Tạp chí Tin học và Điều khiển học T. 11, S. 3 (1995) (41 – 47)

THIẾT KẾ VÀ CÀI ĐẶT HỆ NÉN DỮ LIỆU THÔNG MINH IDCS

NGUYỄN THANH THỦY

Đại học Bách khoa Hà nội

Summary. In this paper we shall investigate some features of the system IDCS (Intelligent Data Compressing System) which, based on the nature and characteristics of a data file f and using a library of compressing routine L, could choose the most convenient compressing program $C \in L$, by infering over a knowledge base K in the form of rules. It is interesting that the more complete the knowledge base K, the more effective the chosen compressing C.

I. MỞ ĐẦU

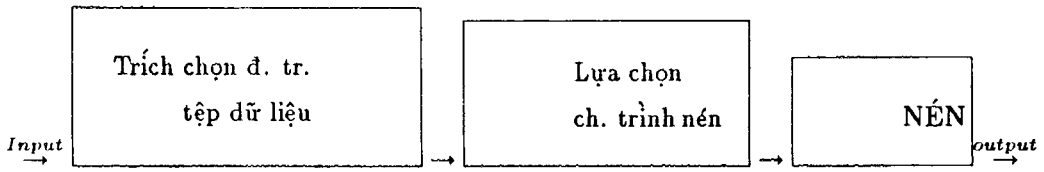
Nén dữ liệu đặc biệt được lưu tâm trong các lĩnh vực có liên quan như lưu trữ và tìm kiếm trong cơ sở dữ liệu; truyền dữ liệu; viễn thông; mật mã và bảo vệ thông tin. Số các sản phẩm nén dữ liệu trên thị trường ngày càng nhiều. Tuy nhiên, hiệu quả nén của chúng hoàn toàn độc lập với bản chất dữ liệu của dòng dữ liệu vào; nó chỉ phụ thuộc vào mô hình và phương pháp mã hoá đã chọn bởi nhà thiết kế. Trong hệ IDCS, với một thư viện các chương trình nén L, tùy thuộc vào bản chất và đặc trưng của tập dữ liệu vào f, hệ sẽ lựa chọn và kích hoạt chương trình nén phù hợp $C \in L$, nhờ vào kết quả suy diễn trên một cơ sở trí thức được nạp vào hệ thống từ trước.

Các kỹ thuật nén dữ liệu nhằm chủ yếu vào việc mã hoá dữ liệu dưới dạng phù hợp sao cho nó có thể dễ dàng khôi phục được dữ liệu ban đầu và tiết kiệm dung lượng nhớ, đồng thời làm giảm thời gian truyền dữ liệu trên mạng, bảo mật thông tin. Người ta chia làm hai lớp kỹ thuật nén: nén logic và nén vật lý. Các kỹ thuật nén dữ liệu có thể làm mất mát hay không mất mát thông tin trong dòng dữ liệu vào. Quá trình nén dữ liệu chia làm hai phần: Mô hình và mã hoá dữ liệu. Trên thực tế, thông thường hai mô hình: Thống kê và Từ điển hay được sử dụng. Tuy vậy, mô hình dự báo được coi là một hướng đi có nhiều triển vọng trong nén hình ảnh và tiếng nói. Sau khi xác định được mô hình, phương pháp mã hoá sẽ được lựa chọn sao cho số bit được dùng trên thực tế để lưu trữ xấp xỉ với số bit được tính toán lý thuyết dựa theo mô hình. Hai phương pháp mã hoá điển hình hay được trích dẫn là: phương pháp Sannon-Fano và phương pháp Huffman. Nhược điểm của hai phương pháp này là số bit được sử dụng để lưu trữ các kí hiệu là số nguyên, tuy rằng về lý thuyết không phải như vậy. Phương pháp mã hoá

số học nhằm khắc phục nhược điểm này, bằng cách tiến hành mã hoá cho từng mẩu tin (message) thay cho từng kí hiệu riêng biệt. Do vậy, tính trung bình các kí hiệu được mã hoá nhờ sử dụng một số lẻ các bit. Các kỹ thuật nén hay được sử dụng trong thực tế bao gồm: kỹ thuật thay thế, kỹ thuật loại bỏ ký tự trống, kỹ thuật ánh xạ bit, kỹ thuật mã loại dài, kỹ thuật sử dụng nửa bit, kỹ thuật mã kép, kỹ thuật mã tương đối, kỹ thuật nén topô, kỹ thuật *LZ**, *LZW*, *LZSS*... Người ta cũng chứng minh được rằng phương pháp Huffman là phương pháp tối ưu theo entropy.

