TẠP CHÍ KHOA HỌC TÍNH TOÁN VÀ ĐIỀU KHIỂN TẬP VI (1990) Số 2 (1 - 8)

VỀ XÂY DỰNG HỆ THÔNG TIN VĂN PHÒNG

NGÔ TRUNG VIỆT

I - GIỚI THIỆU CHUNG

Trong vòng 10-15 năm lại đây, các vấn đề về quản lý và tìm kiếm thông tin đã là những ứng dụng chủ yếu của máy tính. Điều này được thể hiện rõ rệt qua sự phát triển nhanh chóng và việc ứng dụng rộng rãi các hệ quản trị cơ sở dữ liệu, hệ tìm kiếm văn bản. Những phát triển mới đây trong vòng 5-6 năm qua về các hệ thông tin văn phòng đã đặt ra những vấn đề mới, những yêu cầu cao hơn cho việc quản trị và tìm kiếm thông tin trên máy tính.

Về mặt lịch sử, những nghiên cứu về quản trị và tìm kiếm thông tin đã được thực hiện trên hai lĩnh vực chính là các hệ quản trị cơ sở dữ liệu và các hệ tìm kiếm văn bản. Cùng là những hệ thống quản trị và tìm kiếm thông tin nhưng mỗi hệ thống lại có một sự quan tâm khác nhau tới hai khía cạnh cơ bản này.

Trong các hệ quản trị cơ sở dữ liệu, thông tin về thế giới bên ngoài được con người trích lọc một cách thủ công và lưu trữ vào cơ sở dữ liệu một cách có cấu trúc, dưới dạng phiếu ghi hoặc cây. Mọi dữ liệu trong cơ sở dữ liệu đều phải có khuôn dạng thống nhất, phải có một cách tổ chức logic hoàn toàn xác định từ đầu và người sử dụng phải nắm vững cấu trúc này. Dữ liệu phải có nghĩa duy nhất và phải được hiểu thống nhất đối với mọi người sử dụng. Các ràng buộc chặt chẽ này kéo theo yêu cầu người sử dụng phải quen thuộc ít nhiều với tin học, được đào tạo chu đáo về tin học.

Các hệ tìm kiếm văn bản thường được thực hiện trong khuôn khổ việc tiến hành tự động hóa thư viện. Đối tượng xử lý trong những hệ thống này thường là bài báo, tạp chí, sách, luận án v.v., thường không có khuôn dạng cố định. Không thể và cũng chẳng ai muốn tạo khuôn dạng cho loại thông tin như thế. Người ta quan tâm chủ yếu tới toàn văn hơn là các từ khóa hoặc tóm tắt nội dung. Tuy vậy, từ khóa hoặc tóm tắt nội dung thường được sử dụng như một công cụ cho việc tìm kiếm. Người sử dụng có thể không cần am hiểu về tin học nhưng vẫn cần phải có người quản trị hệ thống với các hiểu biết sâu về tin học để vận hành hệ thống và giúp đỡ người sử dụng.

Các hệ thông tin văn phòng phát triển trên cơ sở máy tính xâm nhập vào văn phòng, lúc đầu chỉ thực hiện các công việc đơn lẻ như soạn thảo văn bản, theo dõi công văn v.v., sau đó tiến tới các hệ thống lưu trữ và tìm kiếm thông tin đa phương (multimedia). Thừa hưởng các kết quả đã có của các hệ quản trị cơ sở dữ liệu và tìm kiếm văn bản, các hệ thông tin văn phòng còn đặt thêm các vấn đề mới trước đây chưa có : xử lý tri thức trong văn bản, xử lý thông tin hình ảnh và tiếng nói, truyền thông báo, thư tín điện tử, v.v.. Trong phạm vi bài báo này, chúng tôi chỉ xin đề cập tới một số khía cạnh về quản trị và tìm kiếm thông tin văn bản trong các hệ văn phòng, có so sánh với các hệ quản trị cơ sở dữ liệu và tìm kiếm văn bản.

Một trong các đối tượng chính của hệ thông tin văn phòng là các báo cáo, công văn dưới dạng văn bản, nói chung không có khuôn dạng cố định. Khác với hệ thống thư viện, trong các văn phòng, lãnh đạo quan tâm chủ yếu đến tóm tắt nội dung của các văn bản vì không thể nào đọc chi tiết toàn bộ các thông tin gửi tới. Khác với các hệ quản trị cơ sở dữ liệu và tìm kiếm thông tin thư viện, người sử dụng trong các hệ văn phòng là những người không chuyên về tin học, không có người quản trị hệ thống. Họ cần

lấy thông tin nhưng không chú trọng tới việc tổ chức dữ liệu và cũng không có thời gian để học tập chi tiết cách tổ chức này. Dữ liệu trong các hệ thống văn phòng cũng mang nhiều tính cá nhân tùy thuộc người xử lý.

