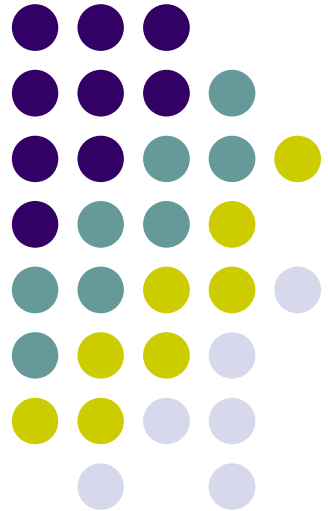
BÀI GIẢNG KHAI PHÁ DỮ LIỆU WEB

CHƯƠNG 5. BIỂU DIỄN WEB

TS. TRẦN MAI VŨ
HÀ NỘI 9-2012
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ
ĐẠI HỌC QUỐC GIA HÀ NỘI



Nội dung

Giới thiệu

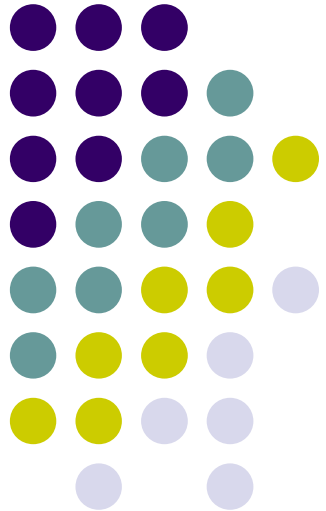
Phân tích văn bản

Biểu diễn Text

Lựa chọn đặc trưng

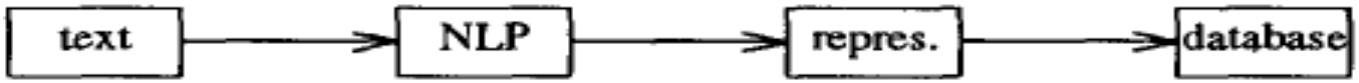
Thu gọn đặc trưng

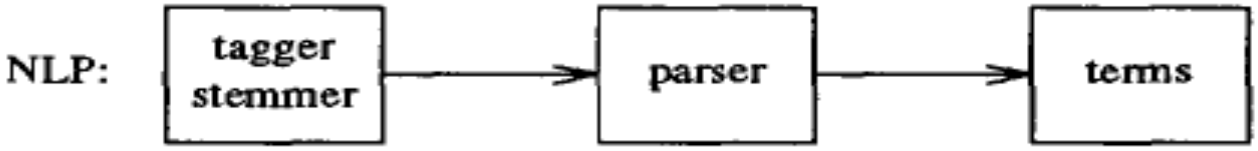
Biểu diễn Web



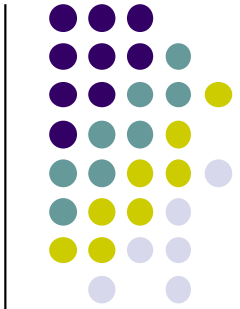
Giới thiệu

- Biểu diễn văn bản
  - Là bước cần thiết đầu tiên trong xử lý văn bản
  - Phù hợp đầu vào của thuật toán khai phá dữ liệu
  - Tác động tới chất lượng kết quả của thuật toán KHDL
  - Thuật ngữ tiếng Anh: (document/text) (representation/indexing)
- Phạm vi tác động của một phương pháp biểu diễn văn bản
  - Không tồn tại phương pháp biểu diễn lý tưởng
  - Tồn tại một số phương pháp biểu diễn phổ biến
  - Chọn phương pháp biểu diễn phù hợp miền ứng dụng
- Một sơ đồ sơ lược: Tomek Strzalkowski: Document Representation in Natural Language Text Retrieval, *HLT 1994*: 364-369





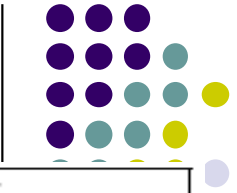
Nghiên cứu về biểu diễn văn bản



- Nghiên cứu biểu diễn văn bản (Text + Web)
  - Luôn là nội dung nghiên cứu thời sự
  - Biểu diễn Web bổ sung một số yếu tố cho biểu diễn Text
- Số công trình liên quan
  - "Document representation”
    - mọi nơi: 8000 bài; tiêu đề: 200 (60 bài từ 2006-nay)
  - “Document indexing”
    - mọi nơi: 5200 bài; tiêu đề: 220 (60 bài từ 2006-nay)
  - “Text representation”
    - mọi nơi: 9200 bài; tiêu đề: 240 (60 bài từ 2006-nay)
  - “Text indexing”
    - mọi nơi: 6800 bài; tiêu đề: 210 (60 bài từ 2006-nay)

Ghi chú: các bài “ở mọi nơi” phần đông thuộc vào các bài toán xử lý văn bản bao gồm bước trình bày văn bản

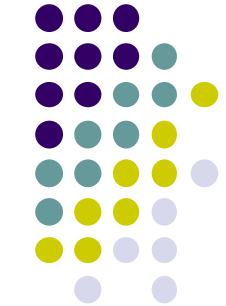
Nghiên cứu về biểu diễn văn bản (2)



Research paper reference	Document Representation	Feature Selection	Learning algorithm
Apté et al. [6]	bag-of-words (freq)	stop list+ frequency	Decision Rules
Armstrong et al. [7]	bag-of-words	informativity	TFIDF Winnow, WordStat
Balabanović et al. [9]	bag-of-words (freq)	stop list+stemming+ keep 10 best words	TFIDF
Bartell et al. [11]	bag-of-words (freq)	latent semantic indexing using SVD	—
Berry et al. [12] Foltz and Dumais [28]	bag-of-words(freq)	latent semantic indexing using SVD	TFIDF
Cohen [21]	bag-of-words	infrequent words pruned	Decision Rules ILP
Joachims [40]	bag-of-words (freq)	infrequent words+ informativity	TFIDF, PrTFIDF, Naive Bayes
Lam et al. [60]	bag-of-words (freq)	mutual info.	Bayesian Network
Lewis et al. [66]	bag-of-words	log likelihood ratio	logistic regression with Naive Bayes
Maes [69]	bag-of-words+ header info.	mail/news header + selecting keywords	Memory-Based reasoning
Pazzani et al. [83, 84]	bag-of-words	stop list+ informativity	TFIDF, Naive Bayes, Nearest Neighbor, Neural Network, Decision Trees
Sorensen and Mc Elligott [97, 25]	n-gram graph (only bigrams)	weighting graph edges	connectionist combined with Genetic Algorithms
Yang [100]	bag-of-words	stop list	adapted k-Nearest Neighbor

Dunja Mladenic' (1998). Machine Learning on Non-homogeneous, Distributed Text Data. *PhD. Thesis*, University of Ljubljana, Slovenia.

Phân tích văn bản



- Mục đích biểu diễn văn bản (Keen, 1977 [Lew91])
  - Từ được chọn liên quan tới chủ đề người dùng quan tâm
  - Gắn kết các từ, các chủ đề liên quan để phân biệt được từ ở các lĩnh vực khác nhau
  - Dự đoán được độ liên quan của từ với yêu cầu người dùng, với lĩnh vực và chuyên ngành cụ thể
- Môi trường biểu diễn văn bản (đánh chỉ số)
  - Thủ công / từ động hóa. Thủ công vẫn có hỗ trợ của công cụ máy tinh và phần mềm
  - Điều khiển: chọn lọc từ làm đặc trưng (feature) biểu diễn) / không điều khiển: mọi từ đều được chọn.
  - Từ điển dùng để đánh chỉ số. Từ đơn và tổ hợp từ.

