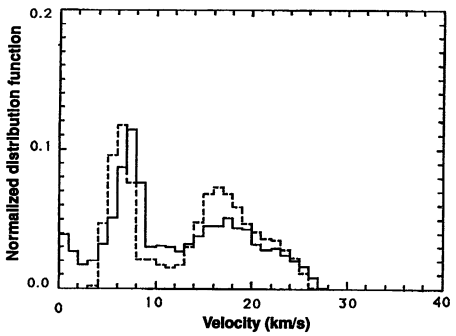

Fig. 4. Velocity distribution function for H atoms leaving the inner coma of comet Halley. The dashed line shows the distribution of velocities of H atoms as initially produced by photodissociation. The solid line shows the actual distribution function for atoms leaving the inner coma after partial collisional thermalization. Thermalization reduces the 20 km/s region and populates the region of low speeds (0 to 4 km/s), which initially contains no atoms.

analysis of these data (*10*), which was based only on the inner region of the coma and did not take into account the optical thickness of the inner coma to solar Lyman-α radiation. The MCPTM analysis of the entire extended image shows that the model (corrected for the optical thickness that occurs only in the inner coma) self-consistently reproduces the two-dimensional shape and gradient of the whole observable inner and outer coma. This result implies that the entire PVOUVS Lyman-α data set should be reevaluated, because all of the production rates determined from the Lyman-α data are likely to be systematically too low, at least where the derived production rates are large. The agreement of the model and data in Fig. 3 is most gratifying in that it both verifies and documents the advantage and value of using the physical model. Together with the analysis of comet Kohoutek (*6*), this analysis represents the second major and successful application of the full H MCPTM for a comet for which the gas production rate is sufficiently large that significant collisional thermalization occurs in the inner coma. Figure 4 shows the distribution in phase space of the velocities of H atoms leaving the inner coma as they are initially produced by photodissociation and after they are partially collisionally thermalized. The MCPTM naturally produces the correct number of low-speed H atoms that are required to explain the shape of the coma. This obviates the need for using a parameterized velocity distribution in the model (*4*).

REFERENCES AND NOTES

1. M. C. Festou, *Astron. Astrophys.* **96**, 52 (1981).
2. D. G. Schleicher, thesis, University of Maryland (1983).
3. E. F. van Dishoeck and A. Dalgarno, *Icarus* **59**, 305 (1984). The initial ejection speed distribution for H has been modified according to the new water solar photodissocation results of J. Crovisier, *Astron. Astrophys.* **213**, 459 (1989).
4. R. R. Meier *et al.*, *Astron. Astrophys.* **52**, 283 (1976).
5. M. R. Combi and W. H. Smyth, *Astrophys. J.* **327**, 1026 (1988).
6. ———, *ibid.*, p. 1044.
7. M. R. Combi, *Icarus* **81**, 41 (1989).
8. A. I. F. Stewart, *IEEE Trans. Geosci. Remote Sensing* **GE-18**, 65 (1980).
9. R. P. McCoy, C. B. Opal, G. R. Carruthers, *Nature* **324**, 439 (1986). The other two Lyman-α images were acquired by rocket payload instruments on 24 February and 13 March 1986.
10. A. I. F. Stewart, *Astron. Astrophys.* **187**, 369 (1987).
11. J. M. Ajello, *J. Geophys. Res.* **95**, 14855 (1990).
12. The data points were mapped into a polar coordinate system in which the polar angle was measured from the plane containing the comet, the sun, and the Pioneer Venus spacecraft. The radial coordinate was the perpendicular distance from Halley's nucleus to the instantaneous line of sight of the ultraviolet spectrometer. A grid of square sort boxes was laid out on the mapping plane; data points falling within the same box were averaged together, and empty boxes were filled by interpolation.
13. M. R. Combi, A. I. F. Stewart, W. H. Smyth, *Geophys. Res. Lett.* **13**, 385 (1986).
14. M. B. McElroy and Y.-L. Yung, *Astrophys. J.* **196**, 227 (1975).
15. M. C. Festou *et al.*, *Nature* **321**, 361 (1986).
16. R. Zwickl, personal communication.

17. The PVOUVS data used here were obtained and reduced under National Aeronautics and Space Administration (NASA) contract NAS2-12318. The modeling analysis and interpretation of the data were supported by the Planetary Atmospheres program of NASA under contracts NASW-3949 and NASW-3387. An earlier version of the Lyman-α image was released to the press in 1986. It was prepared by A. Jain of the University of California at Davis using data supplied by one of us (A.I.F.S.).

