n*-grams (consecutive sequences of *n* characters) with a simple vector-space technique that makes sorting, categorization, and retrieval feasible in a large multilingual collection of documents. No prior information about document content or language is required. Context, as it applies to document similarity, can be accommodated by a well-defined procedure. When an existing document is used as an exemplar, the completeness and accuracy with which topically related documents are retrieved is comparable to that of the best existing systems. The results of a formal evaluation are discussed, and examples are given using documents in English and Japanese.

I report here on a simple, effective means of gauging similarity of language and content among text-based documents. The technique, known as Acquaintance, is straightforward; a workable software system can be implemented in a few days' time. It yields a similarity measure that makes sorting, clustering, and retrieving feasible in a large multilingual collection of documents that span an unrestricted range of topics. It makes no use of words per se to achieve its goals, nor does it require prior information about document content or language. It has been put to practical use in a demanding government environment over a period of several years, where it has demonstrated the ability to deal with error-laden multilingual texts.

Sorting and categorizing the enormous amount of text now available in machine-readable form has become a pressing problem. To complicate matters, much of that text is imperfect, having been derived from existing paper documents by means of an error-prone scanning and character recognition process.

Over the past few decades, many document categorization and retrieval methods [for example, (*1–3*) and references therein] have relied on the self-evident utility of words, sentences, and paragraphs for sorting, categorizing, and retrieving text (*4*), and various means of suppressing uninformative words, removing prefixes, suffixes, and endings, interpreting inflected forms, and performing related tasks have been developed. Depending on the application, these methods share a number of potential drawbacks: They require a linguist (or a polyglot) for initial setup and subsequent tuning, they are vulnerable to variant spellings, misspellings, and random character errors (garbles), and they tend to be both language-specific and topic-specific.

A potentially more robust alternative, the purely statistical characterization of text in terms of its constituent *n*-grams (sequences of *n* consecutive characters) (*5*, *6*), has sporadically been applied to textual analysis and document processing (*7*). Recent examples include spelling and error correction (*8–14*), text compression (*15*), language identification (*16*, *17*), and text search and retrieval (*18–21*).

The literature offers no convincing evidence of the usefulness of either approach for the purpose of categorizing text according to topic in a completely unrestricted multilingual environment, that is, an environment that encompasses many different documents containing a nonnegligible number of character errors. The present paper is intended to provide such a demonstration.

The author is at the Department of Defense, Fort George G. Meade, MD 20755–6000, USA.

Methodology

Except for the Japanese example below, all text shown here has been reduced to a 27-character alphabet (uppercase `A` through `Z` plus `space`). The alphabet size only weakly affects the ultimate outcome (*22*), so there is no loss of generality; consistent results are also obtained for a range of *n*-gram lengths (*22*). I have worked with 5-grams for the English language examples, and 6-grams for the Japanese (*23*), but it should be borne in mind that the size of the alphabet and the *n*-gram length are both flexible.

Given an alphabet and value of n, a naïve calculation of the number of possible *n*-grams can be misleading. It is immaterial that, for example, $27^5 = 14{,}348{,}907$ distinguishable 5-grams can be formed using 27 characters because most of them are never encountered. Huge reserves of computer memory for *n*-gram statistics are therefore unnecessary.

An entire document can be represented as a vector whose components are the relative frequencies of its distinct constituent *n*-grams (the exhaustive list of constituent *n*-grams comprises all *n*-character sequences produced by an *n*-character-wide window displaced along the text one character at a time, and contains many duplications). Let the document contain J distinct *n*-grams, with m_i occurrences of *n*-gram number i. Then the weight assigned to the ith vector component will be

$$x_i = \frac{m_i}{\sum_{j=1}^{J} m_j} \qquad (1)$$

where

$$\sum_{j=1}^{J} x_j = 1 \qquad (2)$$

Because both the size of the alphabet and the length of the *n*-grams are arbitrary, document vectors can be stored conveniently by indexing ["hashing" (*24*)] each *n*-gram in a consistent manner; numerical values of vector components are stored and retrieved using these indices as pointers to memory. For the present work, I have used an 18-bit index (hash key) and ignored collisions (relatively infrequent instances of different *n*-grams being mapped to the same key).

Documents are characterized as follows: (i) Step the *n*-gram window through the document, one character at a time. (ii) Convert each *n*-gram into an indexing key. (iii) Concatenate all such keys into a list and note its length. (iv) Order the list by key value [efficient algorithms will do this in linear time (*25*)]. (v) Count and store the number of occurrences of each distinct key while removing duplicates from the list. (vi) Divide the number of occurrences of each distinct key by the length of the original list.

Because every character of a document (except for the last $n - 1$ characters) is the initial character of some n-gram (which may not necessarily be distinct from previous n-grams encountered in the same document), the number of distinct n-grams will initially closely track the document size in characters. Eventually, as the substance and tone of the document become established, fewer and fewer new n-grams will be introduced, and the initial rise will slow considerably (Fig. 1).