Luật Zipt

- Luật Zipt
  - **Cho dãy dữ liệu được xếp hạng** $x_1 \geq x_2 \geq \ldots \geq x_n$

thì hạng tuân theo công thức

$$x_{(r)} = \frac{C}{r^{\alpha}}$$

C là hằng số, α gần 1; kỳ vọng dạng loga

$$E(\log x_{(r)}) = c - \alpha \log(r)$$

  - Dạng hàm mật độ:

$$p(x) = \frac{C^{1/\alpha}}{\alpha n} \frac{1}{x^{(1/\alpha)+1}} = \frac{A}{x^{\beta}}$$

- Một số dạng khác
  - Phân phối Yule

$$x_{(r)} = \frac{C}{r^{\alpha} B^{r}}$$

  - Mô hình thống kê

$$E(\log x_{(r)}) = c - \alpha \log(r) - be^{\log(r)}$$

c=log(C), b= log(B)

  - Biến thể loga-chuẩn

$$E(\log x_{(r)}) = c - \alpha \log(r) - b(\log(r))^2$$

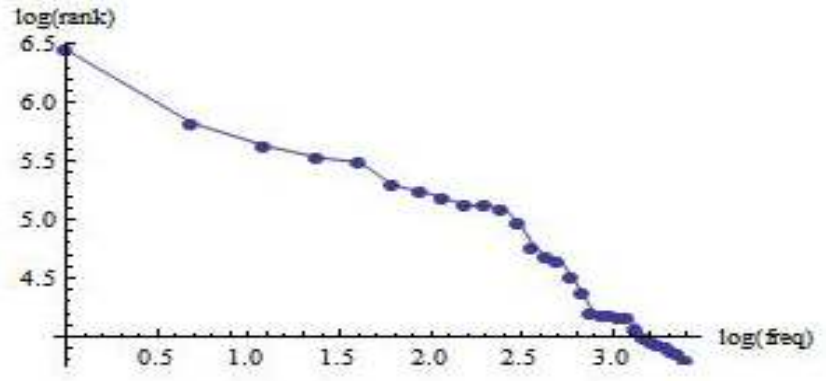
  - Phân phối Weibull với $0<\beta<1$

$$E(\log x_{(r)}) = c - \alpha \log(r) - be^{\beta \log(r)}$$

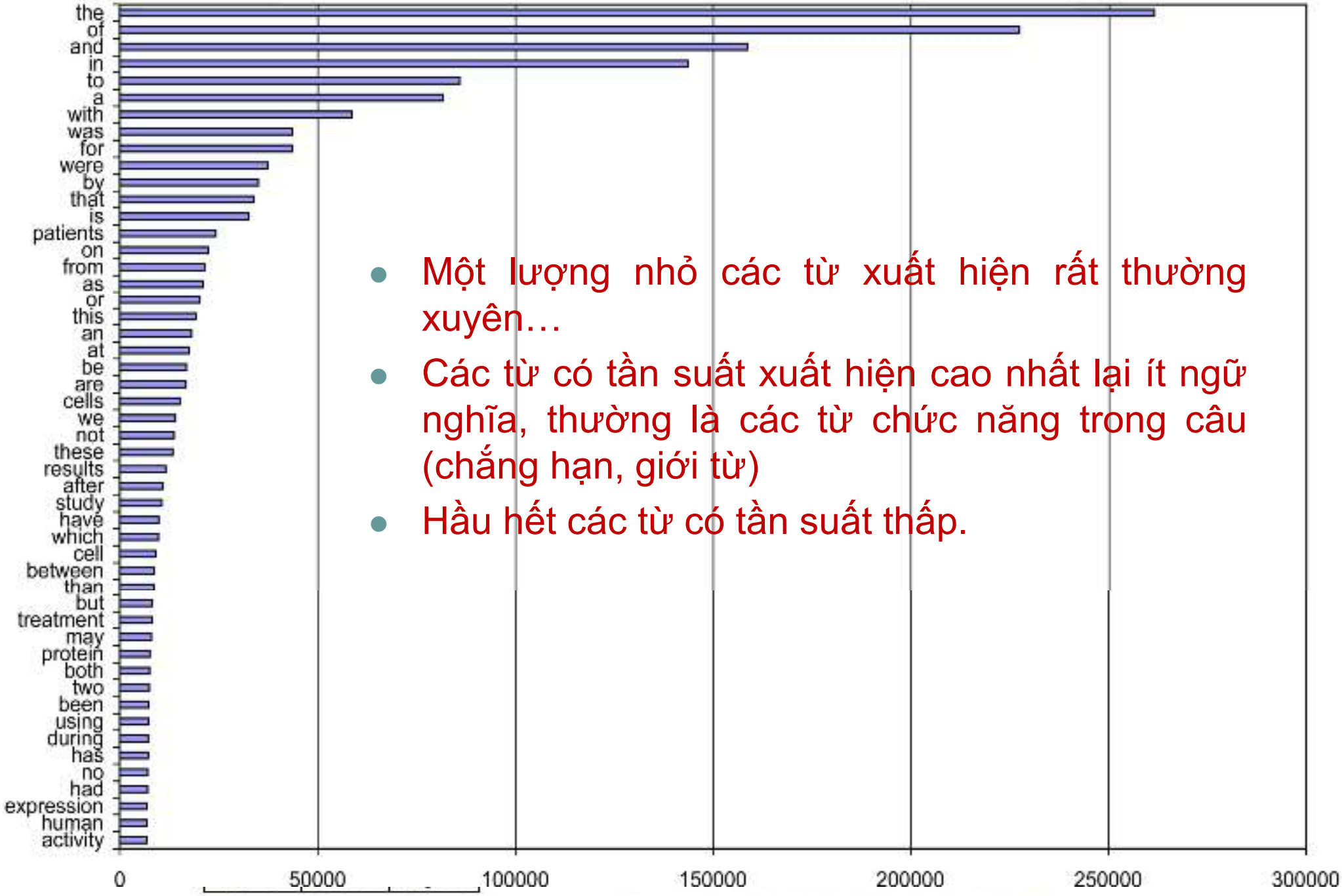
Luật Zipt trong phân tích văn bản

- Trọng số của từ trong biểu diễn văn bản (Luhn, 1958)
  - Dấu hiệu nhấn mạnh: một biểu hiện của độ quan trọng
    - thường viết lặp lại các từ nhất định khi phát triển ý tưởng
    - hoặc trình bày các lập luận,
    - phân tích các khía cạnh của chủ đề. …
  - Các từ có tần suất xuất hiện cao nhất lại ít ngữ nghĩa. Từ xuất hiện trung bình lại có độ liên quan cao.
- Luật Zipt
  - Là một quan sát hiện tượng mà không phải là luật thực sự: xem hình vẽ "Alice ở xứ sở mặt trời"
  - $r_t * f_t = K$ (hằng số): r_t : độ quan trọng của từ t; f_t: tần số xuất hiện từ t. Có thể logarith

the 632	and 338	a 278
to 252	she 242	of 199
it 189	i 178	was 167
alice 167	in 163	said 144
you 118	her 108	that 105
as 91	at 79	with 67
s 66	had 65	all 64
on 64	little 59	out 54
down 52	this 51	t 50
for 48	but 47	they 45