14 March 1991; accepted 25 June 1991

Allerød–Younger Dryas Lake Temperatures from Midge Fossils in Atlantic Canada

IAN R. WALKER,* ROBERT J. MOTT, JOHN P. SMOL

Remains of freshwater midges are abundant in lake sediments, and their species distributions are closely related to the surface-water temperature of lakes; their distributions thus provide a powerful tool for paleoclimatology. The distribution of species in a core from Splan Pond in Atlantic Canada indicates that there were abrupt transitions in late-glacial temperatures between warm and cold states. The transitions are correlative with the well-known warm Allerød and cold Younger Dryas events in Europe. These data thus confirm the inference from palynological data that these events affected regions on both sides of the Atlantic.

THE ALLERØD–YOUNGER DRYAS event, a reversion from the relatively warm climate of the Allerød before 11 ka (thousand years ago) to the much cooler conditions of the Younger Dryas between approximately 11 and 10 ka, is well documented in Europe (*1*). The occurrence of late-glacial temperature fluctuations in Atlantic Canada that are correlative with the European event has been suggested on the basis of palynological and lithological evidence (*2–4*). In Atlantic Canada, from 11 to 10 ka, deposition of organic-rich sediments was interrupted by reversion to mineral deposition in many lakes (*3*). Pollen evidence is interpreted as indicating a reversion to more open vegetation at this time in response to a colder climate (*3*). Nevertheless, other independent records are needed to confirm this interpretation. Although evidence is accumulating for the occurrence of these events in North America outside Atlantic Canada, the evidence is still weak and widely scattered (*5–9*). We therefore used a newly developed technique, weighted averaging calibration of fossil midges to temperature (*10*), to study paleotemperatures in a lake in Canada in which palynological data have indicated that these events occurred.

Two sediment cores with nearly identical stratigraphy were removed from Splan Pond (45°15′15″N, 67°19′50″W), New Brunswick, Canada; this lake [also called Basswood Road Lake (*2, 3, 11*)] is less than 30 km from the coast, near the international border with the United States. The first core (MS 68-27) was sampled for lithostratigraphy and palynostratigraphy (*11*). To evaluate independently the presence of a late-glacial climatic oscillation, we analyzed sediments from the second Splan Pond core (MS 78-3) for fossil aquatic midges. Midge fossils (Diptera: Chironomidae, Ceratopogonidae, and Chaoboridae) were isolated, identified, and enumerated (*10, 12*) from 13 levels in the Splan Pond core [spanning the intervals Mott *et al.* (*3*) believed to represent the Allerød and Younger Dryas].

I. R. Walker and J. P. Smol, Department of Biology, Queen's University, Kingston, Ontario, Canada K7L 3N6.
R. J. Mott, Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, Canada K1A 0E8.

*To whom correspondence should be addressed at Department of Biology, Okanagon College, 1000 K.L.O. Road, Kelowna, British Columbia, Canada V1Y 4X8.