In gauging similarity, I make the basic assumption that two documents whose n-gram vectors are "similar" in some useful sense are likely to deal with related subject matter, and that documents whose vectors are dissimilar are likely to have little to do with one another. As a tentative first step, consider the normalized dot product S between document vectors. For documents m and n drawn from a set of size M ($m, n \in 1, \ldots, M$)

$$S_{mn} = \frac{\sum_{j=1}^{J} x_{mj}x_{nj}}{\left(\sum_{j=1}^{J} x_{mj}^2 \sum_{j=1}^{J} x_{nj}^2\right)^{1/2}} = \cos\theta_{mn} \quad (3)$$

Here x_{mj} is the relative frequency with which key j (out of a total of J possibilities) occurs in document m. The score given by Eq. 3 is the cosine of the angle θ_{mn} between two vectors in the high-dimensional document space, as viewed from the absolute origin. Points (documents) in this vector space are constrained to lie in the hyperplane (subspace) defined by Eq. 2.

This result can easily be visualized in a low-dimensional space. Taking $J = 3$ (ordinary three-space), Eq. 2 means that all document points lie in the plane $x_1 + x_2 + x_3 = 1$ (Fig. 2), and Eq. 3 is in fact the cosine of the angle between document vectors m and n as viewed from the origin.

Equation 3 can provide a gross measure of similarity—in particular, language discrimination is excellent. As a simple example (Fig. 3), I intercompared samples of text in 31 different languages (averaging about 5000 characters each) and displayed the results in a way that divulges clustering among the similarity scores (26). Similarity scores below a threshold determined by the calibration procedure described below were discarded. Within each independent class of languages, which by definition has no above-threshold links to other classes (the five African languages in the lower left corner, for example), the algorithm represents similarity by proximity (27). In this representation, two independent samples from the same language would typically be offset from one another by about 10% of the radius of one of the circular icons. Solely on the basis of their 5-gram content, these samples have been accurately grouped by language family.

The metric Eq. 3 fails at subtler tasks such as topic discrimination, however, because plain-language document vectors are usually dominated by uninformative components (in English, for example, n-grams derived from "is the", "and the", and "with the"), and the simple dot product Eq. 3 is driven primarily by the very strongest vector components. This weakness is shared by conventional word-based vector-space systems, which is why they usually use (language- and domain-dependent) stop lists.

An effective solution to this problem is to translate the origin of the vector space to a location that characterizes the information one wishes to ignore and to compute a similarity score referred to that new vantage point. Because a judiciously chosen origin can represent information common to a given set of documents, it can implicitly define the context in which discrimination is to take place. This is an important step because similarity comparisons become meaningful only when those document characteristics to which the measuring system is sensitive (be they, for example, overall formatting, alphabet type, specific language, or primary topic) have been identi-

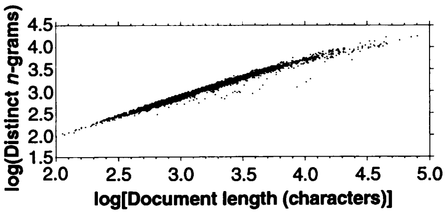

Fig. 1. Number of distinct n-grams (n = 5) as a function of document length for 5050 broad-ranging English-language magazine articles.

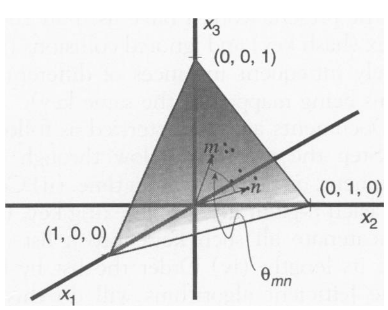

Fig. 2. A realization of the similarity score (Eq. 3) in a three-dimensional document space. Documents are constrained to lie in the plane $x_1 + x_2 + x_3 = 1$.