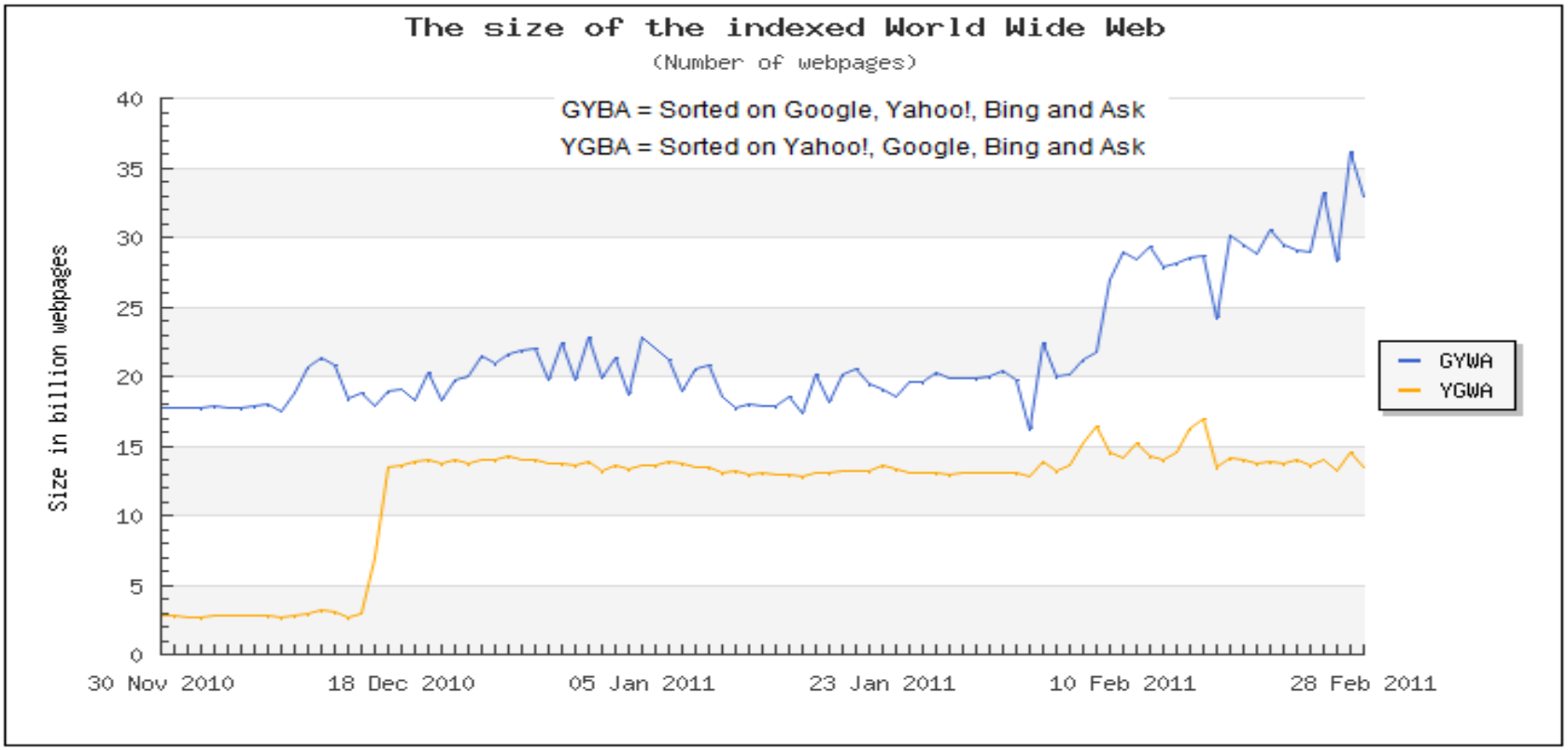
Luật Zipt trong tiếng Anh



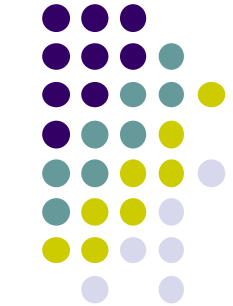
- Một lượng nhỏ các từ xuất hiện rất thường xuyên…
- Các từ có tần suất xuất hiện cao nhất lại ít ngữ nghĩa, thường là các từ chức năng trong câu (chẳng hạn, giới từ)
- Hầu hết các từ có tần suất thấp.

Luật Zipt: ước lượng trang web được chỉ số

- Ước lượng tối thiểu lượng trang web chỉ số hóa
  - http://www.worldwidewebsize.com/
  - Luật Zipt: từ kho ngữ liệu DMOZ có hơn 1 triệu trang web
  - Dùng luật Zipt để ước tính lượng trang web chỉ số hóa.
  - Mỗi ngày: 50 từ (đều ở đoạn logarith luật Zipt) gửi tới 4 máy tìm kiếm Google, Bing, Yahoo Search và Ask.
  - Trừ bớt phần giao ước tính giữa các công cụ tìm kiếm: làm già
  - Thứ tự trừ bớt phần giao → tổng (được làm non)



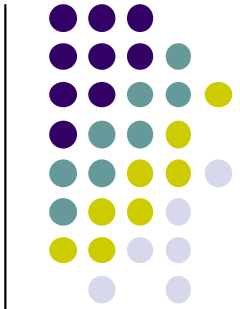
Các mẫu luật Zipt khác



- Dân số thành phố
  - Dân số thành phố trong một quốc gia: có α = 1. Đã xác nhận ở 20 quốc gia.
  - Có thể mở rộng sang: dân cư khu đô thị, vùng lãnh thổ
- Lượt thăm trang web và mẫu giao vận Internet khác
  - Số lượt truy nhập trang web/tháng
  - Các hành vi giao vận Internet khác
- Quy mô công ty và một số số liêu kinh tế khác
  - Xếp hạng công ty theo: số nhân viên, lợi nhuận, thị trường
  - Các hành vi giao vận Internet khác
- …

[Li02] Wentian Li (2002). Zipf's Law Everywhere, *Glottometrics* ***5*** *(2002)*: 14-21

Phương pháp lựa chọn từ Luhn58

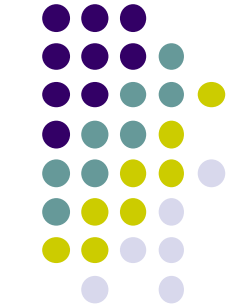


- Bài toán
  - Input: Cho một tập văn bản: có thể coi tất cả các văn bản trong miền ứng dụng; ngưỡng trên, ngưỡng dưới dương.
  - Output: Tập từ được dùng để biểu diễn văn bản trong tập
- Giải pháp
  - Tính tần số xuất hiện mỗi từ đơn nhất trong từng văn bản
  - Tính tần số xuất hiện của các từ trong tập toàn bộ văn bản
  - Sắp xếp các từ theo tần số giảm dần
  - Loại bỏ các từ có tần số xuất hiện vượt quá ngưỡng trên hoặc nhỏ thua ngưỡng dưới.
  - Các từ còn lại được dùng để biểu diễn văn bản
  - "Từ" được mở rộng thành "đặc trưng": n-gram, chủ đề..
- Lưu ý
  - *Chọn ngưỡng*: ngưỡng cố định, ngưỡng được điều khiển
  - Liên hệ vấn đề chọn lựa đặc trưng (mục sau).

