BÁO CÁO: SINH VIÊN NGHIÊN CỨU KHOA HỌC 2011-2012

ஜ★ஜ

XÂY DỰNG ỨNG DỤNG TÌM KIẾM ĐỘNG & MỀM DẺO

Report date: 17/05/2012

G.V hướng dẫn: Phan Trung Huy

Nhóm sinh viên:

1. Lương Thế Dũng

2. Ứng Hoàng Nam

3. Nhữ Bảo Vũ

Lớp: Toán Tin 1 K52

Version : 2012.05.14

NỘI DUNG

BỐI CẢNH CỦA BÀI TOÁN

- So khớp mẫu và tìm kiếm mẫu là bài toán đóng vai trò quan trọng trong lĩnh vực tìm kiếm thông tin.
- Thế hệ máy tính hiện đại hiện nay đang hướng tới khả năng xử lý thông tin định tính (bất định, không chính xác, mờ, …)
- Nhu cầu tìm kiếm động trong các công cụ tìm kiếm nâng cao còn phức tạp và cần được nâng cao chất lượng tìm kiếm.
- Các ứng dụng tìm kiếm cần được thiết kế mềm dẻo, động và dễ dàng nâng cấp.

NHIỆM VỤ NGHIÊN CỨU

- Nghiên cứu các thuật toán tìm kiếm mẫu: mẫu chính xác và mẫu xấp xỉ
- Nghiên cứu về otomat mờ
- Đề xuất thuật toán tìm kiếm mẫu xấp xỉ theo tiếp cận otomat mờ nâng cao.
- Đề xuất giải pháp tìm kiếm mẫu biểu thức, chính quy

CÁC DẠNG TÌM KIẾM

- Phân loại các thuật toán tìm kiếm [3] dựa trên các đặc tính của mẫu
  - Tìm kiếm đơn mẫu
  - Tìm kiếm đa mẫu
  - Tìm kiếm biểu thức: logic, chính quy
- Hai hướng tiếp cận chính
  - Chính xác
  - Xấp xỉ
- Báo cáo này tập trung giải quyết vấn đề tìm mẫu đơn, chính xác và xấp xỉ, và đưa ra phương pháp giải quyết vấn đề tìm kiếm biểu thức động.

PHÁT BIỂU BÀI TOÁN

- **Bài toán 1: Tìm kiếm mẫu đơn chính xác**
  - Cho xâu mẫu P độ dài m, và xâu S độ dài n trên cùng một bảng chữ cái A. Tìm tất cả các xuất hiện của xâu P trong S.
- **Bài toán 2: Tìm kiếm mẫu xấp xỉ**
  - Cho xâu S độ dài n và xâu mẫu P độ dài m trên cùng một bảng chữ cái A. Tìm các vị trí trong S khớp với mẫu, cho phép nhiều nhất k lỗi.
- **Bài toán 3: Tìm kiếm mẫu biểu thức động**
  - Cho xâu S độ dài n và tập các xâu mẫu P_i $(i = 1...n)$, các P_i có quan hệ với nhau theo một biểu thức logic hay biểu thức chính quy nào đó. Tìm các xuất hiện của tập xâu mẫu P_i trong S thỏa mãn biểu thức động (logic, chính quy) đã cho.

TÌM KIẾM MẪU CHÍNH XÁC THEO TIẾP CẬN OTOMAT MỜ (1)

- Ý tưởng chung của tiếp cận otomat mờ [5] [6]
  - Tìm kiếm chính xác sự xuất hiện của mẫu P trong xâu đích S, câu trả lời là chắc chắn: có hoặc không.
  - Cách nhìn mờ cho phép kết quả là "độ mờ xuất hiện mẫu" (hay "mức độ xuất hiện" của mẫu P trong S).
  - Quan điểm hệ mờ về sự xuất hiện mẫu đáp ứng được nhu cầu tìm kiếm mẫu xấp xỉ

TÌM KIẾM MẪU CHÍNH XÁC THEO TIẾP CẬN OTOMAT MỜ (2)