Việc tìm kiếm văn bản nói chung là khó khăn hơn so với việc tìm kiếm dữ liệu có khuôn dạng do không gian tìm kiếm rộng, các văn bản nói chung không có cấu trúc thuần nhất và tùy thuộc vào văn phong của người viết. Câu hỏi cho tìm kiếm văn bản cũng phức tạp hơn câu hỏi cho dữ liệu có khuôn dạng vì cùng một thông tin có thể có nhiều dạng viết khác nhau. Kỹ thuật đánh chỉ số được sử dụng rộng rãi trong các hệ cơ sở dữ liệu và tìm kiếm văn bản để giảm không gian tìm kiếm nhưng đòi hỏi phải có người quản trị thường xuyên bảo trì hệ thống nên trở thành bất tiện cho các hệ thống văn phòng.

Sau đây là liệt kê một số điểm khác biệt giữa các hệ thông tin văn phòng, hệ quản trị cơ sở dữ liệu và hệ tìm kiếm văn bản thư viện:

1. Thường xuyên thực hiện bổ sung thông tin mới trong các hệ thông tin văn phòng. Công việc này do các nhân viên văn phòng đảm nhiệm. Trong hệ thống thư viện, thông tin mới được bổ sung theo lô và do người quản trị hệ thống thực hiện. Trong hệ quản trị cơ sở dữ liệu, các thao tác nạp dữ liệu mới được thực hiện thường xuyên hoặc theo lô, cũng do người quản trị hệ thống thực hiện.

2. Các thao tác xóa bỏ hoặc sửa đổi hiếm khi được thực hiện trong các hệ văn phòng và thư viện do chỗ các thông tin đã đưa vào hệ thống có tính lịch sử, bất biến. Trái lại, trong hệ quản trị cơ sở dữ liệu, các thao tác này lại thường được sử dụng.

3. Hầu hết các tài liệu được lưu trữ trong hệ thống văn phòng gần như không được truy nhập tới. Trong các hệ thư viện và quản trị cơ sở dữ liệu, việc truy nhập vào một tài liệu bất kì là thường xuyên được thực hiện.

4. Kết quả tìm kiếm trả lời cho các câu hỏi trong các hệ thống văn phòng và thư viện được đưa lại trực tiếp cho người yêu cầu, còn trong các hệ quản trị cơ sở dữ liệu thì có thể trả về cho máy. Đối với hệ thống văn phòng, người sử dụng có thể chấp nhận câu trả lời thừa (false hit) nhưng không chấp nhận câu trả lời thiếu (false dismissals). Đối với hệ thống thư viện người sử dụng lại có thể chấp nhận câu trả lời thiếu.

5. Người sử dụng trong hệ thống văn phòng là cán bộ lãnh đạo và các nhân viên văn phòng không am hiểu về tin học. Trái lại, trong các hệ cơ sở dữ liệu và thư viện đều có người quản trị hệ thống am hiểu tường tận về hệ thống cụ thể để giúp đỡ cho người sử dụng.

Trên cơ sở xem xét các đặc thù của hệ thông tin văn phòng: đối tượng người sử dụng, yêu cầu tìm kiếm thông tin văn bản, chúng tôi thấy có hai vấn đề chính cần lưu ý khi xây dựng các hệ thông tin văn phòng:

- Vấn đề xây dựng giao tiếp thân thiện với người sử dụng,

- Vấn đề tìm kiếm các thông tin văn bản theo nội dung.

II - GIAO TIẾP THÂN THIỆN VỚI NGƯỜI SỬ DỤNG

1. Sơ lược lịch sử phát triển

Giao tiếp người - máy là vấn đề phát sinh ngay từ những ngày đầu của sự phát triển tin học. Ban đầu, không phải người ta đã ý thức ngay được tầm quan trọng của vấn đề này bởi vì các máy tính thế hệ đầu tiên chủ yếu dành cho các nhà chuyên môn sử dụng. Việc nạp thông tin vào máy được thực hiện bằng cách đặt dữ liệu luôn trong chương trình hoặc từng thời điểm máy tính tự động đọc dữ liệu từ các

thiết bị nhập. Kết quả tính toán cũng được tự động đưa ra khi hoàn thành xử lí, không cần sự can thiệp của con người. Giao tiếp người - máy trong thời kì này hoàn toàn chưa mang tính chất "thân thiện" mà chỉ là một phần việc do máy tự động thực hiện theo sự bố trí của chương trình.