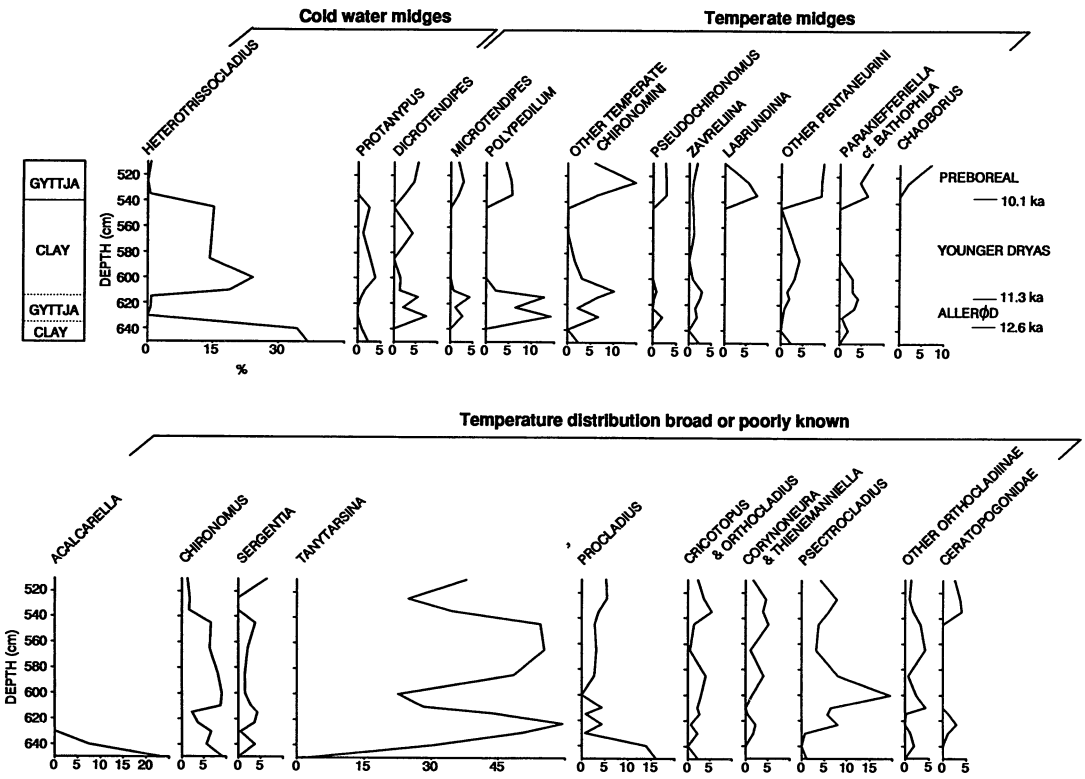

Fig. 1. Percentage abundance of common midge taxa in sediments of Splan Pond, New Brunswick, Canada. For comparison, names of climatic events for correlative European time intervals are included.

Distinct fluctuations in the relative abundance of midge taxa are evident in the data (Fig. 1). Present-day distributions of these midges show that species occurrences are related to surface-water temperature (*10, 13, 14*). Cold-water midges (*Heterotrissocladius* and *Protanypus*) were abundant before 12.5 ka and between 11 and 10 ka but were rare or absent during the interval ~12.5 to 11 ka and after 10 ka. In contrast, many temperate midges were most abundant between 12.5 and 11 ka and after 10 ka (*15*).

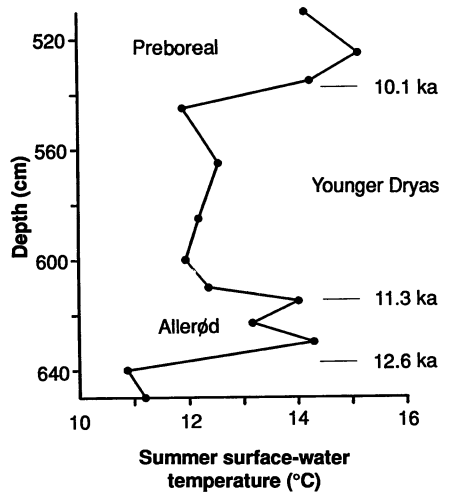

Fig. 2. Summer surface-water paleotemperature reconstruction for Splan Pond. For comparison, names of climatic events for correlative European time intervals are included. The apparent root-mean-square error of the temperature estimates is 1.32°C (*10*).

To further investigate summer surface-water paleotemperature conditions, we used a weighted-averaging transfer function (*10, 16*). Weighted-averaging calibration has been used for many quantitative paleoenvironmental reconstructions; it is theoretically more sound and performs better than many other reconstruction techniques, such as multiple regression (*17, 18*). The paleotemperature reconstruction for Splan Pond indicated that there were three abrupt temperature changes from 12.5 to 10 ka (Fig. 2). Splan Pond was cold during summer and had a surface-water temperature of about 11°C before 12.5 ka. Estimated summer temperatures rose to 14°C during the interval 12.5 to 11 ka before cold conditions returned; this time interval spans that of the Allerød (a warm late-glacial interval) in Europe (*1, 19*). Summer surface-water temperature estimates for the period 11 to 10 ka vary from 11.9° to 12.6°C. This is also the estimated time of the Younger Dryas in Europe (*1, 19*). Thus, the Splan Pond data show that these intervals of warm and cold climate affected both shores of the Atlantic. At Splan Pond, warmer summer temperatures (14° to 15°C) returned at the beginning of the Holocene (~10 ka). The temperature changes bounding the Allerød–Younger Dryas events are abrupt, consistent with rapid shifts between warm and cold climatic states, as envisaged by Broecker *et al.* (*20*).

On the basis of the midge data alone, it is impossible to conclude that the water temperature of Splan Pond was not influenced by local glacial ice or meltwater; other sites must be investigated. Palynological and lithological evidence, however, indicates that similar changes were occurring in other lakes of Atlantic Canada (*2–4*), thus suggesting that the temperature oscillation was not a local phenomenon. Studies of Quaternary geology of New Brunswick also indicate that glaciers had retreated from the area well before the interval from 10 to 11 ka (*21*).