- Otomat mờ dạng tổng quát [7] có mỗi trạng thái mờ là một tập con mờ trên tập nền X = {1, 2, …, m} (X hữu hạn). Có hàm thuộc mở rộng là f: X $\rightarrow$ R.
- Định nghĩa: Otomat mờ so mẫu là bộ $\mathcal{A}(P) = (A, Q, q_0, \delta, F)$
  - Trong đó:
    - Bảng chữ vào $A = A_p \cup \{\#\}$
    - Tập trạng thái $Q = \{0, 1, \ldots, m\}$
    - Trạng thái khởi đầu $q_0 = 0$
    - Trạng thái kết thúc $F = m$
    - Hàm chuyển trạng thái $\delta : QxA \rightarrow Q$

TÌM KIẾM MẪU CHÍNH XÁC THEO TIẾP CẬN OTOMAT MỜ (2)

- Định nghĩa: Cho xâu mẫu P độ dài m và xâu đích S độ dài n. Độ mờ xuất hiện P trên S tại vị trí j là giá trị nguyên $\lambda \geq 0$ thỏa mãn:
  - $\lambda = 0$ nếu $S_j \neq P_1$
  - λ là số lớn nhất sao cho $P_1P_2 .. P_n = S_{j-\lambda+1}S_{j-\lambda+2} .. S_j$

TÌM KIẾM MẪU CHÍNH XÁC THEO TIẾP CẬN OTOMAT MỜ (3)

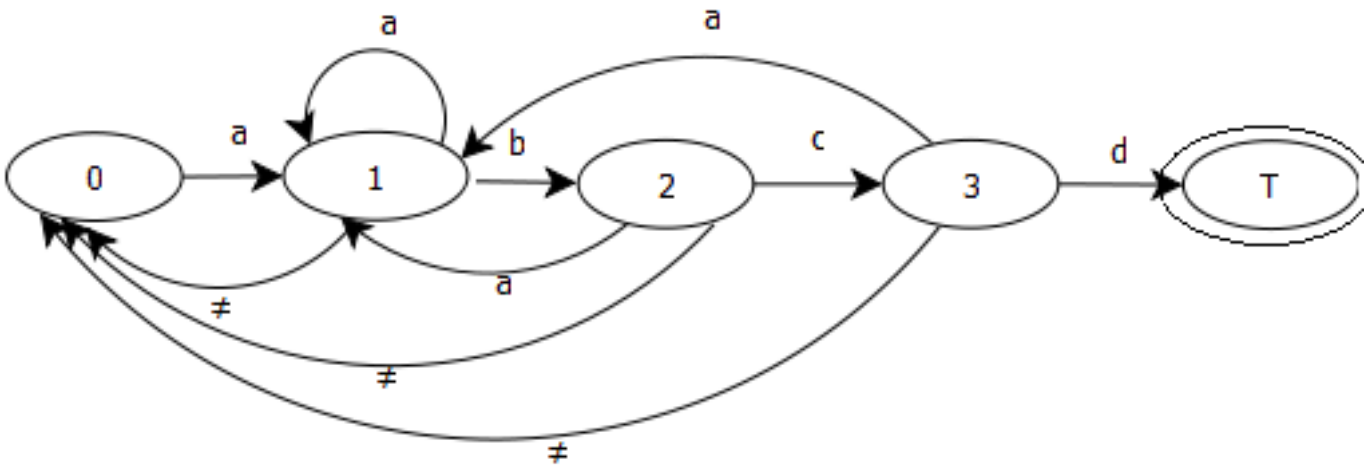
- Bổ đề: Giả sử độ mờ xuất hiện mẫu P tại vị trí S_j là λ, khi đó độ mờ mới λ' tại S_{j+1} được xác định bởi một hàm $\lambda' = TFuzz(\lambda, S_{j+1})$, với *TFuzz* được xác định như sau:

$$TFuzz(0,x)=\begin{cases}0, \forall x \neq P_1\\ 1, x = P_1\end{cases}$$
$$TFuzz(k,x) = 0, \ \ \forall k = 0..m \qquad (1)$$
$$TFuzz(k,x) = \begin{cases}k+1, \ if \ x = P_{k+1}\\ l, \ l \leq k\end{cases}$$