Việc ra đời các ngôn ngữ lập trình bậc cao đánh dấu bước tiến bộ đáng kể trong giao tiếp người - máy. Thay vì phải tự mã hóa và giải mã các dữ liệu, các ngôn ngữ bậc cao cho phép đưa dữ liệu vào và ra dưới dạng gần ngôn ngữ tự nhiên của con người hơn. Bắt đầu xuất hiện các cơ chế trao đổi thông tin trực tiếp giữa máy tính và người sử dụng. Các chương trình lập trong các ngôn ngữ này thoát dần khỏi chế độ tự động hoạt động một mình và hướng tới chế độ đối thoại giữa người và máy. Việc đối thoại không chỉ đơn thuần là cung cấp dữ liệu cho máy khi cần thiết mà còn tham gia dần vào quá trình điều khiển chương trình.

Bắt đầu từ ngôn ngữ BASIC, vấn đề lập trình theo menu được nhấn mạnh, phát triển và mở rộng. Một yêu cầu mới xuất hiện: người làm chương trình phải lưu ý đến người sẽ sử dụng chương trình của họ. Đối tượng người sử dụng máy tính dần dần được mở rộng không chỉ là các nhà tin học chuyên nghiệp mà cả những người không chuyên. Nhất là với việc phát triển nhanh chóng của máy vi tính, sự phổ cập rộng khắp của vi tin học, đối tượng sử dụng không chuyên tin học ngày càng nhiều. Kỹ thuật sử dụng menu không còn đáp ứng được yêu cầu phát triển các hệ phần mềm lớn: hệ thống các menu chồng chất làm cho người sử dụng lúng túng và không bao quát, không nắm được toàn bộ vấn đề.

Vấn đề xây dựng giao tiếp thân thiện người - máy được đặt ra một cách rõ ràng và dứt khoát cho mọi hệ thống phần mềm lớn: sự thành công hay thất bại của hệ phần mềm, việc nó được người sử dụng cuối cùng chấp nhận hay không, tùy thuộc chủ yếu vào việc có dễ dàng sử dụng được phần mềm đó không. Rõ ràng là cần phải đầu tư nhiều suy nghĩ vào việc vạch kế hoạch xây dựng giao tiếp thân thiện hơn làvào bản thân phần mềm ứng dụng, không thể như trước đây chỉ coi việc thực hiện xong phần ứng dụng là được.

2. Các thành phần của giao tiếp người - máy

Giao tiếp người - máy được xem là bộ đệm giữa chương trình ứng dụng và người sử dụng, cho phép người sử dụng chọn dễ dàng công việc mình định làm. Giao tiếp này chứa tương tác giữa người và máy bao gồm các công việc: đối thoại liên tục người - máy, người ra lệnh, cung cấp các thông tin cần thiết còn máy thực hiện mệnh lệnh và đưa lại các kết quả xử lý cho con người. Giao tiếp phải phản ánh được cách suy nghĩ, quan niệm của người sử dụng đối với vấn đề đang được xử lí, các con đường truy nhập vào thông tin. Điều này đặc biệt quan trọng đối với các hệ thống văn phòng vì trong các văn phòng người lãnh đạo không quan tâm tới việc học cách sử dụng hệ thống mà chỉ muốn thu được ngay các thông tin mình cần. Do vậy, giao tiếp người - máy cần phải được xem xét từ hai khía cạnh khác nhau: góc độ quan niệm của người sử dụng và góc độ hệ thống của người lập trình.

a. Giao tiếp người -máy: khía cạnh quan niệm

Cho tới rất gần đây, đại đa số các bộ chương trình đều chỉ có phần giao tiếp không gì nhiều hơn là một bộ thông dịch dòng lệnh. Chương trình cho hiện một kí hiệu báo rằng nó sẵn sàng nhận lệnh của người sử dụng. Người sử dụng trả lời bằng cách gõ vào các thông tin văn bản. Mặc dầu kỹ thuật này cho phép thực hiện việc liên lạc mức thấp giữa người sử dụng và máy tính, nó còn mang nhiều điểm bất tiện như việc người sử dụng phải nhớ hoặc tra cứu về khuôn dạng lệnh và hiệu quả của lệnh.

Giao tiếp thân thiện người - mày thường sử dụng các hình ảnh đồ họa để mô phỏng các đối tượng thực tế. Người sử dụng sẽ dùng một đối tượng đặc biệt, được gọi là con chạy, để thao tác trên các đối tượng khác. Bằng cách điều khiển con chạy, người sử dụng sẽ lựa chọn và kiểm soát các ứng dụng và các đối tượng đi kèm nó. Các hình ảnh trực quan cùng các thông tin văn bản cho phép người sử dụng hình dung dễ dàng mình đang làm gì và sẽ thu được gì. Tất cả đều được thực hiện theo nguyên lí "wyswyg" -

"What you see what you get" (mọi điều bạn thấy được bạn đều có được).