Phương pháp đánh trọng số của từ

- Bài toán
  - *Input*: Cho một tập văn bản miền ứng dụng D và tập từ được chọn biểu diễn văn bản V (sau bước trước đây).
  - *Output*: Đánh trọng số từ trong mỗi văn bản ⇨ Xây dựng ma trận $\{w_{i,j}\}$ là trọng số của từ $w_i \in V$ trong văn bản $d_j \in D$.
- Giải pháp
  - Một số phương pháp điển hình
  - Boolean
  - dựa theo tần số xuất hiện từ khóa
  - Dựa theo nghịch đảo tần số xuất hiện trong các văn bản
- Phương pháp Boolean
  - Đơn giản: trọng số là xuất hiện/ không xuất hiện
  - $w_{i,j} = 1$ nếu w_i xuất hiện trong văn bản d_j, ngược lại $w_{i,j} = 0$.

Các phương pháp đánh trọng số của từ theo tần số



- Dạng đơn giản: TF
  - $w_{i,j} = f_{i,j}$: trong đó $f_{i,j}$ là số lần từ khóa w_i xuất hiện trong văn bản d_j
- Một số phiên bản khác của dạng đơn giản
  - Cân đối số lần xuất hiện các từ khóa: giảm chênh lệch số lần xuất hiện
  - Giảm theo hàm căn $w_{i,j} = \sqrt{tf_{ij}}$
  - Tránh giá trị "0" và giảm theo hàm loga: $w_{i,j} = 1+\log(f_{i,j})$
- Nghịch đảo tần số xuất hiện trong tập văn bản: IDF
  - Từ xuất hiện trong nhiều văn bản thì trọng số trong 1 văn bản sẽ thấp
  - $w_i = \log(\frac{m}{df_i}) = \log(m) - \log(df_i)$

Trong đó $m = |D|$, $df_i = |d \in D: w_i$ xuất hiện trong $d|$

Phương pháp TFIDF

- Tích hợp TF và IDF
  - Dạng đơn giản: $w_{i,j} = f_{i,j} * m/df_i$
  - Dạng căn chỉnh theo hàm loga

$$w_{i,j} = \begin{cases} (1+\log(tf_{ij}))\log(\frac{m}{df_i}) & : tf_{ij} > 0 \\ 0 & : tf_{ij} = 0 \end{cases}$$

  - Ngoài ra, có một số dạng tích hợp trung gian khác

Mô hình biểu diễn văn bản

- Bài toán
  - *Input*: Cho tập văn bản miền ứng dụng D = {d_j }, tập đặc trưng được chọn biểu diễn văn bản V = {w_i }, ma trân trọng số W = ($w_{i,j}$) .
  - *Output*: Tìm biểu diễn của các văn bản $d_j \in$D.
- Một số mô hình
  - Mô hình Boolean
  - Mô hình không gian vector
  - Mô hình túi các từ (Mô hình xác suất)
  - Các mô hình khác
- Mô hình Boolean
  - Tập các từ thuộc V mà xuất hiện trong văn bản

Mô hình không gian vector

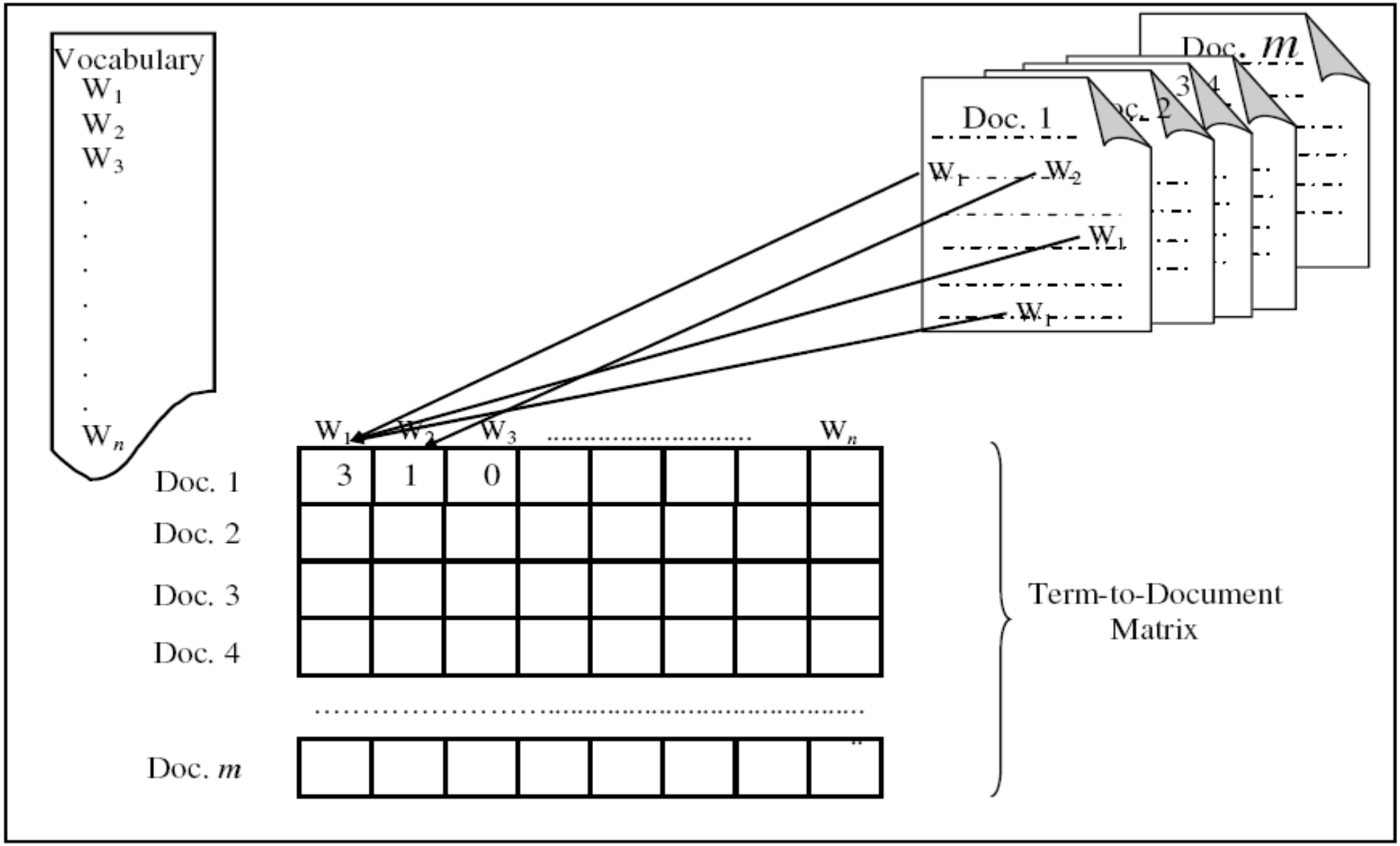
- Nội dung chính
  - Ánh xạ tập tài liệu vào không gian vector n =|V| chiều.
  - Mỗi tài liệu được ánh xạ thành 1 vector

$$d_i \Leftrightarrow (w_{i1}, w_{i2}, \ldots, w_{in})$$

- Độ đo tương tự nội dung văn bản
  - Chuẩn hóa vector: đưa về độ dài 1
    - Độ "tương tự nội dung" giữa hai văn bản ⇔ độ tương tự giữa hai vector
    - Một số phương án sơ khai "các thành phần giống nhau", "nghịch đảo khoảng cách", ..
  - Phổ biến là tính độ đo cosin của góc giữa hai vector: không yêu cầu chuẩn hóa