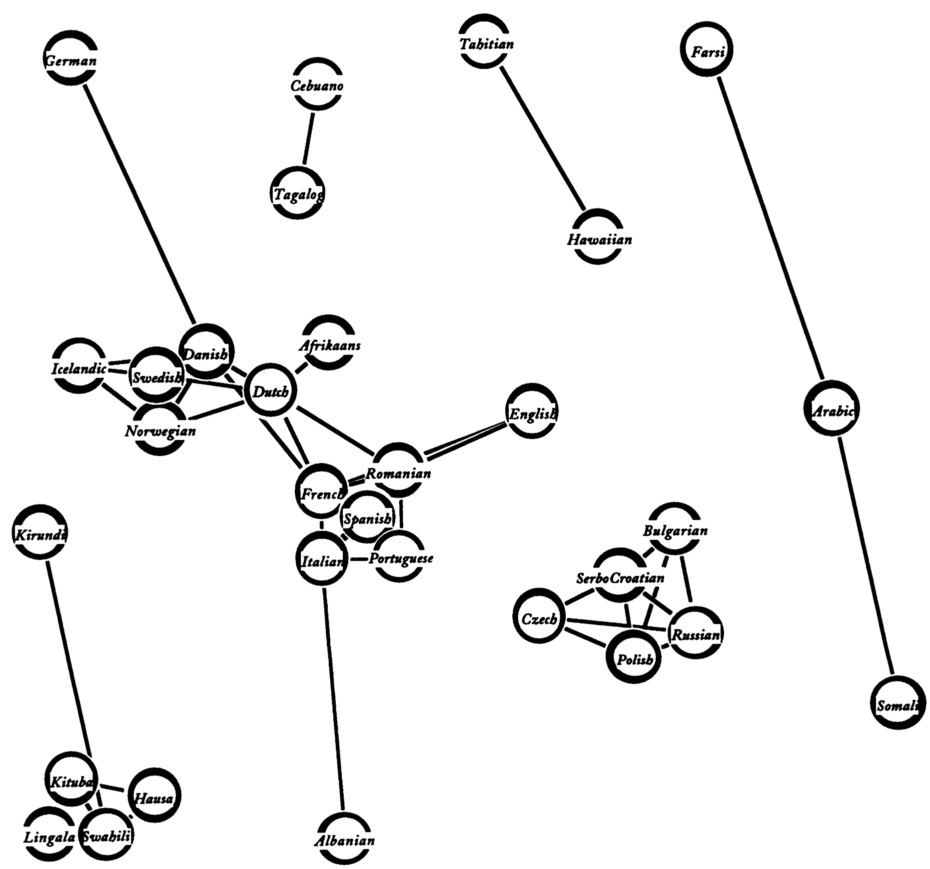

Fig. 3. Clustering of 31 language samples based solely on the normalized dot product (Eq. 3). Proximity within each of the six disjoint classes connotes similarity (only relative distance is meaningful; there are no axes).

fied and accounted for.

One straightforward way to specify this new origin is to average a particular set of document vectors, for example, the document vectors that belong to an identifiable cluster. We may call this average the centroid of the set, whether it be the mean, median, or some other measure of commonality (it is an open question as to which of these might be most effective in any specific application; preliminary experiments indicate that the end result is not a sensitive function of the adopted measure). Note that each of the many axes in the document vector space is associated with a unique n-gram (apart from an arbitrarily small collision factor), and that the proposed transformation is neither a rescaling nor a rotation of the axes, merely a translation. Consequently, documents that belong to sets dominated by nominally different primary topics (for example, health care reform and communicable diseases) but that are related to one another at a subordinate topic level (for example, AIDS epidemiology) can still be recognized as such if the various primary clusters are first "superimposed" by subtracting the corresponding cluster centroid from the documents that define each cluster (thereby centering the clusters themselves about a single origin).

For simplicity, I have chosen the centroid of a given set to be the arithmetic mean of its document vectors. In J-dimensional space, intercomparison of the M documents of one set, represented by vectors $\mathbf{x}_m$, $m \in 1, \ldots, M$, with the N documents of a second set (the second set may of course be identical with the first), represented by vectors $\mathbf{y}_n$, $n \in 1, \ldots, N$, yields the modified score

$$S_{mn} = \frac{\sum_{j=1}^{J}(x_{mj} - \mu_j)(y_{nj} - \nu_j)}{\left[\sum_{j=1}^{J}(x_{mj} - \mu_j)^2 \sum_{j=1}^{J}(y_{nj} - \nu_j)^2\right]^{1/2}} = \cos\theta_{mn} \qquad (4)$$

where

$$\mu_j = \frac{1}{M}\sum_{m=1}^{M} x_{mj} \text{ and } \nu_j = \frac{1}{N}\sum_{n=1}^{N} y_{nj} \qquad (5)$$

Let $\mathbf{x}' = \mathbf{x} - \boldsymbol{\mu}$ and $\mathbf{y}' = \mathbf{y} - \boldsymbol{\nu}$ ($\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are the centroid vectors associated with the two documents $\mathbf{x}$ and $\mathbf{y}$, respectively). By definition, then

$$\sum_{j=1}^{J} x'_j = 0, \qquad \sum_{j=1}^{J} y'_j = 0 \qquad (6)$$

These two conditions constrain all document vectors to a hyperplane parallel to the one shown in Fig. 2 but passing through the absolute origin. The superimposition of clusters referred to above is enforced by the vector differences in the numerator and denominator of Eq. 4.

