PHÂN LOẠI TIN TỰ ĐỘNG CHO BÁO ĐIỆN TỬ

1. Tổng quan

Ứng dụng của Phân loại văn bản

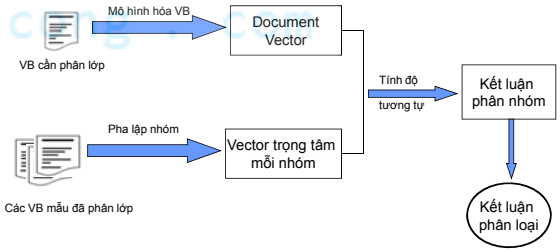
- Phân loại các tài liệu trong các thư viện
- Phân loại trong quá trình tác nghiệp của các báo điện tử.
- Phân chia sắp xếp lại các luận văn, đồ án trong các trường Đại học.
- Bộ máy tìm kiếm muốn phân chia các tài liệu trả về thành các chuyên mục → người đọc dễ nắm bắt được nội dung ban đầu của các kết quả tìm được.

1. Tổng quan

- Ứng dụng **"Phân loại tin tự động cho báo điện tử"** nhằm tìm hiểu và thử nghiệm các phương pháp phân loại văn bản áp dụng trên Tiếng Việt.
- Kết hợp giữa hai phương pháp đã được chứng minh có hiệu quả cao để giải quyết hai bài toán khác nhau là Phân loại và Lập nhóm văn bản → đề xuất một mô hình cải tiến, phù hợp với bài toán

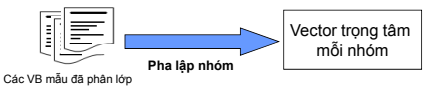
1. Tổng quan

Sơ đồ minh họa quá trình phân loại



2. Các phương pháp thực hiện

Pha lập nhóm



- Pha lập nhóm được thực hiện trước, một cách "offline" →để xác định vector trọng tâm cho mỗi nhóm cùng các thông tin truy hồi

2. Các phương pháp thực hiện (tiếp)

Tại sao cần sử dụng các phương pháp lập nhóm văn bản dựa trên thuật ngữ xuất hiện thường xuyên ?

- Kỹ thuật lập nhóm này phù hợp với yêu cầu "offline", các thuật toán áp dụng cho phương pháp này có độ chính xác cao tuy thời gian xử lý chậm và chi phí lớn, nhưng không cần thiết lắm khi xử lý offline.
- Thuật ngữ thường xuyên là các thuật ngữ xuất hiện nhiều lần trong văn bản hoặc trong một tập văn bản, các thuật ngữ phải có ý nghĩa, chúng đại diện cho nội dung toàn văn bản.
- Các thuật ngữ thường xuyên tạo nền tảng của việc khai thác quy tắc kết hợp.
- Làm giảm được số chiều của vector biểu diễn tài liệu.

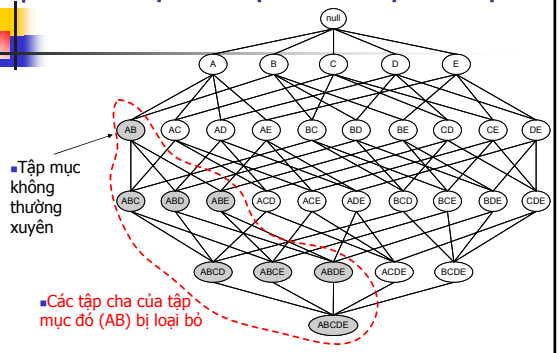
Giảm bớt số lượng các tập mục cần xét

- Nguyên tắc của giải thuật Apriori – Loại bỏ (prunning) dựa trên độ hỗ trợ
  - Nếu một tập mục là thường xuyên, thì tất cả các tập con (subsets) của nó đều là các tập mục thường xuyên
  - Nếu một tập mục là không thường xuyên (not frequent), thì tất cả các tập cha (supersets) của nó đều là các tập mục không thường xuyên
- Nguyên tắc của giải thuật Apriori dựa trên **đặc tính không đơn điệu (anti-monotone) của độ hỗ trợ**

$$\forall X,Y:(X\subseteq Y)\Rightarrow s(X)\geq s(Y)$$

  - Độ hỗ trợ của một tập mục nhỏ hơn độ hỗ trợ của các tập con của nó

Apriori: Loại bỏ dựa trên độ hỗ trợ



2. Các phương pháp thực hiện (tiếp)

- Bước 1 : Giải thuật Apriori – tính toán các tập thuật ngữ thường xuyên

Giải thuật Apriori
Biến C_k: Các tập thuật ngữ ứng cử mức k.
Biến L_k: Các tập thuật ngữ thường xuyên mức k.