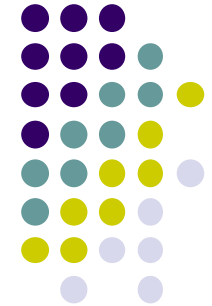
$$sim(d_1, d_2) = \frac{(v_1, v_2)}{\|v_1\| \|v_2\|} = \frac{\sum_{i=1}^{n} w_{1i} * w_{12}}{\sqrt{\sum_{i=1}^{n} w_{1i}^2} * \sqrt{\sum_{i=1}^{n} w_{2i}^2}}$$

Mô hình không gian vector



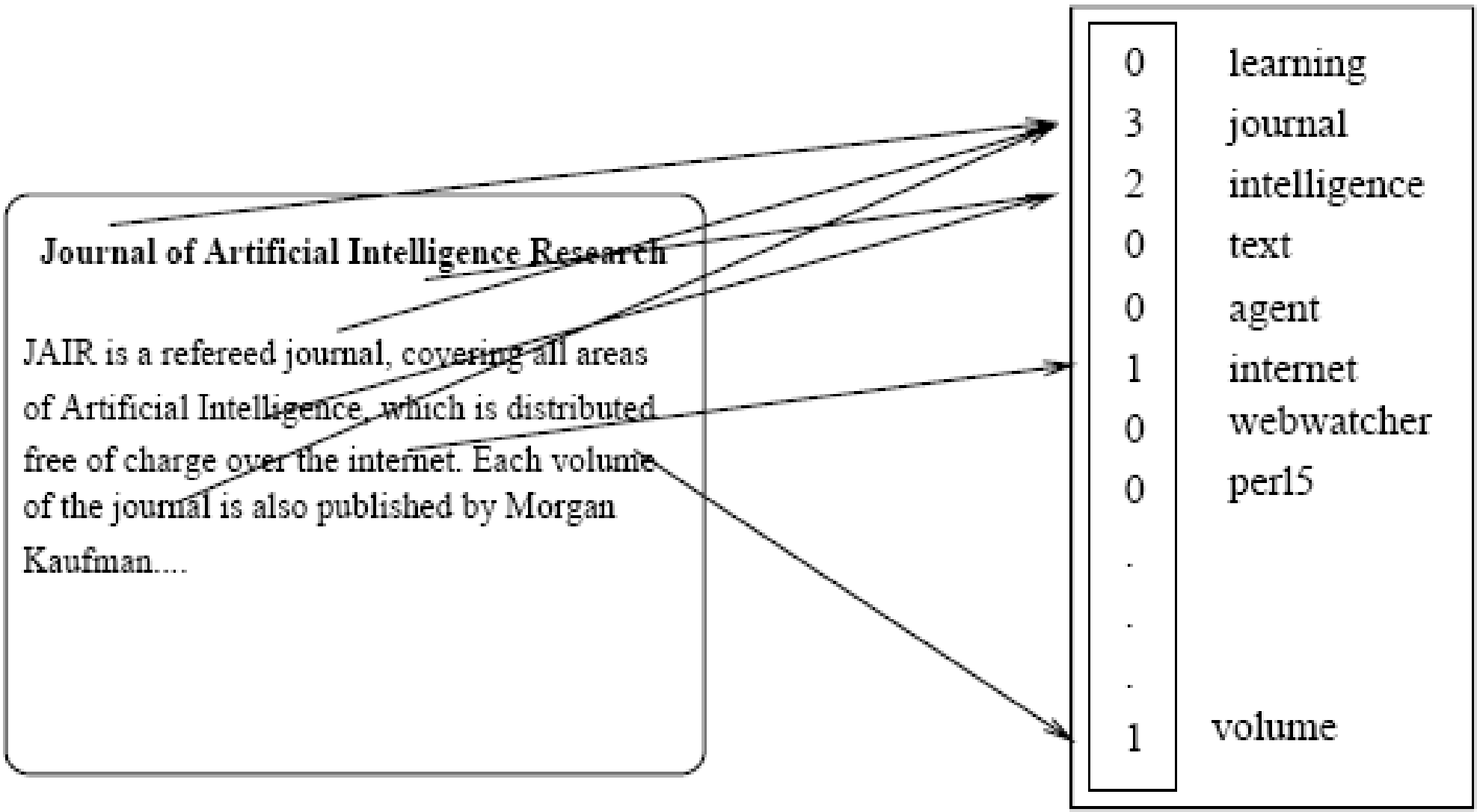
Khaled Shaban (2006). A semantic graph model for text representation and matching in document mining, *PhD Thesis*, University of Waterloo, Canada

Mô hình xác suất



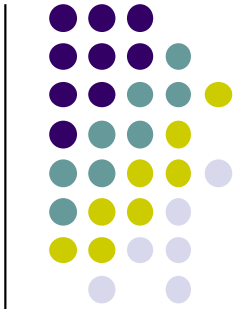
- Giả thiết chính
  - Mô hình xác suất: cặp (Y, P) với Y là tập quan sát được và P là mô hình xác suất trên Y (có thể coi Y là quan sát được các từ/đặc trưng trên văn bản).
  - Các từ xuất hiện trong văn bản thể hiện nội dung văn bản
  - Sự xuất hiện của các từ là độc lập lẫn nhau và độc lập ngữ cảnh
  - Dạng đơn giản: chỉ liệt kê từ, dạng chi tiết: liệt kê từ và số lần xuất hiện
  - *Lưu ý*: Các giả thiết về tính độc lập không hoàn toàn đúng (độc lập lẫn nhau, độc lập ngữ cảnh) song mô hình thi hành hiệu quả trong nhiều trường hợp.
- Độ đo tương tự nội dung văn bản
  - So sánh hai túi từ

Mô hình túi từ (bag-of-word)



Dunja Mladenic' (1998). Machine Learning on Non-homogeneous, Distributed Text Data. *PhD. Thesis*, University of Ljubljana, Slovenia.

Mô hình biểu diễn LSI và theo phân cụm

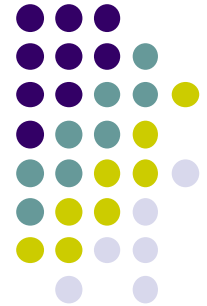


- Giới thiệu
  - Tồn tại nhiều phương pháp biểu diễn khác
  - Tồn tại nhiều phiên bản cho một phương pháp
  - Gần đây có một số phương pháp mới
  - Hai phương pháp phổ biến: LSI và theo phân cụm
  - Lưu ý: Giá phải trả khi *tiền xử lý dữ liệu*
- Mô hình phân cụm
  - Phân cụm các từ trong miền ứng dụng: ma trận trọng số
  - Thay thếtừ bằng cụm chứa nó
- Mô hình biểu diễn LSI
  - LSI: Latent Semantic Indexing biểu diễn ngữ nghĩa ẩn
    - Nâng mức ngữ nghĩa (trừu tượng) của đặc trưng
    - Rút gọn tập đặc trưng, giảm số chiều không gian biểu diễn
    - Không gian từ khóa ⇨ không gian khái niệm (chủ đề).
  - Phương pháp chuyển đổi
    - Ma trận trọng số ⇨ ma trận hạng nhỏ hơn
    - Phép biến đổi đó Từ khóa ⇨ khái niệm. Thay thế biểu diễn.