The score defined by Eq. 4 is sensitive to "noise" (for example, garbled text, stylistic differences among authors, and residual fluctuations in common elements after subtraction of the centroid) in documents that lie close to the origin, with the result that small perturbations of a document vector can drastically alter its similarity scores with other documents (because a small change in some vector component can cause a large change in θ_{mn}). However, because a document close to the origin (in terms of some typical cluster radius) contains little information beyond that represented by the origin itself, and in that sense can be considered fully characterized, there is little to be gained by pursuing more subtle comparisons involving that document. In practice, one can penalize scores involving such documents by associating an overall multiplicative factor with every document vector, such that the closer the document is to the current origin, the smaller the factor becomes. The net result is a similarity score $\lambda_m\lambda_n S_{mn}$, where λ_m and λ_n go smoothly from 0 to 1 as the length of the respective document vector increases (28), with

$$\lambda = f(r), \qquad r = \sqrt{\sum_j x_j^2} \qquad (7)$$

How are these similarity scores distributed over a wide-ranging corpus of documents? Can one distinguish among related documents, unrelated documents, and intermediate cases (29)? Rather than rely on human judges to establish "ground truth," I create a set of closely related document pairs by partitioning each member of a test collection in a special way, extracting alternate sentences into two new "twin" documents (22). With such a test set, one can establish the performance of a similarity metric against documents known to be highly similar (twins) and documents assumed on average to be dissimilar (nontwins). The behavior of intermediate cases can plausibly be interpolated into a number of broad similarity categories, each with associated confidence levels.

For the present test, I partitioned 4000 diverse English-language magazine articles chosen at random from a commercial full-text CD-ROM. The original articles were at least 1000 characters long and produced twins containing at least 20 sentences apiece. Each of these altered documents was compared with all others by its modified score (4), producing two score distributions (document versus twin, document versus nontwin) (Fig. 4). If in fact these test sets faithfully model strongly related and unrelated documents, then the risk of severely misclassifying strongly related documents by calling them totally unrelated (and vice versa) is low (less than 1%). By interpolating between the two distributions, one can likewise reduce the likelihood of subtler misclassifications to an acceptable level.

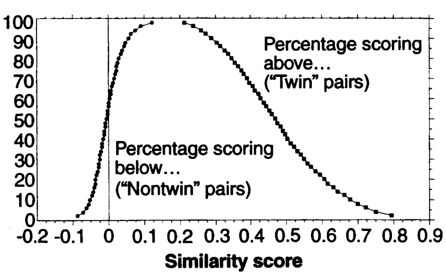

Fig. 4. Cumulative distributions of test-pair scores based on a broad sample of English-language magazine text.

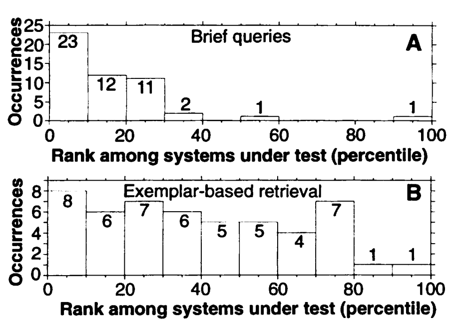

Fig. 5. Rank among retrieval systems versus number of queries in which that rank was achieved for (**A**) the ad hoc and (**B**) routing tasks.

Formal Evaluation

The annual Text Retrieval Conference (TREC) sponsored by the National Institute of Standards and Technology (30) is a well-attended forum for the comparison of document retrieval and categorization methodologies, with more than 90 participants in 1994 from government, industry, and academia. Two high-volume test protocols (run against over a million broad-ranging English-language documents) have been devised to assess participating systems, which are ranked according to their conformity with human evaluators' judgments of the "relevance" of retrieved documents to a prescribed set of queries. The results of these assessments are characterized in terms of recall (the fraction of all "relevant" documents in a corpus that are actually retrieved) and precision (the fraction of the documents retrieved that are tagged "relevant"). Obviously, the significance of such results depends on the care with which "relevance" is defined and determined (31).

The purpose of taking part in this year's TREC activity (32) was to compare the performance of Acquaintance with that of state-of-the-art document retrieval systems. The two main tasks undertaken by participants were (i) document retrieval based on brief stylized descriptions of the desired documents (known as the ad hoc task), and (ii) document retrieval based on the full text of exemplars certified to be of interest (known as the routing task).

The performance of Acquaintance on the ad hoc task is shown in Fig. 5A. Out of 50 queries considered, the measured precision exceeded the median (across 34 participating systems) only twice. It was far below the median in almost all other cases, although it fared better than 10% of the participants more than half the time. Such queries fail to model desired documents well enough to serve as the sole input for retrieval by Acquaintance.

Performance on the exemplar-based routing task is plotted in Fig. 5B. Acquaintance scored at least as well as half of the 34 systems addressing this task in more than one-third of the queries (18 out of 50); it scored better than two-thirds of the systems in one-fifth of the queries (10 out of 50).