L_1 = {Các thuật ngữ thường xuyên mức 1};
For (k=1; L_k!=Ø; k++) do Begin
// Lặp lại cho đến khi không có thêm bất kỳ tập mục thường xuyên nào mới
//Bước kết hợp: Kết hợp L_k với bản thân nó để tạo ra C_{k+1}
//Bước cắt tỉa: Loại bỏ (k+1)-itemsets từ C_{k+1} chứa k-itemsets không thường xuyên
  C_{k+1} = các ứng cử viên được tạo ra từ L_k
  For mỗi tài liệu t trong tập văn bản do
    Tăng số lượng của tất cả các ứng cử viên trong C_{k+1} có chứa trong t
    L_{k+1} = các ứng cử viên trong C_{k+1} có GS > min_support
End
Return L_k

2. Các phương pháp thực hiện (tiếp)

Bước 2 : sử dụng thuật toán FIHC để phân nhóm các tập thuật ngữ thường xuyên ra. **(Frequent Item-based Hierarchical Clustering)**

Thuật toán FIHC bao gồm hai giai đoạn :

- **Xây dựng các *Cluster* khởi tạo.**
- **Dựng cây *Cluster*.**

3.Chương trình thực nghiệm

Mô hình

- Phần tiền xử lý văn bản làm các công việc như tách thuật ngữ, phân tích tổ chức dữ liệu, tổ chức từ điển.
- Pha lập nhóm văn bản, sử dụng thuật toán Apriori và FIHC.
- Khi phân loại một văn bản mới ứng dụng chỉ việc đọc các thông tin về vector trọng tâm, so sánh với văn bản đầu vào đã được vector hóa → quyết định phân loại.

3.Chương trình thực nghiệm

Phần tiền xử lý văn bản.

- **Tách thuật ngữ tiếng Việt** : Sử dụng thuật toán đối sánh thuật ngữ dài nhất từ bên phải qua.

Ví dụ : *Ban công tác đã xác định được vấn đề.*

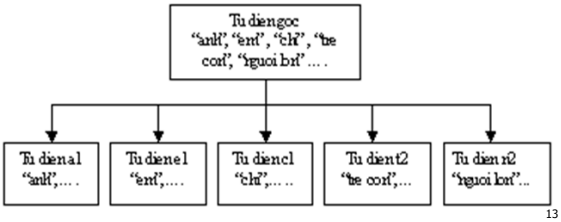
Khi sử dụng thuật toán từ phải qua, ta sẽ tách được chính xác câu này. Kết quả như sau : *vấn đề, được, xác định, đã, công tác, ban*. Và ta chỉ cần đảo ngược lại thứ tự này.

3.Chương trình thực nghiệm

Phần tiền xử lý văn bản.

- **Phân tích tổ chức dữ liệu: (1)**

Tổ chức từ điển dưới dạng cấu trúc như sau:



3.Chương trình thực nghiệm

Phân tích tổ chức dữ liệu: Xây dựng 3 File đầu vào

1. File **ClassID.txt** là file chứa ID và tên của các class, được tạo bằng cách duyệt qua tất cả các thư mục con của thư mục chứa tập văn bản mẫu.

Ví dụ nội dung 1 file ClassID.txt

0: Dulich
1: Giaoduc
2: Oto xe may
3: Suckhoe
4: The thao
5: Vitinh
6: Kinhdoanh

3.Chương trình thực nghiệm

2. **File ThreeLine.txt** chứa các thông số chung của quá trình lập nhóm, gồm 3 dòng:
   - Tổng số nhóm phân ra từ tập văn bản mẫu
   - Số lớp (số thư mục con) của tập văn bản mẫu.
   - Số lượng các nhóm phân bổ vào từng lớp tương ứng bên file *ClassID.txt.*
- **Ví dụ** nội dung một file *ThreeLine.txt* :

  174
  8
  20 22 22 16 27 14 14 39

3.Chương trình thực nghiệm

3. File **InputForYou.txt** chứa các vectơ trọng tâm của tất cả các nhóm, 1 vectơ / dòng.
- Thông tin trên 1 dòng
  - Số văn bản thuộc nhóm/vectơ trọng tâm đó;
  - ID của lớp mà nhóm đó thuộc về;
  - ID của nhóm đó trong lớp;
  - Các cặp (Term ID – Trọng số) thể hiện cho các chiều của vector trọng tâm

4. Đánh giá kết quả

Xây dựng mẫu kiểm thử

- Tập kiểm thử được xây dựng từ các bài báo thuộc các lĩnh vực khác nhau của báo điện tử VnExpress (http://www.vnexpress.net)
- Dữ liệu kiểm thử là 56 bản tin mới nhất trên VNExpress thuộc các chủ đề Giáo dục, Du lịch, Kinh doanh, Ô tô xe máy, Thể Thao, Pháp luật, Vi Tính, Sức khoẻ (theo sự phân chia chủ đề của báo) đã được ghi lại theo chủ đề từ trước.
- Độ chính xác : 94,64%.

4. Đánh giá kết quả

- Mô hình cải tiến đạt được độ chính xác cao.
- Dữ liệu nói chung đã tối ưu
- Các chức năng được phân tách rõ ràng làm giảm chi phí tài nguyên và tăng tốc độ phân lớp lên rất nhiều.
- Hai thuật toán *Apriori*, *FIHC* tuy đạt được độ chính xác cao nhưng chưa ổn định.

Hướng phát triển

- Các thuật toán *Apriori*, *FIHC* tuy được cài đặt để sử dụng trong thời gian xử lý "offline" nhưng chi phí tính toán cũng khá lớn. → cải tiến các thuật toán này để giảm chi phí lập nhóm
- Việc tiền xử lý văn bản như xử lý thống nhất font chữ, định dạng file đầu vào và đặc biệt là quá trình tách thuật ngữ có ảnh hưởng quan trọng đối với hệ thống xử lý văn bản nói chung và ứng dụng phân loại tin tự động nói riêng. Đây cũng là một vấn đề cần được nghiên cứu sâu hơn và đưa ra các giải thuật tốt hơn

19