Truy hồi dung nạp
(Tolerant Retrieval)

Nguyễn Mạnh Hiển
hiennm@tlu.edu.vn

Câu truy vấn ký tự đại diện (wildcard query)

- Dấu * biểu thị một xâu ký tự tùy ý (có thể rỗng).
- Ví dụ:
  - Câu truy vấn *a*e*i*o*u* tìm những văn bản chứa một từ gồm tất cả 5 nguyên âm a, e, i, o và u.
  - Câu truy vấn automat* tìm những văn bản chứa một trong các từ automatic, automation và automated.
  - Nếu người dùng không nhớ chính xác tên một thành phố ở Úc là Sydney hay Sidney, thì có thể gõ câu truy vấn S*dney.
- Ta sẽ dùng **chỉ mục k-gram** để xử lý câu truy vấn ký tự đại diện.

Chỉ mục k-gram

- k-gram là một dãy k ký tự.
- Ký tự $ biểu thị điểm đầu hoặc cuối của một từ.
- Ví dụ: Các 3-gram cho từ castle là: $ca, cas, ast, stl, tle, le$.
- Chỉ mục k-gram:
  - Từ điển của chỉ mục k-gram chứa các k-gram trích rút ra từ các từ trong từ điển của trong chỉ mục chuẩn.
  - Mỗi k-gram trỏ tới một danh sách thông báo gồm các từ chứa k-gram đó.

Xử lý câu truy vấn ký tự đại diện

- Xét câu truy vấn ký tự đại diện: re*ve
  - Nghĩa là tìm những văn bản chứa một từ bắt đầu bằng re và kết thúc bằng ve.
- Giả sử ta dùng chỉ mục 3-gram (k = 3).
- Tách ra các 3-gram từ câu truy vấn ký tự đại diện, thu được câu truy vấn Boole: $re AND ve$.
- Tra cứu trong chỉ mục 3-gram, giả sử thu được danh sách các từ gồm relive, remove và retrieve.
- Với mỗi từ như vậy, lại tra cứu tiếp trong chỉ mục chuẩn.
- Hợp các kết quả (cho các từ khác nhau), thu được các văn bản cần tìm cho câu truy vấn ký tự đại diện ban đầu.

Bước hậu lọc (post-filtering)

- Chỉ mục k-gram có thể trả về những từ không đúng:
  - Xét câu truy vấn ký tự đại diện red*.
  - Chạy câu truy vấn Boole $re AND red trên chỉ mục 3-gram.
  - Kết quả trả về có thể bao gồm từ retired. Từ này chứa cả hai $re và red, nhưng không khớp với red* → cần loại bỏ từ retired.
- Bước hậu lọc:
  - Kiểm tra mỗi từ được trả về bởi chỉ mục 3-gram xem từ đó có khớp với câu truy vấn gốc red* hay không.
  - Loại những từ không khớp với câu truy vấn gốc, như từ retired.
  - Tra cứu những từ còn lại trên chỉ mục chuẩn.

Sửa lỗi chính tả

- Ví dụ:
  - Người dùng gõ vào câu truy vấn carot (một chữ "r", sai chính tả).
  - Ta muốn hệ trả về những văn bản chứa từ carrot (hai chữ "r", đúng chính tả).
- Hai nguyên tắc cơ bản khi sửa lỗi chính tả:
  1. Nếu có nhiều cách sửa lỗi cho một câu truy vấn sai chính tả, chọn cách gần nhất → cần định nghĩa **độ đo gần**.
  2. Nếu có hai cách sửa lỗi gần như nhau (ví dụ, grunt và grant đều sửa được grnt như nhau), chọn từ phổ biến hơn (trong tập văn bản hoặc trong lịch sử truy vấn).

Sửa lỗi chính tả

Có thể trình diễn kết quả tìm kiếm với người dùng theo một trong các cách sau:

1. Khi truy vấn carot, trả về các văn bản chứa carot cũng như bất cứ phiên bản sửa chính tả nào của carot, bao gồm carrot và tarot.
2. Như cách 1, nhưng chỉ khi từ truy vấn carot không có trong từ điển.
3. Như cách 1, nhưng chỉ khi câu truy vấn gốc trả về số văn bản ít hơn số đã định trước (ví dụ, ít hơn 5).
4. Khi câu truy vấn gốc trả về số văn bản ít hơn số đã định trước, giao diện người dùng hiện ra gợi ý sửa chính tả, ví dụ: “Did you mean carrot?”.

Các kiểu sửa lỗi chính tả

- Sửa cách ly: Sửa mỗi từ trong câu truy vấn một cách riêng biệt. Một số cách sửa cách ly:
  1. Dùng **khoảng cách biên tập** (edit distance).
  2. Dùng **dãy con chung dài nhất** (longest common subsequence).
  3. Dùng chỉ mục k-gram để tăng tốc hai cách bên trên.
- Sửa nhận biết ngữ cảnh: Sửa cách ly không phát hiện được lỗi chính tả trong câu truy vấn flew form Heathrow, trong đó từ form đúng chính tả khi đứng riêng nhưng sai chính tả khi xét cả cụm từ đó (đúng phải là flew from Heathrow).

Khoảng cách biên tập

- Số lượng nhỏ nhất các phép chèn/xóa/thay một ký tự để biến xâu thứ nhất thành xâu thứ hai.
- Ví dụ:
  - misspell với mispell là 1.
  - misspell với mistell là 2.
  - misspell với misspelling là 3.