In the latter task, and in terms of these widely adopted metrics, it would appear that Acquaintance can perform on a par with some of the best existing retrieval systems. Aside from the utter simplicity of its approach, however, one feature that sets it apart is its complete language independence: At no time was the system informed that it was processing English. Not surprisingly, then, useful practical results have been obtained during the past several years of development, with no modification of the algorithm, in close to two dozen languages.

Examples

The following three examples illustrate the use of Acquaintance as a basis for blind clustering, the grouping of documents according to language, topic, and subtopic (and even finer subdivisions) with no prior information about document content. Other applications—such as sorting (redirecting documents, that is, forwarding them to end users, in accordance with previously specified categories), categorization (labeling documents, but not necessarily redirecting them, in accordance with previously specified categories), and retrieval—can be viewed as procedural modifications of this basic process. In sorting and categorization, incoming documents are compared with previously established references, which may be existing documents or groups of documents (and which may well have been identified by previous stages of processing).

For retrieval, Acquaintance facilitates an iterative refinement process in which one maps relationships and labels entire clusters of documents at each stage of retrieval. (It is not necessary to intercompare all documents in a large corpus beforehand because relationships can quickly be mapped among just the documents retrieved at any stage.) Query-matching requirements in the initial stage of a search can be relaxed significantly (so as to enhance recall), and entire clusters of inappropriate documents can be discarded strictly on the basis of their labels (thereby enhancing precision). If appropriate documents are identified in this way, they can be merged and resubmitted as a far more focused follow-up query, and the mapping and labeling procedure can be repeated. In the process, the context (that is, the origin in document space) can be redefined after each round of retrieval, if desired, enhancing discrimination in successive rounds (*33*).

For the first example, Acquaintance processed two days' worth of Associated Press (AP) wire service news articles (19 and 20 February 1990) from the TREC collection. All possible pairs of articles were intercompared, and the resulting scores were used to construct Fig. 6, which reflects the interrelationships among a subset of the 392 articles (*26*). Guided by a related *n*-gram text-profiling technique (*34*), clusters were automatically labeled with an informative set