- Với *l* là số lớn nhất sao cho $P_1P_2 .. P_{k-1} = P_{k-l+2}P_{k-l+3} .. P_k$

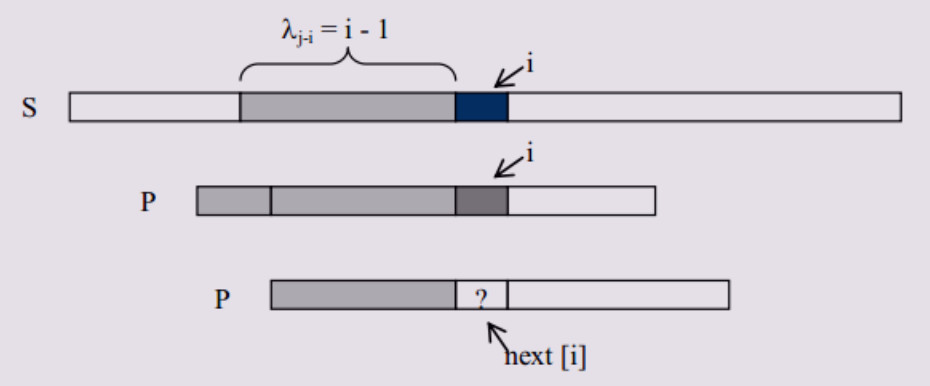
TÌM KIẾM MẪU CHÍNH XÁC THEO TIẾP CẬN OTOMAT MỜ (3)

- Thuật toán KMP mờ [7], Otomat so mẫu



TÌM KIẾM MẪU CHÍNH XÁC THEO TIẾP CẬN OTOMAT MỜ (3)

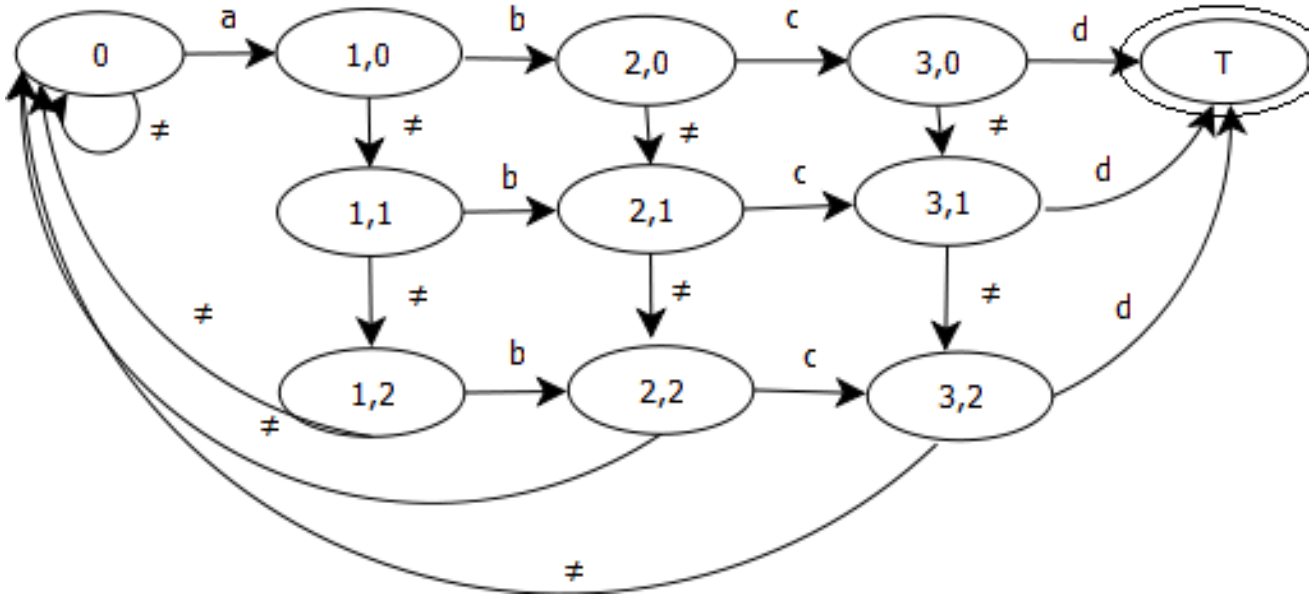
- Ví dụ: Cho mẫu P = aababaab, A = {a, b, #}, A_p = {a, b}. Theo (1) bảng TFuzz được tính toán dựa trên mảng Next như sau:



Q \ A	a	b	#
0	1	0	0
1	2	0	0
2	0	3	0
3	4	0	0
4	2	5	0
5	6	0	0
6	7	0	0
7	2	8	0
8	4	0	0