Con chạy là một đối tượng được sử dụng để tập trung sự chú ý của người sử dụng trong việc lựa chọn các công việc cần thực hiện trên màn hình. Con chạy có thể lấy nhiều hình dạng khác nhau, từ dạng mũi tên, chữ thập cho tới đồng hồ, bàn tay .. tùy theo từng loại ứng dụng, tùy theo chức năng đang thực hiện. Người sử dụng có thể điều khiển con chạy bằng các phím mũi tên trên bàn phím hoặc thông qua cơ chế con chuột, bút quang học...

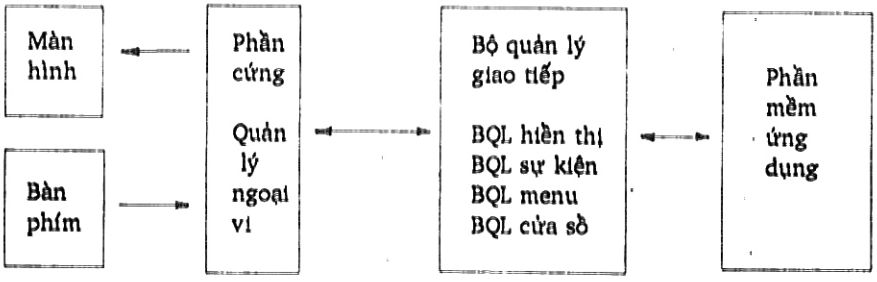
Toàn bộ màn hình được sử dụng để mô phỏng cho bàn giấy làm việc với hình ảnh các công cụ văn phòng : cặp hồ sơ, tài liệu, đồng hồ, lịch, bồ rác...Các hình ảnh dữ liệu này được gọi là hoa văn (icon) biểu diễn cho các chức năng và sự chọn lựa. Người sử dụng dùng con chạy để lựa chọn một hoa văn, khi đó hoa văn sẽ tự đổi màu báo hiệu việc chọn lựa. Có thể dời chỗ hoặc sao chép các hoa văn một cách trực tiếp. Việc mở một hoa văn tạo ra sự xuất hiện trên bàn giấy ứng dụng do hoa văn tượng trưng.

Các ứng dụng thông thường được thực hiện trong một đối tượng đồ họa đặc biệt được coi là cửa sổ. Cửa sổ là biểu tượng cho một tờ giấy trắng đặt trên bàn giấy. Có thể có nhiều cửa sổ chồng chất lên nhau như thể các tờ giấy đè lên nhau, cũng có nghĩa nhiều ứng dụng đang trong quá trình thực hiện. Luôn luôn có một cửa sổ đang hoạt động còn các cửa sổ khác thì không hoạt động. Người sử dụng có thể tùy ý kích hoạt lại các cửa sổ khác nhau cũng như thay đổi kích thước của chúng, di chuyển chúng tới các vị trí khác nhau. Bên trong các cửa sổ là dữ liệu của từng ứng dụng cụ thể.

Việc kiểm soát các ứng dụng và giao tiếp với người sử dụng được thực hiện thông qua cơ chế menu. Các menu bao gồm một số mục có quan hệ logic với nhau, thực hiện một loạt các nhiệm vụ cụ thể có trong hệ thống giao tiếp với người sử dụng. Để chọn các nhiệm vụ này, chỉ cần chọn các mục của menu thông qua việc sử dụng con chạy. Có ba loại menu cơ bản: menu tĩnh, menu ma và menu chạy. Menu tĩnh xuất hiện cố định trên màn hình. Menu ma chỉ xuất hiện khi con chạy đi đến chỗ nó và chương trình đợi người sử dụng chọn các mục tương ứng. Menu sẽ biến mất khi thực hiện xong nhiệm vụ. Việc tổ hợp hai kiểu menu trên sẽ cho một kiểu lai gọi là menu chạy. Kiểu menu này liên kết tính nhìn được của menu tĩnh với hiệu quả hiển thị của menu ma.

b. Giao tiếp người - máy: khía cạnh hệ thống

Giao tiếp người - máy dưới khía cạnh hệ thống được thể hiện là một nhóm các bộ quản lý chức năng. Mỗi bộ quản lý như vậy lại bao gồm một loạt các hàm kiểm soát các khía cạnh khác nhau của sự tương tác giữa người sử dụng và chương trình ứng dụng. Đó là các bộ quản lý hiển thị, quản lí sự kiện, quản lý menu và quản lý cửa sổ (xem hình sau)



Bộ quản lý hiển thị thực hiện các thao tác đặc biệt về việc hiển thị bàn giấy và con chạy. Bàn giấy chiếm toàn bộ màn hình và trên đó có các hoa văn tượng trưng cho các ứng dụng, các tài liệu và các menu tĩnh. Có một số menu chính dành cho việc kết thúc làm việc, đóng mọi cửa sổ đã hiện và in màn hình. Con chạy có thể thay đổi hình dạng và trạng thái dưới sự kiểm soát của bộ quản lý này.