Dãy con chung dài nhất

- Chiều dài của dãy con dài nhất xuất hiện trong cả hai xâu ký tự.
- Dãy con của một xâu thu được bằng cách xóa 0, 1 hoặc nhiều ký tự.
- Ví dụ:
  - misspell với mispell là 7.
  - misspelled với misinterpretted là 7 (mis…p…e…ed).

Tìm các từ tương tự

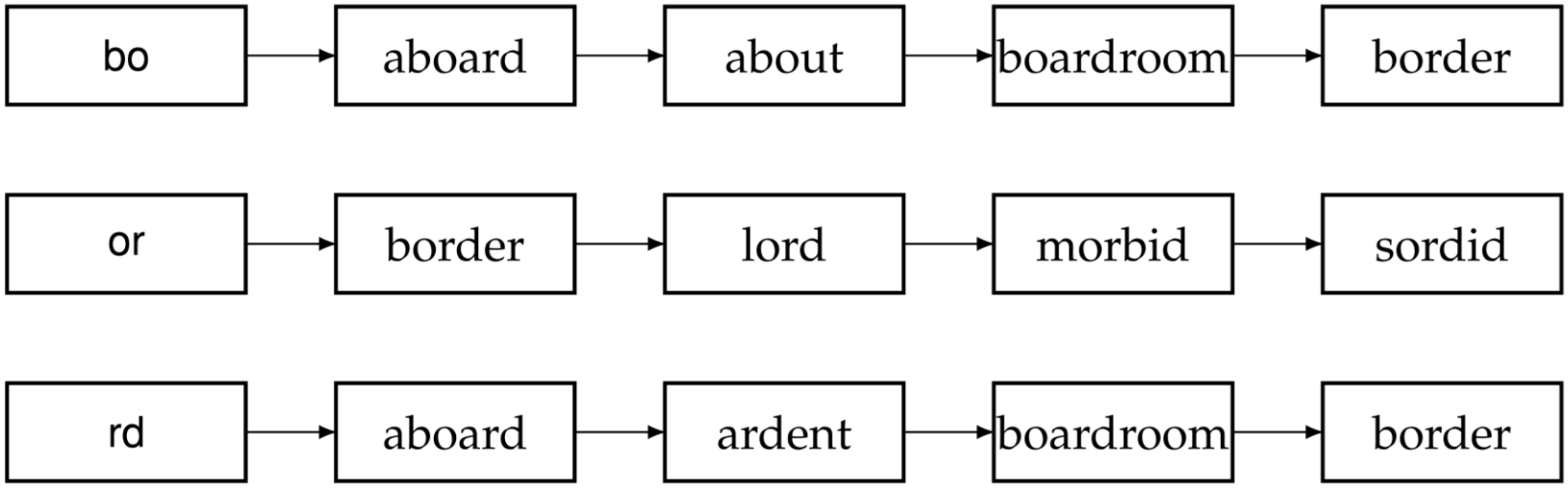
- Khi đang sửa chỉnh tả cho một từ truy vấn, sẽ không hiệu quả (chạy chậm) nếu làm như sau:
  1. Quét tuần tự qua mọi từ trong từ điển.
  2. Tính khoảng cách biên tập hoặc dãy con chung dài nhất giữa từ truy vấn và mỗi từ quét qua.
  3. Chọn ra từ tương tự nhất.
- Cách giải quyết:
  - Cách 1: Chỉ xét những từ trong từ điển có cùng chữ cái đầu với từ truy vấn. Cách này giả thiết rằng chữ cái đầu ít khi bị gõ nhầm $\rightarrow$ không đảm bảo luôn tìm ra từ gần nhất.
  - Cách 2: Dùng chỉ mục k-gram.

Tìm các từ tương tự với chỉ mục k-gram

- Xét một từ truy vấn:
  1. Sinh ra các k-gram của nó.
  2. Tìm trên chỉ mục k-gram những từ chứa các k-gram đó.
  3. Tính độ tương tự giữa các từ tìm được với từ truy vấn.
- Ví dụ: Xét câu truy vấn bord.
  1. Sinh ra các 2-gram (k = 2) của bord: { b, bo, or, rd, d }
  2. Tìm trên chỉ mục 2-gram các từ chứa ít nhất hai 2-gram ở bước 1, giả sử được: { aboard, boardroom, border }
  3. Tính điểm số (ví dụ, khoảng cách biên tập) cho các từ tìm được ở bước 2, sau đó chọn từ có điểm số tốt nhất.

Minh họa tìm từ trên chỉ mục 2-gram

Tìm từ khớp với ít nhất hai trong ba 2-gram của câu truy vấn bord. Kết quả: { aboard, boardroom, border }



Bài tập

1. Trong chỉ mục k-gram, các từ trong một danh sách thông báo được sắp xếp tăng dần. Vì sao cần điều này?

2. Xét câu truy vấn ký tự đại diện fi*mo*er. Nếu dùng 2-gram thì câu truy vấn Boole sinh ra là gì?

3. Xét câu truy vấn ký tự đại diện mon*h. Nếu dùng 2-gram và không có bước hậu lọc, hỏi các từ sau đây có khớp với câu truy vấn Boole sinh ra hay không? Từ nào khớp với câu truy vấn gốc?

(a) month

(b) moonish