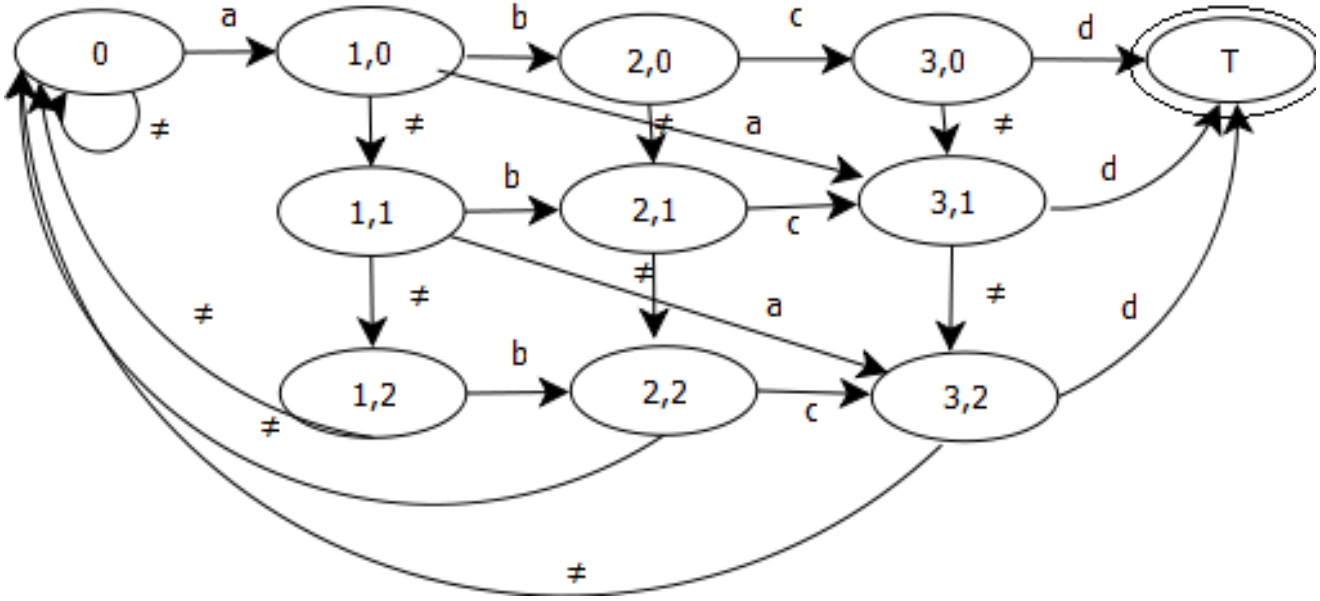
OTOMAT NÂNG CAO TÌM KIẾM XẤP XỈ (1)

- Otomat cải tiến lưu thêm trạng thái tích lũy lỗi.
- Bỏ qua lỗi để tìm kiếm xấp xỉ
- Giới hạn lỗi phạm sai bởi *ngưỡng phạm sai* π



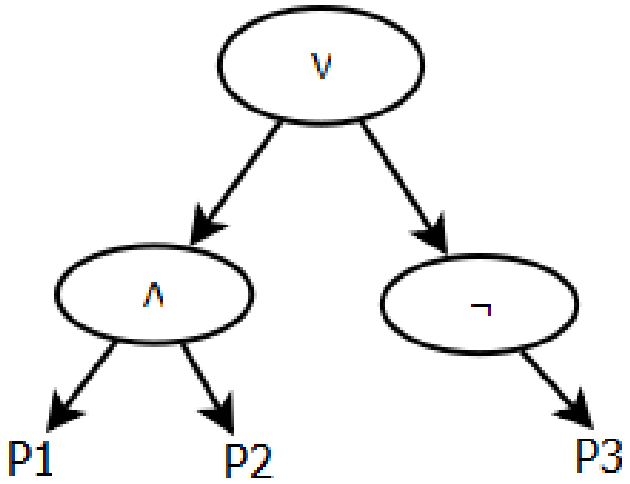
OTOMAT NÂNG CAO TÌM KIẾM XẤP XỈ (2)

- Trường hợp lỗi được bơm thêm vào P một đơn vị ε làm kích thước mẫu tăng không bình thường, giả sử ε < 2.
- Ví dụ mẫu P = “th.ông tư” có kí tự ‘.’ chèn bất thường vào P



TÌM KIẾM MẪU BIỂU THỨC ĐỘNG: LOGIC, CHÍNH QUY (1)