Bộ quản lý sự kiện là một phần của chương trình nhận mọi tác động của người sử dụng, thông dịch và chuyển chúng cho phần còn lại của hệ thống. Nhờ có bộ quản lý này, mọi sự kiện đặc thù cho phần

cứng tỉ như di chuyển con chạy, nhấn phím... đều được phát hiện. Việc gõ phím được lọc qua một bảng tra cứu đơn giản để định nghĩa lại giá trị và chuyển giao cho chương trình xử lí. Mỗi khi phát hiện ra một sự kiện, bộ quản lí sự kiện sẽ nạp dữ liệu tương ứng vào cấu trúc dữ liệu chứa các thông tin về kiểu sự kiện, vị trí con chạy và phím vừa được gõ. Tiếp đó, con trỏ về cấu trúc này sẽ được gửi trả lại hàm đã gọi tới bộ quản lý sự kiện.

Bộ quản lý menu cũng có hàm cho phép tạo ra và kích hoạt menu cũng như các hành động cất giữ hoặc sử dụng lại bộ nhớ. Menu ma được hình thành qua lời gọi hàm quản lý menu với các đối: số mục trong menu và con trỏ tới bảng các con trỏ dòng văn bản. Nếu một mục được người sử dụng lựa chọn thì hàm sẽ cho lại chỉ số tương ứng trong bảng.

Bộ quản lý cửa sổ là phức tạp nhất trong tất cả các bộ quản lý trên. Nó kiểm soát việc tạo ra, hiển thị, thay đổi và xóa các cửa sổ trên bàn giấy. Bộ quản lý cửa sổ cũng khôi phục các vùng đặc biệt trên màn hình, kể cả mọi cửa sổ và các phần tử trên bàn giấy. Bộ quản lý này đưa ra nhiều thao tác dữ liệu trong từng cửa sổ, phát hiện sự kiện đang xảy ra trong cửa sổ nào...

III - TÌM KIẾM VĂN BẢN THEO NỘI DUNG

Việc tìm kiếm văn bản đã được đặt ra từ lâu và cho đến nay đã có nhiều phương pháp, thủ tục, thuật toán được áp dụng trong thực tế. Có thể phân loại tạm thời các phương pháp tìm kiếm văn bản thành bốn nhóm sau. Ba nhóm đầu đã được nghiên cứu nhiều còn nhóm thứ tư mới được đề cập tới trong lĩnh vực nhận dạng và phân tích dữ liệu, dầu vậy nó cũng đem lại nhiều kết quả ứng dụng khả quan. Chúng ta sẽ xem xét chi tiết các nhóm này cùng những ưu nhược điểm của chúng.

1. Các phương pháp duyệt toàn bộ văn bản

Cách tìm kiếm văn bản thông thường và đơn giản nhất, không tốn công tiền xử lí trước, là duyệt toàn bộ các văn bản để phát hiện sự xuất hiện của một dãy kí tự mẫu nào đó. Việc xác định sự xuất hiện mẫu được thực hiện bằng cách so sánh lần lượt các kí tự của dãy mẫu và các kí tự trong văn bản bắt đầu từ kí tự đầu tiên. Nếu gặp sự khác biệt thì lặp lại việc so sánh nhưng từ kí tự tiếp kí tự lấy làm gốc trước đó. Mặc dầu đơn giản trong cài đặt, thuật toán này quá chậm: nếu m là chiều dài của dãy kí tự mẫu, n là chiều dài của văn bản thì sẽ cần khoảng 0 (m*n) phép so sánh.

Thuật toán của Knuth, Morris và Pratt cho việc xác định sự xuất hiện mẫu trong một dãy kí tự chỉ cần tới 0 (m+n) phép so sánh. Tư tưởng chính ở đây là chuyển dịch dãy cần so sánh đi nhiều hơn 1 kí tự về bên phải mỗi khi gặp sự khác biệt. Việc kiểm tra sự xuất hiện mẫu được đồng thời từ điểm đầu và điểm cuối của mẫu (xem [3]).

Nói chung ưu điểm của phương pháp duyệt toàn bộ là đơn giản, không đòi hỏi không gian nhớ phụ, không đòi hỏi phải chuẩn bị gì trước cho việc tìm kiếm. Hơn nữa, vì văn bản được duyệt trực tiếp nên mọi thông tin có sẵn trong văn bản đều có thể được sử dụng làm tiêu chuẩn tìm kiếm, do vậy cho phép đưa vào những câu hỏi tìm kiếm phức tạp (chẳng hạn như các nửa từ hoặc một đoạn các từ v.v..). Nhược điểm hiển nhiên của phương pháp này là thực hiện rất nhiều các thao tác xử lí và truy nhập đĩa khi cơ sở dữ liệu là lớn, do vậy thời gian trả lời cho người sử dụng cũng rất lâu. Vì những nhược điểm như vậy mà phương pháp này chỉ được sử dụng cho các phần cứng đặc biệt hoặc được dùng kèm với các phương pháp truy nhập thông tin khác để thu hẹp không gian duyệt xét. .