Số các sản phẩm nén dữ liệu trên thị trường ngày càng nhiều: COMPACT, COMPRESS (trong môi trường UNIX); SQ, ARS, PKZIP, ARJ, LHA, ICE, LZH, chương trình Norton (trong môi trường DOS). Tuy nhiên chúng đều có chung một đặc điểm là không tính đến các đặc điểm và bản chất của tệp dữ liệu đầu vào. Thông thường, người sử dụng đầu cuối dùng chung một chương trình nén cho các loại tệp dữ liệu có bản chất khác nhau.

Ngược lại, trong hệ IDCS, với một thư viện các chương trình nén L, tùy thuộc vào bản chất và đặc trưng của tệp dữ liệu vào f và dựa vào kết quả suy diễn trên một cơ sở trí thức luật, hệ sẽ lựa chọn và kích hoạt chương trình nén C, $C \in L$, phù hợp nhất đối với f.



II. CẤU TRÚC HỆ NÉN DỮ LIỆU THÔNG MINH
IDCS(Intelligent Data Compressing System)

Thông thường, mỗi khi sử dụng một chương trình để nén một tập dữ liệu nào đó, hiệu quả nén hoàn toàn độc lập với bản chất và đặc trưng dữ liệu của tệp dữ liệu; nó chỉ phụ thuộc vào mô hình và phương pháp mã hoá đã được lựa chọn lúc cài đặt.

Quá trình nén trong hệ IDCS được tiến hành theo các giai đoạn sau:

Giai đoạn 1: Tiền xử lí tệp dữ liệu.

Tiền xử lý tệp dữ liệu nhằm trích chọn ra đặc trưng cơ bản của tệp dữ liệu (kiểu loại dữ liệu, phân bố xắc suất các nhóm kí hiệu...)

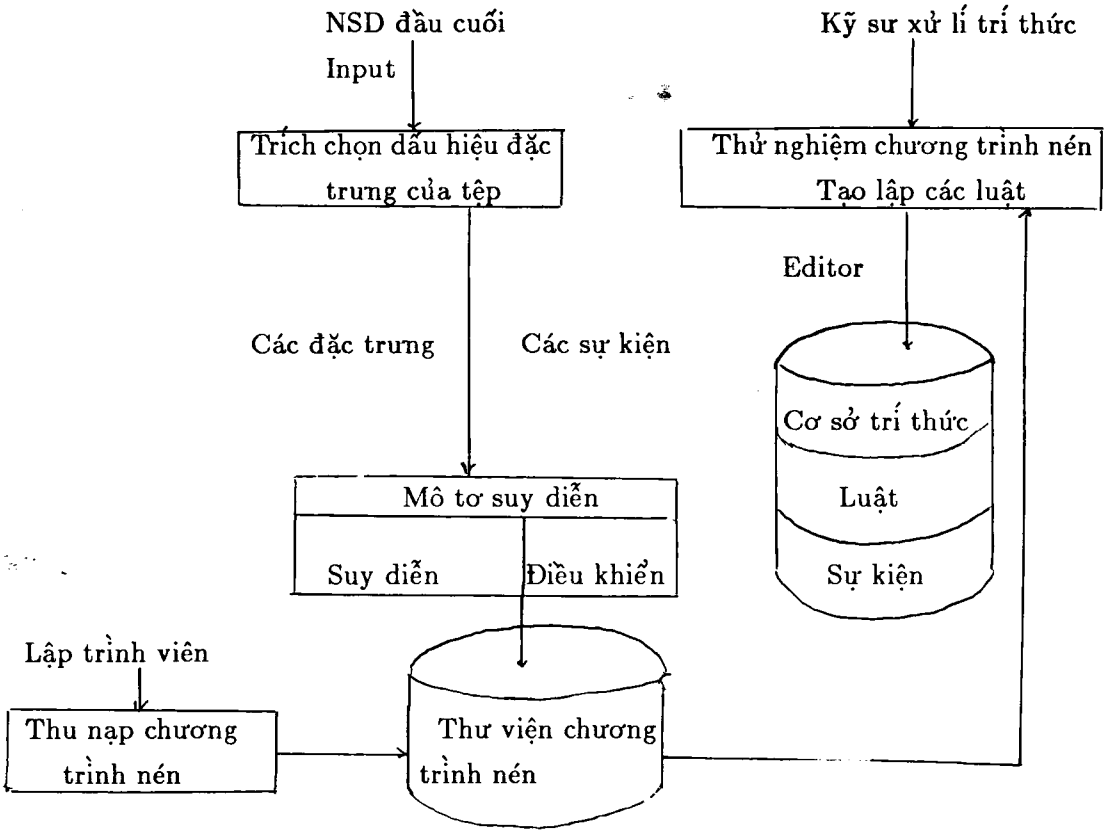
Giai đoạn 2: Lựa chọn chương trình nén phù hợp.