The midge data thus support evidence for climatic oscillations in Atlantic Canada correlative with the Allerød and Younger Dryas. Fossil midges provide sensitive paleoclimatic markers, especially near tree line in arctic and alpine regions (*10*). Although paleoclimatic data are available from diverse evidence, including isotopic records, geomorphological evidence, and floral and faunal remains, paleoclimatic reconstructions for continental regions have relied mainly on data from one source: palynology. Insects and most aquatic organisms have short lifespans and efficient means for dispersal; they respond more rapidly than vegetation to climatic change (*22, 23*). Paleoclimatic proxy data from midges should be routinely used in combination with other lines of evidence to document past climatic change. Midge fossils offer an independent and quantitative test of paleoclimatic scenarios reconstructed from other indicators.

Understanding natural climatic variability has gained new urgency as we ponder the direction and magnitude of future global climatic change. As Schindler *et al.* (*24*) have shown, lake surface temperatures are likely to vary as a result of global warming. Evidence of past changes in lake temperature, such as shown here, is important for understanding the response of lakes to future anthropogenic climatic change.

REFERENCES AND NOTES

1. W. A. Watts, in *Studies in the Lateglacial of North-West Europe*, J. J. Lowe, J. M. Gray, J. E. Robinson, Eds. (Pergamon, Oxford, 1980), pp. 1–21.
2. R. J. Mott, *Syllogeus* **55**, 281 (1985).
3. _____, D. R. Grant, R. Stea, S. Occhietti, *Nature* **323**, 247 (1986).
4. R. R. Stea and R. J. Mott, *Boreas* **18**, 169 (1989).
5. H. E. Wright, Jr., *Quat. Sci. Rev.* **8**, 295 (1989).
6. D. M. Peteet, in *Abrupt Climatic Change—Evidence and Implications*, W. H. Berger and L. D. Labeyrie, Eds. (Reidel, Dordrecht, 1987), pp. 185–193.
7. D. M. Peteet *et al.*, *Quat. Res.* **33**, 219 (1990).
8. W. S. Broecker *et al.*, *ibid.* **30**, 1 (1988).
9. D. R. Engstrom, B. C. S. Hansen, H. E. Wright, Jr., *Science* **250**, 1383 (1990).
10. I. R. Walker, J. P. Smol, D. R. Engstrom, H. J. B. Birks, *Can. J. Fish. Aquat. Sci.* **48**, 975 (1991).
11. R. J. Mott, *Can. J. Earth Sci.* **12**, 273 (1975).
12. For midge analysis, 1 to 8 ml of sediment were deflocculated in warm 5% KOH and sieved on a 95-μm mesh. Head capsules retained on the sieve were hand sorted from the sediment at ×50 magnification from a Bogorov counting tray. Relative abundances of midge taxa are based on a minimum of 50 chironomid head capsules. For a more detailed summary of methods, see (*10*).
13. I. R. Walker and R. W. Mathewes, *J. Paleolimnol.* **2**, 61 (1989).
14. I. R. Walker, *Hydrobiologia*, in press.
15. The core chronology is based on ^{14}C dates reported (*3*) for core MS 68-27 (GSC-1643, 9.460 ± 0.220 ka, 5 cm of sediment collected 28 to 33 cm above the Younger Dryas clay; GSC-3862, 10.100 ± 0.130 ka, 5 cm of sediment collected immediately above the Younger Dryas clay; GSC-1645, 11.300 ± 0.180 ka, 5 cm of sediment collected immediately below the Younger Dryas clay; and GSC-1067, 12.600 ± 0.270 ka, 5 cm of gyttja deposited immediately above the basal clay), and stratigraphic correlation of lithologies between cores. Comparable deposits, distributed throughout Atlantic Canada, yield similar dates (*3*).
16. The weighted-averaging transfer function was developed from a temperature-constrained canonical correspondence analysis of midge distributions among 26 Labrador lakes. We implemented this analysis using the computer program CANOCO, Version 3.0. The Labrador lakes were distributed among arctic tundra, forest-tundra, and forest sites, with summer surface-water temperatures ranging from 7.6° to 18°C. For a more detailed summary of the temperature inference procedure, see (*10*).
17. J. M. Line and H. J. B. Birks, *J. Paleolimnol.* **3**, 170 (1990).
18. C. J. F. ter Braak and H. van Dam, *Hydrobiologia* **178**, 209 (1989).
19. T. C. Atkinson *et al.*, *Nature* **325**, 587 (1987).
20. W. S. Broecker *et al.*, *ibid.* **315**, 21 (1985).
21. V. N. Rampton, R. C. Gauthier, J. Thibault, A. A. Seaman, *Geol. Surv. Can. Mem.* **416**, 1 (1984).
22. G. R. Coope and W. Pennington, *Philos. Trans. R. Soc. London Ser. B* **280**, 337 (1977).
23. B. Ammann, *Eclogae Geol. Helv.* **82**, 183 (1989).
24. D. W. Schindler *et al.*, *Science* **250**, 967 (1990).