2. Các phương pháp tệp ngược

Ngược lại với phương pháp duyệt toàn bộ, phương pháp tệp ngược cố gắng thu hẹp khối lượng văn bản cần phải duyệt tìm vào chỉ các đối tượng có liên quan. Mọi văn bản đều được biểu diễn thành một

danh sách các từ (khóa), coi như các từ đó diễn tả được cho nội dung của văn bản cần tìm kiếm. Có thể thực hiện được việc tìm kiếm nhanh nếu tạo lập được các bảng chỉ số cho các từ khóa đó. Với các từ để hỏi được đưa vào, chỉ cần tìm đúng từ khóa là có thể rút ra được tất cả các chỉ số cho các văn bản chứa từ đó. Có nhiều cách tổ chức bảng chỉ số này: tệp được sắp (các từ khóa) B-cây, TRIE, băm, hoặc các biến thể và tổ hợp của những cách này (xem [3]).

Ưu điểm của phương pháp này là tương đối dễ cài đặt, nhanh và cũng thuận tiện cho việc xử lí các từ đồng nghĩa. Do những lý do này mà phương pháp tệp ngược được sử dụng nhiều trong các hệ thư viện.

Nhược điểm chủ yếu của phương pháp này là không gian nhớ phụ dùng cho các tệp chỉ số khá lớn (từ 50 - 300% kích thước tệp gốc). Phải thường xuyên tiến hành bảo trì, cập nhật và tổ chức lại các tệp chỉ số nếu môi trường có nhiều thay đổi. Thời gian xử lí tăng lên nhanh chóng khi biểu thức hỏi trở nên phức tạp vì phải lấy giao các danh sách chỉ số. Cuối cùng là câu hỏi bị giới hạn bởi vốn từ được dùng trong việc đánh chỉ số. Không thể xử lí trực tiếp được việc hỏi theo nửa từ hoặc đoạn từ. Những nhược điểm này cũng làm giới hạn đáng kể các ứng dụng của tệp ngược trong các hệ tìm kiếm văn bản.

3. Phương pháp mã trùm và tệp dấu hiệu (superimposed coding and signature file)

Các phương pháp dựa trên mã trùm tỏ ra thích hợp hơn cả đối với việc tìm kiếm văn bản. Về mặt lịch sử, người đầu tiên đã sử dụng mã trùm trong việc tìm kiếm là C. N. Mooers năm 1947. Ông đã phát minh ra thiết bị cơ khí để thực hiện việc tìm kiếm. Về sau nhiều người đã tiếp tục ý kiến của ông. Ngoài ra, cách tổ chức tệp dấu hiệu có thể được coi là một thỏa hiệp giữa hai phương pháp duyệt toàn bộ và tệp ngược đã trình bày ở trên.

Trong phương pháp mã trùm, mỗi từ của văn bản đã cho đều được băm để tạo ra một số vị trí bit mang giá trị 1 trong mẫu bit với chiều dài cố định. Các mẫu bit được đặt trùm lên nhau (lấy theo phép toán OR cho các mẫu bit của tất cả các từ trong văn bản), kết quả ta thu được một mẫu bit là dấu hiệu của văn bản. Có thể dùng dấu hiệu này để xác định vị trí văn bản (như kiểu băm khóa chính), hoặc có thể lưu trữ tất cả các dấu hiệu của văn bản vào một tệp riêng, dùng làm bộ lọc cho việc tìm kiếm.

Khi thực hiện tìm kiếm, mẫu cũng được xử lý tương tự như các văn bản gốc để tạo nên dấu hiệu. Việc tìm kiếm được thực hiện theo hai giai đoạn. Ban đầu, duyệt qua toàn bộ tệp dấu hiệu để lọc ra các văn bản có dấu hiệu có chứa dấu hiệu của mẫu. Sau khi đã loại được phần lớn các văn bản không thỏa dấu hiệu, còn phải thực hiện duyệt toàn bộ các văn bản nào có chứa đựng mẫu đã đưa vào.

Ưu điểm của phương pháp này là đơn giản trong cài đặt. Thời gian xử lí được rút gọn đáng kể do chỗ chỉ phải duyệt tìm tệp dấu hiệu và một phần nhỏ cơ sở dữ liệu. Kích thước của tệp dấu hiệu thông thường chỉ bằng 10 - 20% kích thước của cả cơ sở dữ liệu. Hơn nữa, với phương pháp này có thể xử lí được các câu hỏi phức tạp trên nửa từ hoặc cả đoạn từ, có thể cho phép tha thứ cho các lỗi chính tả hoặc gõ nhầm kí tự. Và một điểm nữa không thể bỏ qua là đối với các hệ thống thường xuyên thực hiện bổ sung dữ liệu như văn phòng, phương pháp này không cần đòi hỏi có người quản trị để thực hiện bảo trì, tổ chức dữ liệu. Nhược điểm có thể nói đến của phương pháp này là thời gian trả lời câu hỏi sẽ lâu nếu cơ sở dữ liệu là rất lớn.