Đây là phần thông minh của hệ thống. Nó đảm nhiệm chức năng mô tơ suy diễn. Quá trình suy diễn được thực hiện với các thông tin vào (sự kiện) do giai đoạn 1 cung cấp và các thông số do người sử dụng đưa vào, trên cơ sở áp dụng các luật được nạp vào hệ thống từ trước. Ở đầu ra, mô tơ suy diễn sẽ xác định các phương pháp nén chấp nhận được theo yêu cầu của người sử dụng và hệ thống.

Giai đoạn 3: Nén dữ liệu.

Đây là phần hành động của hệ thống. Tệp dữ liệu sẽ được nén và giải nén theo chương trình đã được lựa chọn.

Toàn bộ cấu trúc của hệ được cho trên hình vẽ sau:



III. TRÍCH CHỌN CÁC DẤU HIỆU ĐẶC TRƯNG CỦA TỆP DỮ LIỆU

Phần tiền xử lí tệp dữ liệu gồm 2 giai đoạn chính:

+ Chia cắt tệp dữ liệu, tạo ra các vùng dữ liệu thuần nhất.

+ Đối với các vùng dữ liệu thuần nhất, tiến hành các thao tác phân tích dữ liệu để trích xuất các đặc trưng thống kê.

Các dấu hiệu đặc trưng của tệp dữ liệu chia làm hai loại:

+ Các đặc trưng chung: entropy, kiểu loại dữ liệu.

+ Các đặc trưng riêng với tệp dữ liệu. Cần chú ý một điểm là đối với các tệp dữ liệu có bản chất khác nhau, các đặc trưng có thể khác nhau. Chẳng hạn, đối với tệp văn bản đó là tần xuất xuất hiện các tổ hợp kí hiệu (chữ, cụm chữ, từ, cụm từ, đoạn câu, câu...). Đối với tệp ảnh, người ta quan tâm đến các thông tin ảnh đen, trắng/ màu, ảnh, ảnh điểm, ảnh đường nét, ảnh vùng, ảnh có màu vân lặp, ảnh có thể, dự đoán ...

IV. TẠO LẬP CƠ SỞ TRÍ THỨC

Trong hệ IDCS, sử dụng phương pháp biểu diễn trí thức chuyên gia dưới dạng luật. Mỗi luật tương ứng với một kinh nghiệm của chuyên gia con người trong lĩnh vực nén dữ liệu. Các kinh nghiệp này có được sau khi sử dụng các chương trình nén nhiều lần đối với các thể loại dữ liệu và các tệp dữ liệu mẫu khác nhau. Do vậy, các luật này mang tính chất động, phụ thuộc vào thống kê của chuyên gia trong quá trình sử dụng.

Cách thử nghiệm xây dựng cơ sở trí thức trong hệ IDCS như sau:

Giả sử có thư viện chương trình nén $L = \{C_1, ..., C_n\}$. Với mỗi xâu $u = C_{i_1}, ..., C_{i_k}$ đối với các tệp dữ liệu mẫu f, $f \in F = \{f_1, ..., f_n\}$ nghiã là $C_{i_1}(C_{i_2}(...C_{i_k}(f)...))$.