25. This work was funded by the Natural Sciences and Engineering Research Council of Canada, the Geological Survey of Canada, and the Atmospheric Environment Service of Environment Canada. Geological Survey of Canada Contribution 44290.

20 February 1991; accepted 4 June 1991

Global Text Matching for Information Retrieval

GERARD SALTON* AND CHRIS BUCKLEY

An approach is outlined for the retrieval of natural language texts in response to available search requests and for the recognition of content similarities between text excerpts. The proposed retrieval process is based on flexible text matching procedures carried out in a number of different text environments and is applicable to large text collections covering unrestricted subject matter. For unrestricted text environments this system appears to outperform other currently available methods.

THE PROBLEM OF TEXT RETRIEVAL from large heterogeneous text databases, in which the vocabulary varies widely and the subject matter is unrestricted, becomes increasingly important every year. These databases include newspaper articles, newswire dispatches, textbooks, dictionaries and encyclopedias, manuals, magazine articles, and so on. The normal text analysis and text indexing approaches that are based on the use of available thesauruses and other vocabulary control devices are difficult to apply in unrestricted text environments, because the word meanings are not stable in such circumstances and the interpretation varies depending on context. The applicability of more complex text analysis systems that are based on the construction of knowledge bases covering the detailed structure of particular subject areas, together with inference rules designed to derive relationships between the relevant concepts, is even more questionable in such cases. Complete theories of knowledge representation do not exist, and it is unclear what concepts, concept relationships, and inference rules may be needed to understand particular texts (*1*).

We take advice from Wittgenstein and others who suggest that text understanding must be based on a study of how text words are used in the language (*2*). In so doing, we are not primarily interested in deriving detailed descriptions of text content, but instead we want to recognize text portions within which the text meanings are homogeneous. An identification of semantically homogeneous text excerpts makes it possible to supply text links between related text portions, leading to the generation of structured text representations that lend themselves to selective reading and perusal of large texts. In addition, the recognition of related text portions leads to the retrieval of relevant texts in answer to available search requests.

In the identification of related text portions in unrestricted text environments, the texts themselves must necessarily form the basis for the text analysis. The following text comparison process is used. Each text in a collection is broken down into individual text units—for example, sections, paragraphs, and sentences. A standard automatic indexing procedure is utilized to assign to each text unit (or each available search request) a set of weighted content identifiers, or terms, collectively used to represent text content. If t terms in all are available, a given text item $\mathbf{D}_i$ may then be represented by a term vector as $\mathbf{D}_i = (w_{i1}, w_{i2} \ldots, w_{it})$, where w_{ik} is the weight of term T_k assigned to document $\mathbf{D}_i$. A weight of zero is used for terms that are absent from a particular document and a positive weight for terms actually assigned. Similarities between particular text items (or between text items and information requests) are detected by comparison of the term vectors representing the texts at various levels of detail. For example, two texts may be assumed to be related when a sufficient global similarity between them is accompanied by local similarities between included text paragraphs or text sentences.

The similarity measurement between two texts must depend on the types and weights of the coinciding terms in the respective term vectors. In retrieval, the most valuable terms for document content representation are those best able to distinguish particular texts from the remainder of the collection. Thus, the term weighting system should assign low weights to high-frequency terms that occur in many documents of a collection and high weights to terms that are important in particular documents but unimportant in the remainder of the collection. The weight of terms that occur rarely in a collection is of no consequence, because such terms contribute little to the text similarity.

A well-known term weighting system following that prescription assigns weight w_{ik} to term T_k in document $\mathbf{D}_i$ in proportion to the frequency of occurrence of a term in $\mathbf{D}_i$ and in inverse proportion to the number of documents to which the term is assigned (*3*). When the texts are represented by term vectors, the similarity (*sim*) between two items $\mathbf{D}_i$ and $\mathbf{D}_j$ is conveniently obtained as the inner product between corresponding vectors, or $sim(\mathbf{D}_i, \mathbf{D}_j) = \sum_{k=1}^{t} w_{ik}\, w_{jk}$. Thus, the similarity between two texts depends on the weights of coinciding terms

Department of Computer Science, Cornell University, Ithaca, NY 14853.

*To whom correspondence should be addressed.