4. Phương pháp phân chùm

Phương pháp này dựa trên ý chính là gộp nhóm các văn bản tương tự nhau thành từng chùm: các văn bản có liên quan chặt chẽ với nhau có khuynh hướng đáp ứng được cho cùng một yêu cầu. Nhiều công trình về nhận dạng cũng chú trọng tới vấn đề này.

Phương pháp phân chùm bao gồm hai thủ tục: sinh ra chùm và tìm kiếm chùm. Các văn bản sẽ được

xử lí và từ đó các từ quan trọng được tự động rút ra. Người ta sử dụng các bộ tự điển phủ định để khử bỏ các từ chung kiểu như "và", "thì",... Các bộ từ điển đồng nghĩa được sử dụng để chuyển các từ về lớp các khái niệm. Một số thủ thuật biến đổi các từ được sử dụng để đưa chúng về dạng cố định. Mỗi văn bản được chuyển thành một vectơ t chiều với t là số các khái niệm đã ấn định. Việc xuất hiện một từ được đánh dấu là 1 hoặc một số dương (trọng số của từ), phản ánh tầm quan trọng của từ trong văn bản; còn việc thiếu từ đó được đánh giá là 0. Sau đó sẽ thực hiện phân hoạch các vectơ này thành các nhóm có quan hệ gần gũi với nhau. Việc tìm kiếm có phần đơn giản hơn việc sinh chùm. Câu hỏi đưa vào cũng được biểu diễn dưới dạng vectơ t chiều và được so sánh với các tâm chùm để xác định ra các chùm gần nhất. Từ đó có thể thông qua đối thoại với người sử dụng để đi dần tới câu trả lời.

IV - XÂY DỰNG HỆ THỐNG TIN VĂN PHÒNG OFFICIN

Trong thời gian vừa qua chúng tôi đã phát triển thử nghiệm một hệ thông tin văn phòng với tên gọi OFFICIN. Hệ thống này được thiết kế dựa trên cơ sở phân tích chi tiết các chức năng và nhiệm vụ của văn phòng trong quan hệ với cả tổ chức và người lãnh đạo ([9]). Hệ thống nhằm đáp ứng cho nhiều văn phòng xử lí thông tin, tạo ra công cụ để lưu trữ kết quả làm việc của các nhân viên văn phòng và cung cấp phương tiện cho lãnh đạo sử dụng trực tiếp các thông tin đã được chuẩn bị đó. Sau đây chúng tôi xin giới thiệu một số vấn đề chính đã được thực hiện trong hệ thống này và các hướng nghiên cứu có thể triển khai tiếp.

1. Một số vấn đề cài đặt hệ thông tin văn phòng

Trong việc cài đặt hệ thống phần mềm, điều rất quan trọng là xác định rõ tư tưởng quán xuyến của hệ thống. Cả hệ thống phần mềm cần được xây dựng dựa trên việc thể hiện rõ tư tưởng đó. Mặt khác, hệ thống cần phải đi sát với các đối tượng, con người mà nó phục vụ, phải thể hiện được quan niệm cũng như cách nhìn nhận, xử lý của con người. Do vậy giai đoạn phân tích hệ thống, xác định rõ lớp người sử dụng, xác định các cách xem xét và xử lý dữ liệu là rất quan trọng và có thể nói sẽ quyết định đến chất lượng của toàn bộ hệ thống, cần phải được đầu tư khá nhiều thời gian và công sức.

Trên cơ sở các phân tích chi tiết như vậy, chúng tôi đã hình thành kế hoạch phát triển giao tiếp thân thiện người - máy, thiết lập mô hình truy nhập thông tin qua cấu trúc các cặp hồ sơ lưu trữ. Người sử dụng sẽ được giới thiệu về các cặp hồ sơ tài liệu mà hệ thống quản lý. Tùy theo vai trò của họ trong cơ quan, là nhân viên hoặc cán bộ lãnh đạo, sẽ có các con đường truy nhập thông tin riêng để đáp ứng yêu cầu và nhiệm vụ xử lý thông tin của họ.

Chúng tôi đã sử dụng một số kỹ thuật quản lý màn hình ở chế độ chữ số để xây dựng giao tiếp thân thiện người - máy. Do chưa có nhiều kinh nghiệm trong việc sử dụng chế độ đồ họa nên chúng tôi tạm gác vấn đề sử dụng các đối tượng đồ họa trong việc xây dựng giao tiếp cho các lần cải tiến sau. Đã thực hiện việc cài đặt các bộ quản lý cửa sổ, hiển thị và menu đồng thời đưa vào một số yếu tố hoạt hình và âm nhạc để làm sinh động hệ thống.