- Với tệp f, xác định đặc trưng $a_1, ..., a_l$ với các giá trị tương ứng $v_1, ..., v_l$.

- Tính toán các đặc trưng của tổ hợp nén v: entropy trung bình của các kiểu loại dữ liệu trước và sau khi nén, tỉ số nén entropy, hiệu quả nén thời gian nén, chi phí giải nén...

- Tạo lập các luật sản xuất có dạng

Nếu < điều kiện 1 > và

.....

< điều kiện n >

thì < Kết luận >

Ví ụ 1 Nếu Yêu cầu nén = giảm bộ nhớ và

Tỷ số nén entropy của tổ hợp nén u là cực đại

Thì Tập phương pháp nén phù hợp = $\{u\}$

Cơ sở trí thức trong hệ IDCS được chia thành 4 lớp như sau:

+ Lớp các luật chung

+ Lớp các luật hướng mục đích nén dữ liệu

+ Lớp các luật phụ thuộc kiểu loại dữ liệu

+ Lớp các luật phụ thuộc đặc trưng thống kê của tệp dữ liệu

Ví dụ 2 Luật 75 (Thuộc nhóm các luật chung)

Nếu Kiểu tệp dữ liệu vào = ảnh và
yêu cầu nén = giảm bộ nhớ và
đã áp dụng = RL
Thì Ph = Ph {RL, RL - LZW },
RL - Phương pháp nén mã loại dài (Run Length)
LZW - Phương pháp nén Lempel-Ziv-Welch
RL-LZW - là tổ hợp nén RL và LZW.

V. MÔ TƠ SUY DIỄN

Quá trình suy diễn trong hệ IDCS bao gồm cả suy diễn tiến (forward chaining) và suy diễn lùi (backward chaining).

Bài toán suy diễn:

Vào: Tập luật sản xuất $R = \{r_1, ..., r_m\}$ (xem mục 4)

Tập các sự kiện đã biết GT (các đặc trưng của tệp dữ liệu, các thông số do người sử dụng cung cấp)

Ra: Tập các chương trình nén phù hợp Ph.

Yêu cầu: Sử dụng tập luật R, xác định xem có thể suy ra giá trị của Ph, biết rằng ban đầu đã có GT.

Quá trình suy diễn được mô tả như sau: Ban đầu biết GT. Sau đó dựa vào các luật thỏa mãn, dẫn thêm các sự kiện mới, cho đến khi Ph được định nghiã. Nghiã là

i) $TG_1 = GT$

ii) $TG_1 \subseteq TG_2 \subseteq ... \subseteq TG_k$, $Ph \in TG_k$, ở đây với mọi $j = 1,...,k-1$ có luật r : $p_1 \wedge ... \wedge p_n \rightarrow q$ sao cho $p_1, ..., p_n \in TG_j$, và $TG_{j+1} = TG_j \cup \{q\}$.

Ngược lại, ở mỗi bước, suy diễn lùi tìm cách thay việc chứng minh q trong luật $p_1 \wedge ... \wedge p_n \rightarrow q$ bởi việc chứng minh các $p_1, ..., p_n$. Quá trình cứ tiếp tục cho đến khi tất cả các sự kiện cần chứng minh đều có trong GT. Nghiã là:

I) $TG_1 = \{Ph\}$

ii) $TG_1 \Leftarrow TG_2 \Leftarrow ... \Leftarrow TG_k$, $TG_k \subseteq GT$, ở đây đối với mọi $j = 1,...,k-1$, tồn tại luật r : $p_1, ..., p_n \rightarrow q$ sao cho $q \in TG_j$ và $TG_{j+1} = TG_j \setminus \{q\} \cup \{p_1, ..., p_n\}$.

Tập các luật được dùng trong quá trình suy diễn được gọi là vết suy diễn.

VI. TỔ CHỨC THƯ VIỆN CHƯƠNG TRÌNH NÉN

Điểm quan trọng trong hệ IDCS là tận dụng được tất cả các chương trình nén hiện có. Thư viện chương trình nén càng lớn và cơ sở trí thức đầy đủ và hiệu quả nén càng cao. Việc cập nhật thư viện nén hoàn toàn không làm tồi đi chất lượng của hệ IDECS.