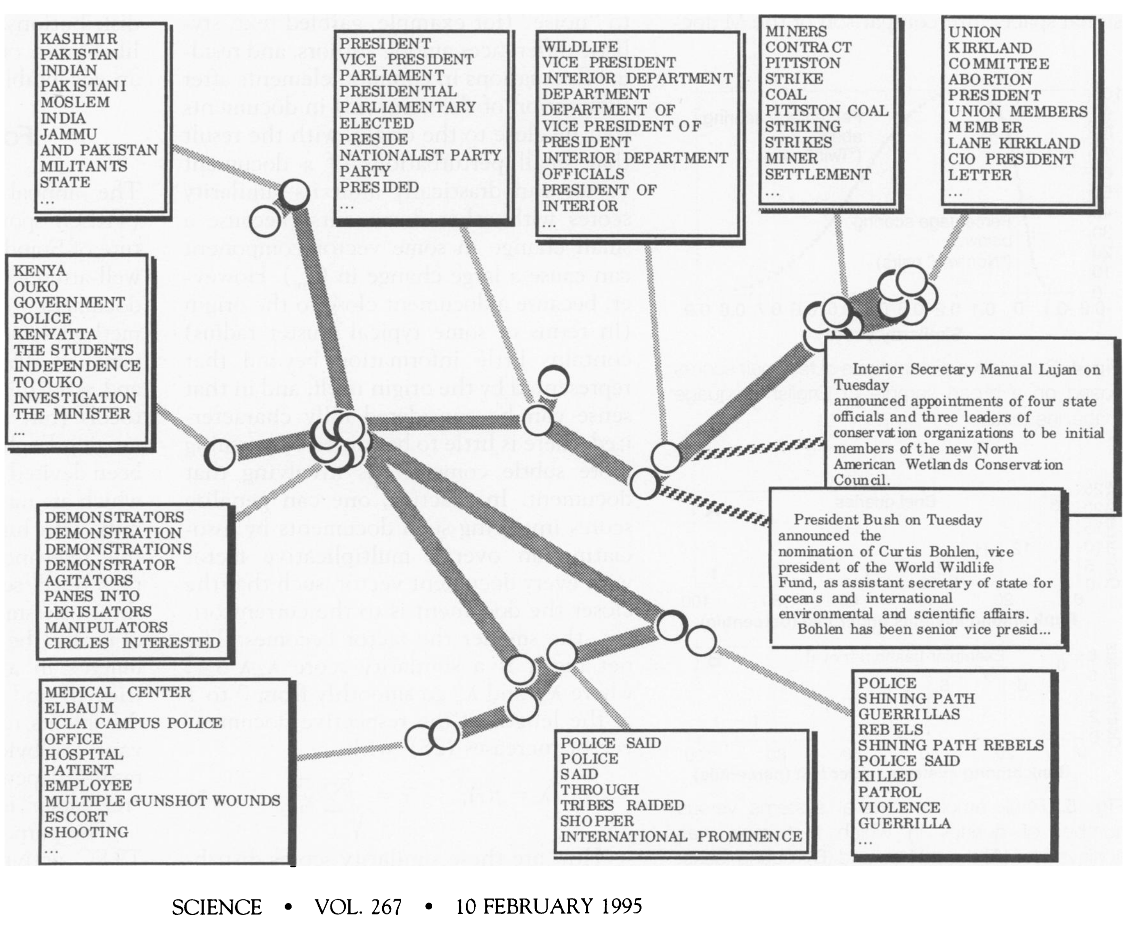

Fig. 6. A subset of AP wire service articles automatically grouped by topic with no prior information on document content. The clusters have been labeled with *n*-gram–derived word and phrase highlights (*34*). In addition, text excerpts label two of the articles to the right of center.

of "highlights" (words and phrases that characterize a document or group of documents in the context of a specified set; the context in this case was the full set of 392 articles). The resulting labels clearly delineate the topic or topics addressed by those documents. Two ostensibly related articles have been further labeled (by explicit user request) with excerpted text.

To illustrate the robustness of the algorithm in the face of severe degradation in text quality, I compared the roster of a particular cluster of news items (museum and art gallery announcements) identified within the full data set used in the preceding example with that of the corresponding cluster in an artificially corrupted version of that same text (the clean and corrupted text versions of the file were processed completely independently of one another). The character error rate was approximately 15%: On average, 15 of every 100 original characters were randomly modified either by substitution, addition, or deletion. The cluster of corrupted announcements was found still to contain 17 of the 20 members originally found in the clean version. Figure 7 shows four corresponding pairs from the two clusters.

Figure 8 illustrates the results of blind clustering of Japanese newspaper text. The test set of 394 articles was drawn from the Defense Advanced Research Projects Agency (DARPA) TIPSTER Japanese collection [16-bit shifted Japanese Industrial Standards (JIS) code], which deals exclusively with the general topic of joint ventures. Successful discrimination must therefore take place at what would commonly be viewed as the subtopic level or better.

The figure provides a view into one branch of a set of 119 computer-related articles (other prominent subject areas covered by the full set included joint ventures between Japanese and German auto manufacturers, and the world of *haute couture*). The articles in Fig. 8 relate mainly to computer memory and report on developments among companies such as Intel, Motorola, NKK, and Siemens. The articles within each cluster are mutually consistent to the same extent as the English-language examples shown above.

Summary

The capabilities of the Acquaintance technique include (i) language sorting based solely on brief reference samples and an "unknown" sample some dozens of characters long, (ii) topical partitioning of a large collection of documents in any language, which need not be specified in advance, with no prior specification of subject matter and with no user intervention, (iii) document sorting in any language, based on reference documents constructed in a self-consistent manner, and (iv) natural-language exemplar-based document retrieval in any language, at a performance level that compares favorably with state-of-the-art retrieval systems. These derive directly from an n-gram–based measure of document similarity and a means of automatically mitigating the effects of uninformative text. The latter has made possible a practical definition of the context in which document similarity is to be judged and is a generally applicable vector-space technique, rather than being specifically tailored to n-grams.

An off-the-shelf personal computer can fully intercompare hundreds to thousands of documents in a matter of minutes (35). The time for full intercomparison is of course quadratic in the number of items compared, but the iterative refinement process described above obviates the need for full intercomparison of more than several thousand documents at a time in all but the very highest volume applications. Sorting and retrieval are linear in the number of documents investigated.

The time required to set up and tune a specific application is measured in minutes to hours, rather than weeks to months. The approach lends itself well to experimentation and is straightforward to adapt to a wide variety of problems. Research into the characteristics, capabilities, and limitations of this technique is ongoing.

The results obtained thus far underscore the desirability of distinguishing between topic and meaning in text processing. Although words in isolation may well be ambiguous, one can assume that documents like the ones discussed here are intended to communicate useful information. The author of a coherent document will attempt to mitigate ambiguities by supplying related words and phrases—and therefore n-

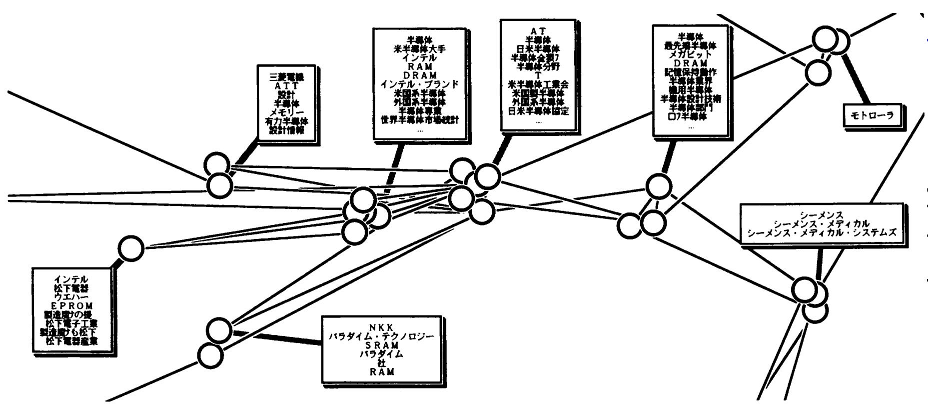

Fig. 8. One branch of a set of computer-related Japanese newspaper articles. The articles in this branch deal mainly with computer memory hardware.