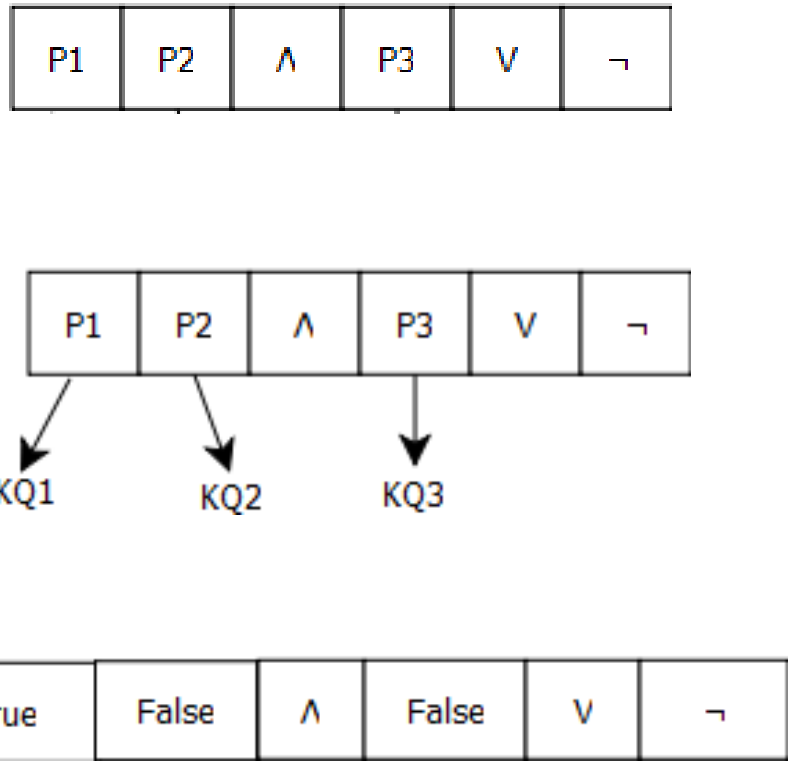
- Yêu cầu thực tế là các mẫu biểu thức động
- Ví dụ: (<thông tư> $\vee$ <văn bản>) % <nộp lệ phí>



TÌM KIẾM MẪU BIỂU THỨC ĐỘNG: LOGIC, CHÍNH QUY (2)