Chúng tôi đã tiến hành nghiên cứu và cài đặt thành công kỹ thuật tìm kiếm văn bản theo nội dung, dùng phương pháp mã trùm và tệp dấu hiệu.

Một số kỹ thuật xử lý danh sách đã được áp dụng trong việc tìm kiếm thông tin để thể hiện mối quan hệ hữu cơ giữa các đối tượng khác nhau trong hệ thống. Một số kỹ thuật tạo lập con trỏ theo ngày tháng đã được sử dụng để nâng tốc độ xử lí thông tin và giảm việc truy nhập đĩa.

Cũng cần nói thêm rằng việc đưa hệ thống vào làm việc thực sự là cả một quá trình lâu dài và kiên nhẫn. Vấn đề chủ yếu ở đây không còn là tin học nữa mà đã bước sang lĩnh vực của tổ chức và con người. Hệ thông tin văn phòng không phải là một hệ thống hoàn toàn tự động để thay thế con người mà trái

lại, nó chỉ là một công cụ không hơn không kém để giúp cho lao động xử lý thông tin của con người có chất lượng hơn. Hệ thống chỉ thực sự hoạt động có chất lượng chừng nào có được sự ủng hộ mạnh mẽ của cả lãnh đạo lẫn các nhân viên thực hiện, chừng nào hình thành được một cơ cấu tổ chức xử lý thông tin mới, năng động và hữu hiệu. Do vậy, ngoài yếu tố tin học, còn cần chuẩn bị nhiều yếu tố khác như tổ chức, con người, tâm lí, thời gian... để đưa hệ thống vào vận hành.

2. Các hướng phát triển tiếp theo của hệ OFFICIN

Để nâng cao khả năng sử dụng hệ thống trong nhiều môi trường khác nhau, việc nghiên cứu xây dựng giao tiếp thân thiện người - máy trong chế độ đồ họa là rất quan trọng. Vấn đề này cũng đòi hỏi nhiều kiến thức mới và cao cấp về các kỹ thuật lập trình hiện đại.

Các yêu cầu về việc thực hiện tự động tóm tắt và phân loại văn bản là những yêu cầu thực tế của công tác văn phòng. Ở đây đòi hỏi nhiều kiến thức về nhận dạng và trí tuệ nhân tạo, nhất là trong vấn đề xử lý tri thức có trong mọi văn bản.

Việc lưu trữ các thông tin hình ảnh và tiếng nói là vấn đề tự nhiên ở mọi văn phòng. Các kỹ thuật nén thông tin để thuận tiện cho lưu trữ và tìm kiếm là rất quan trọng và chúng ta còn ít có kinh nghiệm xử lý.

Việc hình thành cơ sở dữ liệu chung cho nhiều người sử dụng là yêu cầu tất yếu của các hệ thống văn phòng. Vấn đề bảo vệ dữ liệu trong quá trình sử dụng chung là cần thiết và phải được thực hiện.

Vấn đề đưa vào sử dụng các mạng máy tính và việc triển khai các hệ thống thư tín điện tử sẽ là những yêu cầu cần có nhiều nghiên cứu từ bây giờ và trong tương lai.

Nhận ngày 15-8-1988

TÀI LIỆU THAM KHẢO

1. Y. Gauthier, Macintosh, Une philosophia bureautique", Soft and Micro, October 87, pp. 130-133.

2. C. Faloutsos, S. Christodoulakis, "Access Methods for Documents", Office Automation, Springer-Verlag, 1985, pp. 317-338.

3. E. Horowitz, S. Sahni, "Fundamentals of Data Structures", PITMAN, 1981; pp. 190-197, 485-555.

4. A. Lee, F. H. Lochovsky, "User Interface Design", Office Automation, Springer-Verlag, 1985, pp. 3-20.

5. D. L. Lee, F. H. Lochovsky, "Text Retrieval Machines", Office Automation, Springer-Verlag, 1985, pp. 339-375.

6. Najah Naffah, "La vvraie convivialité reste à inventer", InforPC, No. 37, Dec. 87/Janv. 88, pp. 84-89.

7. Ron Person, "Aspects avancées du langage C", InterEditions 1987, pp. 155-187 .

8. C. C. Woo, F. H. lochovsky, A. Lee, "Document Management Systems", Office Automation, Springer-Verlag, 1985, pp. 21-40.

9. Ngô Trung Việt, "Tự động hóa công tác văn phòng", 1988, chưa xuất bản.

ABSTRACT

ABOUT THE OFFICE INFORMATION SYSTEMS

Some characteristics of office information systems are briefly presented in this paper. The user friendly interface problems and methods for text retrieval are discussed in details. Some experiments in implementation of these systems are given.