Fig. 7. Comparison of four articles from a prominent cluster with their counterparts, which were also grouped into a significant cluster. The average character error rate in the latter case is 15%.

```
     High Museum of Art:
     ``French Ceramics: Masterpieces From Lorraine.'' Through Jan. 6.
     ``Poster Art of the Soviet Union: A Window Into Soviet Life.''
Through Feb. 8.

     Baltimore Museum of Art:
     ``Ndebele Beadwork.'' Through Jan. 13.
     ``Lalique: A Century of Glass for a Modern World.'' Through Jan.
13.

     Walters Art Gallery:
     ``Islamic Art and Patronage: Selections From Kuwait.'' Through
Feb. 17.

     Museum of Fine Arts:
     ``Rediscovering Pompeii.'' Through Jan. 27.
     ``Adolph Menzel, 1815-1905: Master Drawings From Berlin.''
Through Jan. 27.
     ``The Pen and the Sword: Winslow Homer, Thomas Nast and the
American Civil War.'' Through Feb. 3.
     ``The Sculpture of Indonesia.'' Through March 17.

     Nelson-Atkins Museum of Art:
     ``Organic Abstraction.'' Through Feb. 10.
     ``The Modern Poster: The Museum of Modern Art.'' Through Feb.
10.
     ``South Asian Textiles From the Permanent Collection: Woven
Patterns.'' Through Feb. 17.
```

```
     Hgh Musum of ArrIt:
     ``French Cewraamic: Masterpi4eces F@rom; Lorraine.b'N' Through
Jan. 6.
     `~`Poster XIA~rFtt ofU thC bSovietH Uion: A Windw Ito Soiet
Life.''l Thoug~h Feby.N 8.

     J N +BaltimoeMuseum of Art:
H o  ``NdbLl{e Beadwork.'('% Thrjough Jan 13.  ``Lliqqu~:1 A Centuy of
GlYa=s for a Modern &W#orld.'' Throug Jan).w 13.u $
      Wate#rRs Ar Galler:
     ``Islamic A[r!t and Pat[rqoHnbag: SlecUt7ions From Kuwait.''
Through Feib. 17.

     M Museum Wof Fu"i!nez aArts:
     ``RediscoeriDn}g PoWmpei.'' Troungh Jan. 27.
     `~`Adolph M(nbzl, 181{5c-1905: Mast%er Drawvings FKrom/
uBerlin.o'7'
hrough cJUan. 27.
     ``The+P@e;nY[amndd te S%w5or{:7E MWinldoG*wc Homer, Thomas@NNast
and@ th
American CliSviWar.' Through Fe.n3.
     ``The Sculkpture of IndoneWsXia'' ThroughS =Marc+h 17.

     ! oNelson-Atkins Museum of ArtV:s
     ``Organic bstraction.'' iTkhrUorugh Feb. 0.
     ``The ModePrfn Posterj:L HThLe Musge6um of\ 2Mo.d8ern
rt.'=hyrougWh FebH. 120.
     i v``Sou$thAsan TexXt+iles\ 'From ith Permanent Collection:
W,oIven PattGernnms.' Thrfough Feb. t1(7.
```

grams—as necessary, no matter what the language. The evidence presented here suggests that the mere presence of such terms, rather than their linguistic interrelationships, can usefully constrain the topic.

The distinction between topic and meaning is of practical interest because many document-handling tasks are facilitated by the ability to sort solely according to topic, deferring the appreciation of meaning to a late stage of processing. We now possess a versatile tool to do just that.