Lựa chọn từ trong biểu diễn văn bản

- Loại bỏ từ dừng
  - Những từ được coi là không mạng nghĩa
  - Có sẵn trong ngôn ngữ
- Đưa về từ gốc
  - Các ngôn ngữ có biến dạng từ: Anh, Nga…
  - Thay từ biến dạng về dạng gốc
- Chon đặc trưng n-gram
  - Các âm tiết liền nhau n-gram
    - Uni-gram: chỉ chứa một âm tiết
    - Bigram: chứa không quá 2 âm tiết
    - Trigram: chứa không quá 2 âm tiết
    - N-gram: Thường không quá 4 gram
  - Một số đặc trưng
    - Chính xác hơn về ngữ nghĩa
    - Tăng số lượng đặc trưng
    - Tăng độ phức tạp tính toán

Một số đô đo cho lựa chọn đặc trưng



- Giới thiệu chung
  - Lựa chọn đặc trưng: lợi thế chính xác, lợi thể tốc độ hoặc cả hai
  - Các độ đo giúp khẳng định lợi thế
- Phân nhóm độ đo
  - Hai nhóm: theo tần số và theo lý thuyết thông tin
- Một số độ đo điển hình
  - Xem hai trang sau

Một số đô đo cho lựa chọn đặc trưng

P(t_k, c_i) kí hiệu là xác suất của từ t_k có trong chủ đề c_i và P(t_k, $\bar{c}_i$) là xác suất của từ t_k không có trong chủ đề c_i.

1. DIA (Darmstadt Indexing Approach – Tiếp cận đánh chỉ số Darmstadt): Được đề xuất bởi Fuhn và đồng nghiệp [FHK91].

$$f(t_k, c_i) = z(t_k, c_i) = P(c_i|t_k)$$

2. Độ đo IG (Information Gain).

$$f(t_k, c_i) = IG(t_k, c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} P(t, c) \cdot \log \frac{P(t, c)}{P(t) \cdot P(c)}$$

3. Độ đo thông tin tương hỗ (mutual information).

$$f(t_k, c_i) = \log \frac{P(t_k, c_i)}{P(t_k) \cdot P(c_i)}$$

4. Độ đo Khi -bình phương (Chi-square).

$$f(t_k, c_i) = \chi^2(t_k, c_i) = \frac{|Tr| \cdot [P(t_k, c_i) \cdot P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i \cdot P(\bar{t}_k, c_i))]^2}{P(t_k) \cdot P(\bar{t}_k) \cdot P(c_i) \cdot P(\bar{c}_i)}$$

5. Độ đo liên quan (Relevancy score).

$$f(t_k, c_i) = RS(t_k, c_i) = \log \frac{P(t_k|\bar{c}_i) + d}{P(\bar{t}_k|\bar{c}_i) + d}$$

6. Tỷ lệ dư (Odd Ratio).

$$f(t_k, c_i) = OR(t_k, c_i) = \frac{P(t_k|c_i) \cdot (1 - P(t_k|\bar{c}_i))}{(1 - P(t_k|c_i)) \cdot P(t_k, |\bar{c}_i)}$$

Một số đô đo cho toàn bộ các lớp

Các độ đo trên là tính cho từng lớp. Độ đo cho toàn bộ các lớp trong tập hợp có thể được tính theo nhiều cách khác nhau,