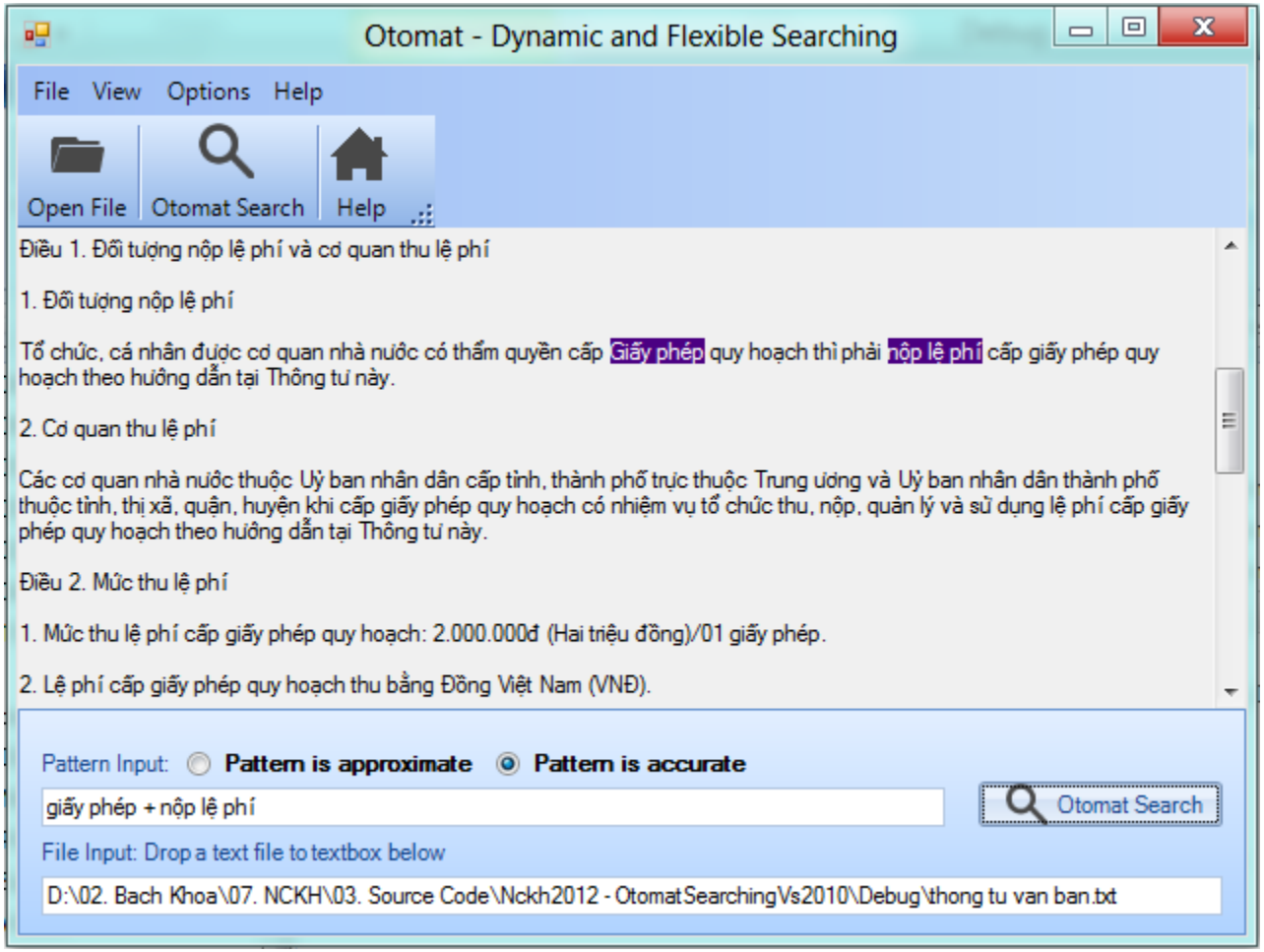
- Phương pháp giải
  - Bước 1: Chuyển mẫu biểu thức sang dạng hậu tố
  - Bước 2: Thiết lập các otomat so mẫu A_i đoán nhận P_i tương ứng
  - Bước 3: Tính giá trị biểu thức hậu tố với xâu gồm các toán tử (logic, chính quy) và toán hạng là các mẫu P_i.

TÌM KIẾM MẪU BIỂU THỨC ĐỘNG: LOGIC, CHÍNH QUY (2)