REFERENCES AND NOTES

1. G. Salton, *Science* **253**, 974 (1991).
2. ——— and C. Buckley, *ibid.*, p. 1012.
3. G. Salton, J. Allan, C. Buckley, A. Singhal, *ibid.* **264**, 1421 (1994).
4. "The first step of any language analysis system is necessarily recognizing and identifying individual text words." G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer* (Addison-Wesley, Reading, MA, 1989), p. 379.
5. C. E. Shannon, *The Mathematical Theory of Communication* (Univ. of Illinois Press, Urbana, IL, 1949).
6. ———, *Bell Syst. Tech. J.* **30**, 50 (1951).
7. The term *n*-gram has occasionally been used to mean a sequence of *n* words or other composite units. Throughout this paper, except for the Japanese example, *n*-gram will refer exclusively to a sequence of *n* characters; in the case of Japanese, it refers to *n* bytes of text.
8. C. Y. Suen, *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 164 (1979).
9. A. Zamora, *J. Am. Soc. Inf. Sci.* **31**, 51 (1980).
10. J. L. Peterson, *Commun. ACM* **23**, 676 (1980).
11. E. M. Zamora, J. J. Pollock, A. Zamora, *Inf. Process. Manage.* **17**, 305 (1981).
12. J. J. Hull and S. N. Srihari, *IEEE Trans. Pattern Anal. Mach. Intell.* **4**, 520 (1982).
13. J. J. Pollock, *J. Doc.* **38**, 282 (1982).
14. R. C. Angell, G. E. Freund, P. Willett, *Inf. Process. Manage.* **19**, 255 (1983).
15. E. J. Yannakoudakis, P. Goyal, J. A. Huggill, *ibid.* **18**, 15 (1982).
16. J. C. Schmitt, U.S. Patent No. 5,062,143 (1990).
17. W. B. Cavnar and J. M. Trenkle, "*N*-Gram-Based Text Categorization," *Proceedings of the 1994 Symposium on Document Analysis and Information Retrieval* (Univ. of Nevada, Las Vegas, 1994), p. 161.
18. P. Willett, *J. Doc.* **35**, 296 (1979).
19. C. P. Mah and R. J. D'Amore, "DISCIPLE Final Report," *PAR Rep. 83-121* (PAR Technology Corporation, New Hartford, NY, 1983).
20. J. Scholtes, in *International Joint Conference on Neural Networks, Singapore* (IEEE Press, New York, 1991), vol. 1, p. 95.
21. W. B. Cavnar, in (*30*), pp. 171–179.
22. S. M. Huffman, in preparation.
23. The Japanese text is encoded as two-byte integers (specifically, 16-bit shifted JIS code), and the current processing system is byte-oriented; a 6-gram in Japanese therefore refers to a sequence of three kanji or kana.
24. D. E. Knuth, *Sorting and Searching*, vol. 3 of *The Art of Computer Programming* (Addison-Wesley, Reading, MA, 1973), sec. 6.4.
25. T. H. Cormen, C. E. Leiserson, R. L. Rivest, *Introduction to Algorithms* (McGraw-Hill, New York, 1990), sec. 9.3.
26. J. D. Cohen, in preparation.
27. The visualization algorithm assumes that the various relationships are transitive: The fact that French resembles Spanish and that Spanish resembles Portuguese, for example, strengthens the imputed relationship between French and Portuguese. Only relative distance is meaningful in Fig. 3, that is, there are no axes in this diagram. Note that Fig. 3 does not represent the same space as Fig. 2. It is effectively a distillation in two dimensions of some of the useful information found in the high-dimensional document vector space.
28. The specific form used for the work illustrated here is $f(r) = \{1 + \exp[(r_0 - r)/\Delta]\}^{-1}$, but any similarly behaved function would likely do as well. Note that Eq. 2 implies that the maximum possible Euclidean length of a vector is 1.
29. To skirt the contentious issue of "relevance" [L. Schamber, M. B. Eisenberg, M. S. Nilan, *Inf. Process. Manage.* **26**, 755 (1990)], I adopt an operational definition of "strong relatedness": Two documents are strongly related if (but not only if) they have been produced by the document-splitting procedure described in this section. As with all operational definitions, the utility of this definition is ultimately to be judged by the utility of its consequences.
30. D. K. Harman, Ed., *The Second Text Retrieval Conference (TREC-2)* (NIST Spec. Publ. 500-215, National Institute of Standards and Technology, Gaithersburg, MD, 1994).
31. "The possibility, for example, that two irrelevant documents might become relevant if put together has never been adequately considered, so far as I know." D. R. Swanson, *J. Am. Soc. Inf. Sci.* **26**, 755 (1990).
32. D. K. Harman, Ed., *The Third Text Retrieval Conference (TREC-3)* (National Institute of Standards and Technology, Gaithersburg, MD, in preparation).
33. This process is ideally suited to the ad hoc task in the TREC environment but was not implemented in time for the 1994 conference.
34. J. D. Cohen, *J. Am. Soc. Inf. Sci.* **46**, 162 (1995).
35. In a typical run, 941 newspaper articles were exhaustively compared with one another, generating a total of 443,211 pairwise scores, in 726 seconds, including the time required to save the scores to disk (Apple Macintosh Quadra 950).

36. I thank J. D. Cohen, S. M. Huffman, J. M. Kubina, and C. Pearce for their enthusiastic contributions to this project, and D. E. Brown and M. W. Goldberg for their support and encouragement. The technique described in this paper is the subject of a pending U.S. patent and of France's patent no. 2,694,984.