$$f(t_k) = \sum_{i=1}^{|C|} f(t_k, c_i)$$

hoặc

$$f(t_k) = \sum_{i=1}^{|C|} P(c_i) f(t_k, c_i)$$

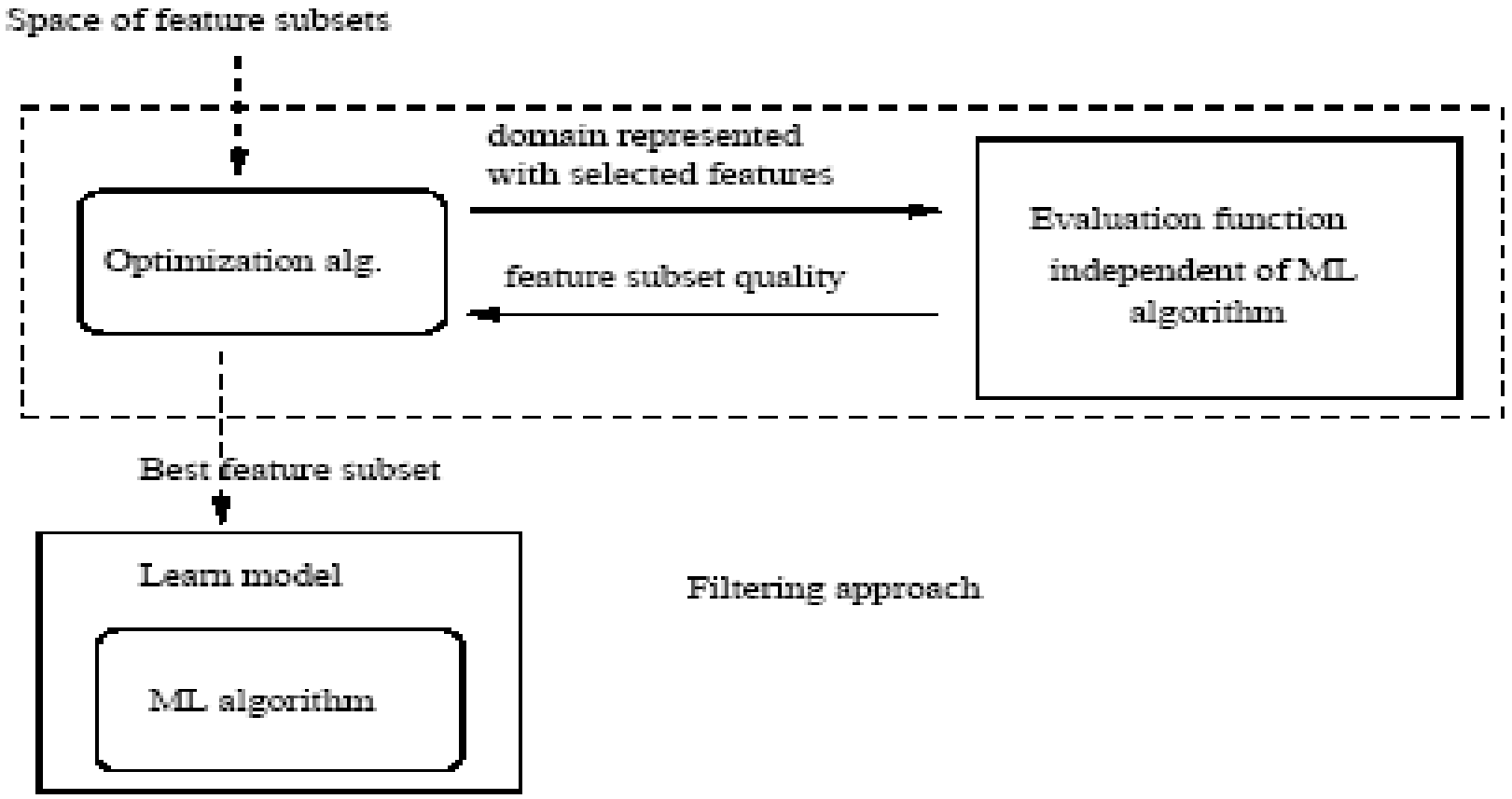
hoặc

$$f(t_k) = \max_{i=1}^{|C|} f(t_k, c_i).$$

Thu gọn đặc trưng

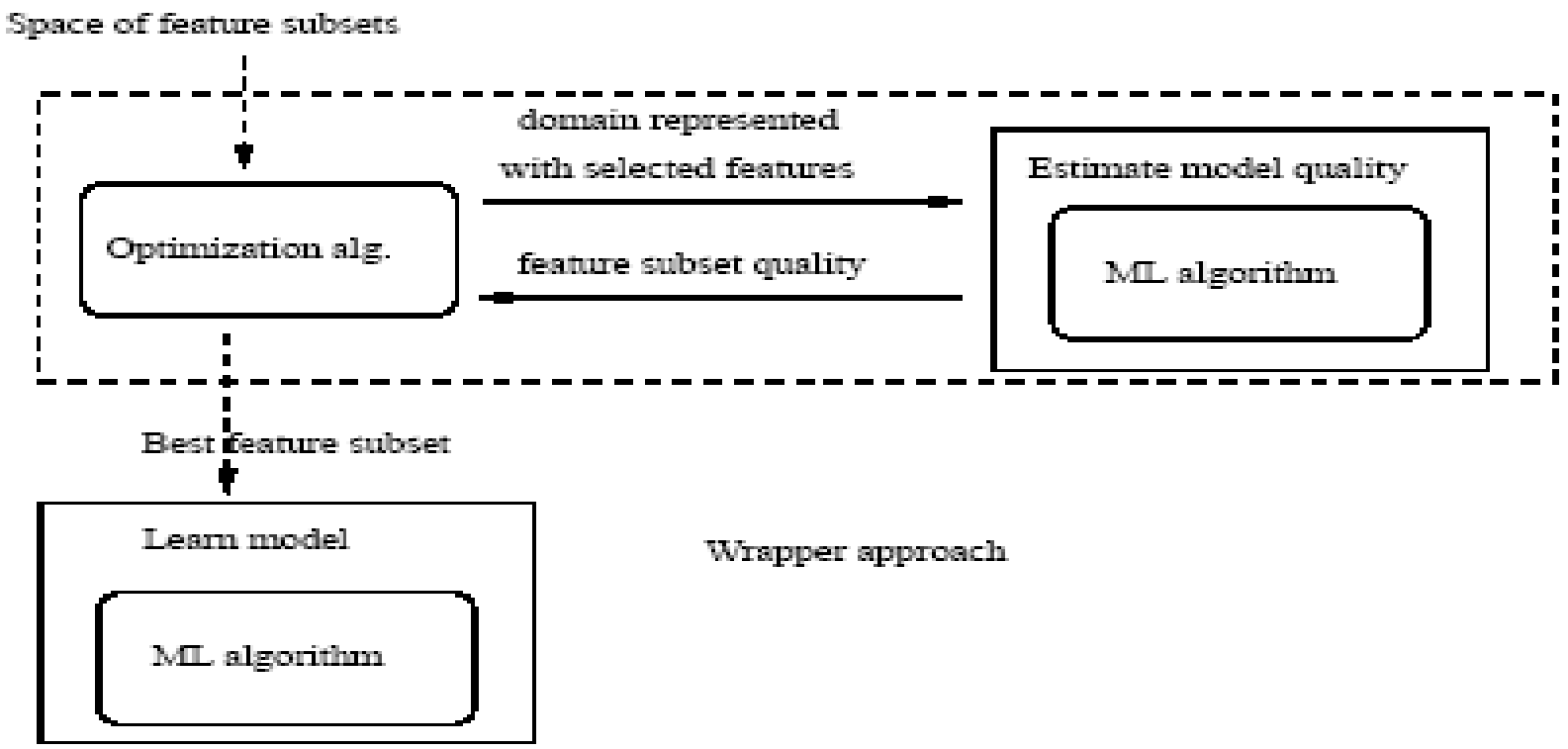
- Giới thiệu chung
  - "Tối ưu hóa" chọn tập đặc trưng
    - Số lượng đặc trưng nhỏ hơn
    - Hy vọng tăng tốc độ thi hành
    - Tăng cường chất lượng khai phá văn bản. ? Giảm đặc trưng đi là tăng chất lượng: có các đặc trưng "nhiễu"
    - Hoặc cả hai mục tiêu trên
- Hai tiếp cận điển hình
  - Tiếp cận lọc
  - Tiếp cận bao gói
- Với dữ liệu văn bản
  - Tập đặc trưng: thường theo mô hình vector
  - Tính giá trị của từng đặc trưng giữ lại các đặc trưng được coi là "tốt".