Ở đây, chúng ta có thể làm giầu thư viện chương trình nén theo các hướng sau:

- Thu nạp các chương trình nén hiện có trên thị trường
- Thiết kế và thử nghiệm các chương trình nén chuyên dụng hiệu quả
- Thử nghiệm các tổ hợp chương trình nén $u \in L^*$, chẳng hạn $LZH = RL + LZSS +$ Huffman.

VII. CÀI ĐẶT HỆ THỐNG IDCS

Hệ IDCS được cài đặt bằng ngôn ngữ C, Borland $C++3.1$. Hệ IDCS định hướng xử lí các loại tệp: văn bản, ảnh, .EXE, .DBF. Hệ thống có modun nhận dạng tự động kiểu loại tệp, thông qua truy nhập vào cấu trúc đầu tệp. Đối với các tệp văn bản tiếng Việt, hệ nhận biết dựa trên nhận biết các từ của nó (mỗi từ gồm các phần: phụ âm đầu, vần, phụ âm cuối và dấu). Ngoài ra còn có modun Data - Analysis nhằm xác định các đặc trưng thống kê của tệp dữ liệu. Bên cạnh đó, modun compr - Analysis sẽ tiến hành các tính toán thống kê đối với các phương pháp nén. Chẳng hạn tỉ số entropy

$$RE_c = (E_t - E_s)/E_t,$$

ở đây E_t và E_s tương ứng là entropy của tệp dữ liệu trước và sau khi nén nhờ áp dụng một chương trình nén C nào đó.

Cơ sở trí thức của IDCS gồm 75 luật, chia thành 4 lớp như đã nói ở trên. Modun Know-Editor cho phép soạn thảo tệp luật dưới dạng khá tự nhiên, dễ dàng đối với người sử dụng. Mô tơ suy diễn tiến và lùi For-Inference và Back-Inference được cài đặt gọn nhẹ, độc lập hoàn toàn với nội dung của các luật trong cơ sở trí thức. Ở giai đoạn đầu, IDCS version 0.1 đã tiến hành thử nghiệm với thư viện chương trình nén $L = \{PKZIP, ARJ, LHA, ICE, LZH, LZW, LZ77, LZ78,$ norton$\}$, cho kết quả khả quan.

NHỮNG VẤN ĐỀ PHÁT TRIỂN TIẾP

Hệ IDCS có thể pháp triển tiếp theo các hướng sau:

- Hoàn thiện các phần trong hệ thống: phân tích và trích chọn các đặc trưng tệp dữ liệu, thử nghiệm có thống kê các chương trình nén, xem xét các tổ hợp nén, làm mịn cơ sở tri thức.

- Giữ nguyên cấu trúc hệ, xây dựng các hệ con IDCS chuyên dụng, cho một kiểu loại dữ liệu nào đó, chẳng hạn tệp văn bản tiếng Việt, tệp dữ liệu .DBF có sử dụng tiếng Việt, tệp .EXE, tệp ảnh ...

- Tiếp cận một số thiết bị phần cứng chuyên dụng, trong đó đã bao gồm cả quá trình nén dữ liệu, chẳng hạn máy FAX.

REFERENCES

1. G. Held, T.R. Marshall , Data Compression, John Willey & sons Ltd., New York, USA 1990.
2. A.K. Jain, Fundamental of Digital Image Processing, Prentice-Hall, N.J. 1989.
3. X. Marsault, Compression et cryptage eninformatique, Hermes, Paris, France 1992.
4. M. Nelson, La Compression de donneés, M & T Publ., Inc. California 1992.
5. P. Plume, La Compression de donneés, Ed. Eyrolles, Paris 1993.
6. D.A. Waterman, A guid to Expert System, Addision-Wesley Publ. comp. N.J. 1987.
7. S.M. Weiss, C.A. Kulikowski, A pratical guide to designing expert systems, Rowman & Allanhed Publishers, N.J. 1988.

Đại học Bách khoa Hà nội
Khoa Công nghệ Thông tin