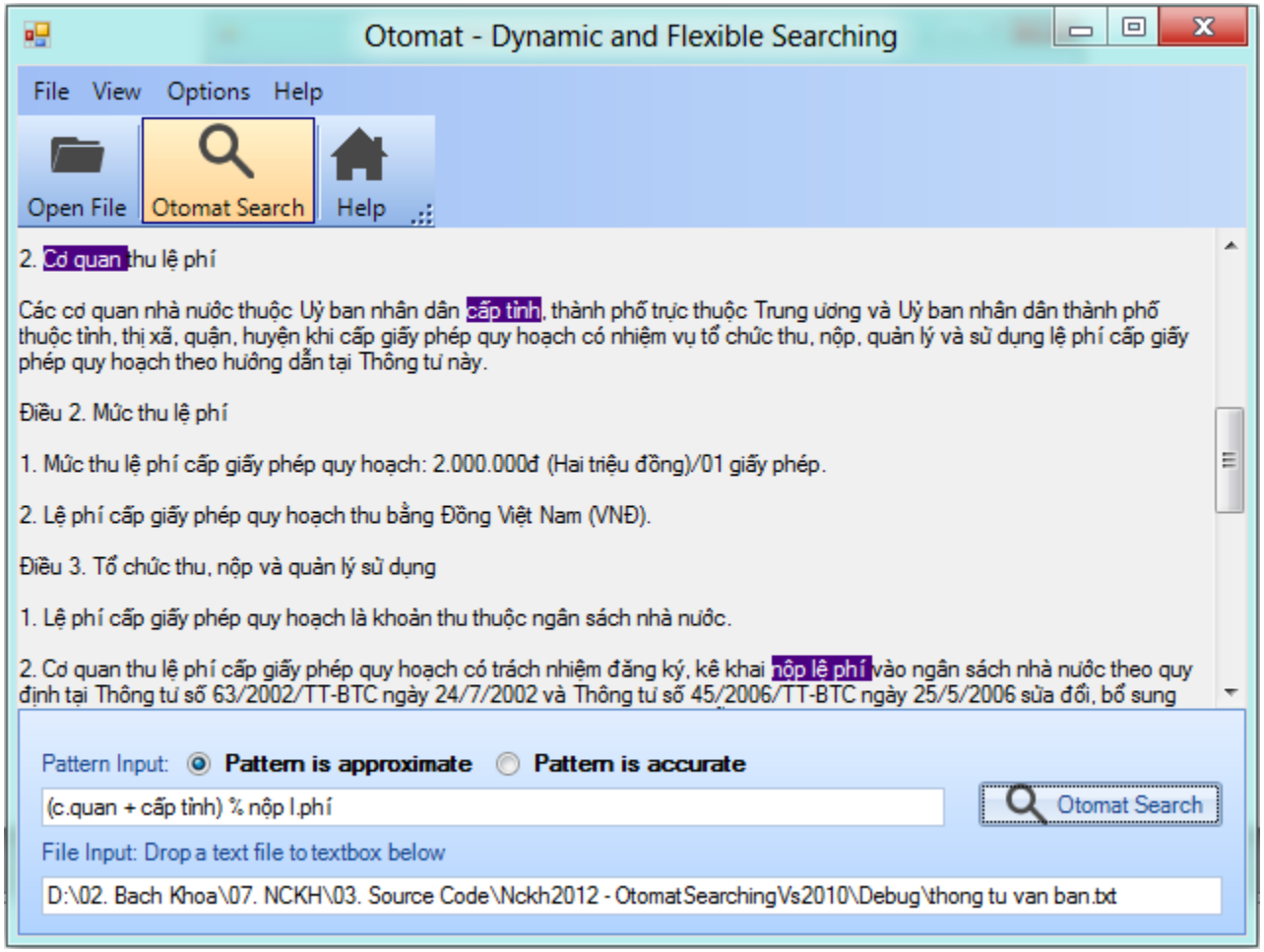
Kết quả: KQ = KQ1 $\wedge$ KQ2 $\vee$ ($\neg$ KQ3) = True $\wedge$ False $\vee$ ($\neg$ False) = True

CHƯƠNG TRÌNH THỬ NGHIỆM (1)



Tìm kiếm mẫu chính xác

CHƯƠNG TRÌNH THỬ NGHIỆM (2)



Tìm kiếm mẫu xấp xỉ

KẾT LUẬN

- Kết quả đạt được
  - Nghiên cứu tìm kiếm theo tiếp cận Otomat mờ
  - Đề xuất phương pháp tìm kiếm xấp xỉ theo tiếp cận Otomat mờ
  - Đề xuất phương pháp tìm kiếm mẫu biểu thức động: Logic, Chính quy
  - Xây dựng được chương trình thử nghiệm theo 2 phương pháp đề xuất

TÀI LIỆU THAM KHẢO

[1] Alfred V. Aho, *Algorithms for Finding Pattern in Strings*, Chapter V, Handbook of Theoretical Computer Science, Vol. A, Jan Van Leeuwen, Algorithms and Complexity, The MIT Press., 1990, pp 257-299.

[2] Christian Charras, Thierry Lecroq, *Hanbook of Exact String- matching Algorithms* , http:// www-igm.univ-mlv.fr/~lecroq/string/index.html.

[3] Maxime Crochemore, Thierry Lecroq, *Pattern matching and text compression algorithms*, 2004.

[4] Nguyễn Văn Ba, *Ngôn ngữ hình thức*, NXBKH&KT, 2002.

[5] Nguyễn Thị Thanh Huyền, Phan Trung Huy, *Tiếp cận mờ trong một số thuật toán so mẫu*, Tạp chí Tin học và điều khiển học, số 3, tập 18 (2002).

[6] Nguyễn Thị Thanh Huyền, Bùi Kiên Cường, Phan Trung Huy, *Các thuật toán tìm kiếm xâu con và tìm kiếm tựa ngữ nghĩa có sử dụng Otomat mờ* , Kỷ yếu Hội thảo quốc gia lần thứ VI về "Một số vấn đề chọn lọc của Công nghệ thông tin", Thái Nguyên, 2003.

[7] Đỗ Thị Hạnh, Tìm kiếm mờ và ứng dụng tìm kiếm thông tin trong các văn bản nén, ĐH Thái Nguyên, 2009.

EM XIN CẢM ƠN!