Tiếp cận tổng quát: lọc



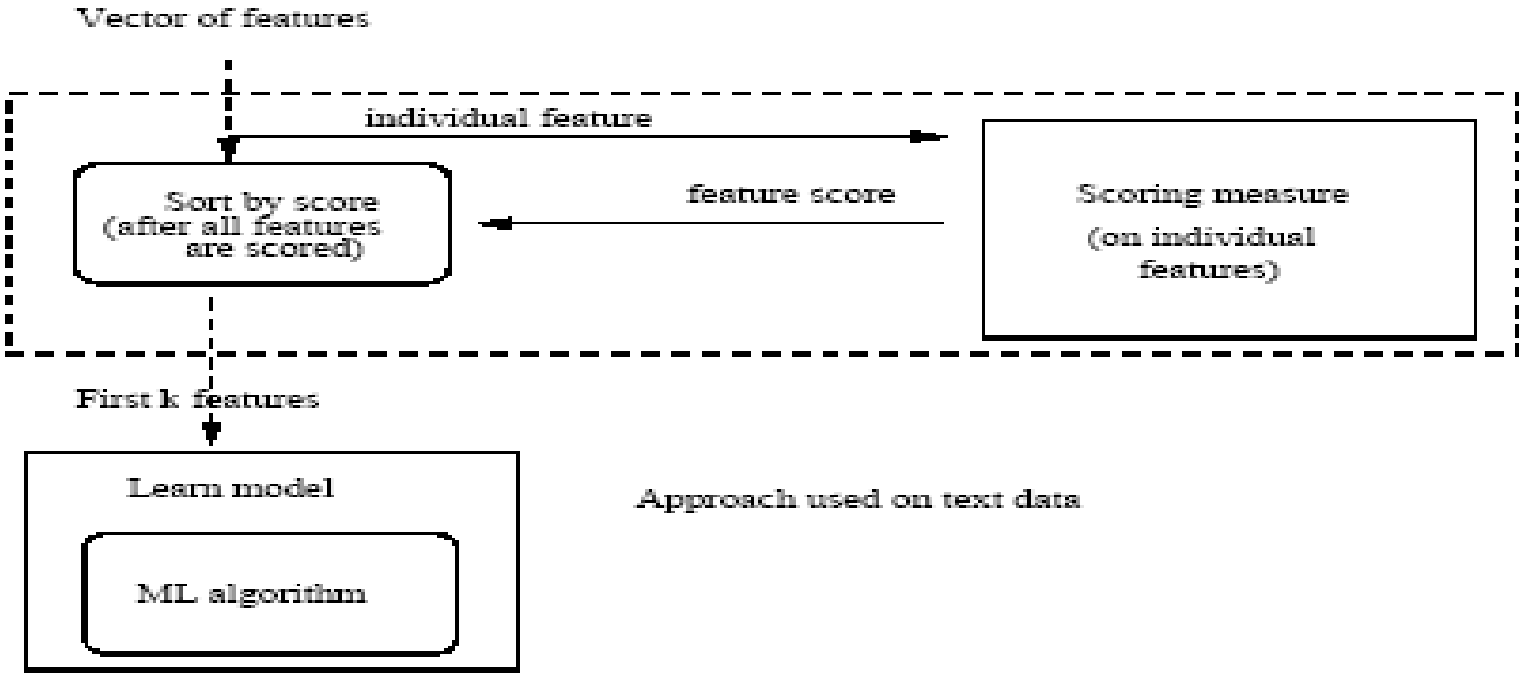
- Tiếp cận lọc
  - Đầu vào: Không gian tập các tập đặc trưng
  - Đầu ra: Tập con đặc trưng tốt nhất
  - Phương pháp
    - Dò tìm "cải tiến" bộ đặc trưng: Thuật toán tối ưu hóa
    - Đánh giá chất lượng mô hình: độc lập với thuật toán học máy

Tiếp cận bao gói tổng quát



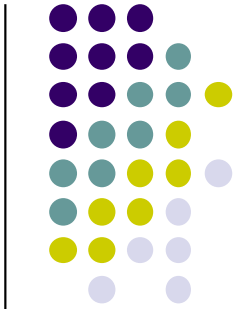
- Tiếp cận bao gói
  - Đầu vào: Không gian tập các tập đặc trưng
  - Đầu ra: Tập con đặc trưng tốt nhất
  - Phương pháp
    - Dò tìm "cải tiến" bộ đặc trưng: Thuật toán tối ưu hóa
    - Đánh giá chất lượng mô hình: Dùng chính thuật toán học để đánh giá

Thu gọn đặc trưng văn bản text



- Thu gọn đặc trưng
  - Đầu vào: Vector đặc trưng
  - Đầu ra: k đặc trưng tốt nhất
  - Phương pháp (lùi)
    - Sắp xếp các đặc trưng theo độ "tốt" (để loại bỏ bớt)
    - Tính lại độ "tốt" của các đặc trưng
    - Chọn ra k-đặc trưng tốt nhất
- Các kiểu phương pháp
  - Tiến / Tiến bậc thang (có xem xét thay thế khi tiến)
  - Lùi / Lùi bậc thang (có xem xét thay thế khi lùi)

Thu gọn đặc trưng phân lớp text nhị phân



```
SELECTFEATURES(𝔻, c, k)
1  V ← EXTRACTVOCABULARY(𝔻)
2  L ← []
3  for each t ∈ V
4  do A(t, c) ← COMPUTEFEATUREUTILITY(𝔻, t, c)
5     APPEND(L, ⟨A(t, c), t⟩)
6  return FEATURESWITHLARGESTVALUES(L, k)
```

- Một thuật toán lựa chọn đặc trưng text
  - V: Bảng từ vựng có được từ tập văn bản D
  - c: lớp đang được quan tâm
  - giá trị A(t,c): một trong ba thủ tục tính toán
- Ba kiểu thủ tục tính toán A(t,c)
  - Thông tin tương hỗ
  - Lựa chọn đặc trưng theo khi-bình phương (chi-square)
  - Lựa chọn đặc trưng theo tần suất

Thu gọn đặc trưng: thông tin tương hỗ

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t)P(C = e_c)}$$

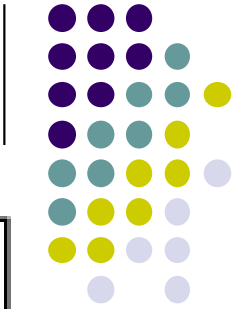
$$I(U;C) = \frac{N_{11}}{N} \log_2 \frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N} \log_2 \frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N} \log_2 \frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N} \log_2 \frac{NN_{00}}{N_{0.}N_{.0}}$$

- Công thức MI (Mutual Information)
  - Biến ngẫu nhiên U: từ khóa t *xuất hiện*/*không xuất hiện*
  - Biến ngẫu nhiên c: tài liệu *thuộc*/*không thuộc* lớp c
  - Ước lượng cho MI
- Ví dụ: Bộ dữ liệu Reuter-RCV1
  - Lớp poultry, từ khóa export

	$e_t = e_{\text{export}} = 1$	$e_t = e_{\text{export}} = 0$
$e_c = e_{\text{poultry}} = 1$	$N_{11} = 49$	$N_{01} = 141$
$e_c = e_{\text{poultry}} = 0$	$N_{10} = 27{,}652$	$N_{00} = 774{,}106$

$$I(U;C) = \frac{49}{801{,}948} \log_2 \frac{801{,}948 \cdot 49}{(49+27{,}652)(49+141)} + \frac{141}{801{,}948} \log_2 \frac{801{,}948 \cdot 141}{(141+774{,}106)(49+141)}$$
$$+ \frac{27{,}652}{801{,}948} \log_2 \frac{801{,}948 \cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)} + \frac{774{,}106}{801{,}948} \log_2 \frac{801{,}948 \cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)}$$
$$\approx 0.000105$$

10 đặc trưng tốt nhất cho 6 lớp



UK	
london	0.1925
uk	0.0755
british	0.0596
stg	0.0555
britain	0.0469
plc	0.0357
england	0.0238
pence	0.0212
pounds	0.0149
english	0.0126

China	
china	0.0997
chinese	0.0523
beijing	0.0444
yuan	0.0344
shanghai	0.0292
hong	0.0198
kong	0.0195
xinhua	0.0155
province	0.0117
taiwan	0.0108

poultry	
poultry	0.0013
meat	0.0008
chicken	0.0006
agriculture	0.0005
avian	0.0004
broiler	0.0003
veterinary	0.0003
birds	0.0003
inspection	0.0003
pathogenic	0.0003

coffee	
coffee	0.0111
bags	0.0042
growers	0.0025
kg	0.0019
colombia	0.0018
brazil	0.0016
export	0.0014
exporters	0.0013
exports	0.0013
crop	0.0012

elections	
election	0.0519
elections	0.0342
polls	0.0339
voters	0.0315
party	0.0303
vote	0.0299
poll	0.0225
candidate	0.0202
campaign	0.0202
democratic	0.0198

sports	
soccer	0.0681
cup	0.0515
match	0.0441
matches	0.0408
played	0.0388
league	0.0386
beat	0.0301
game	0.0299
games	0.0284
team	0.0264

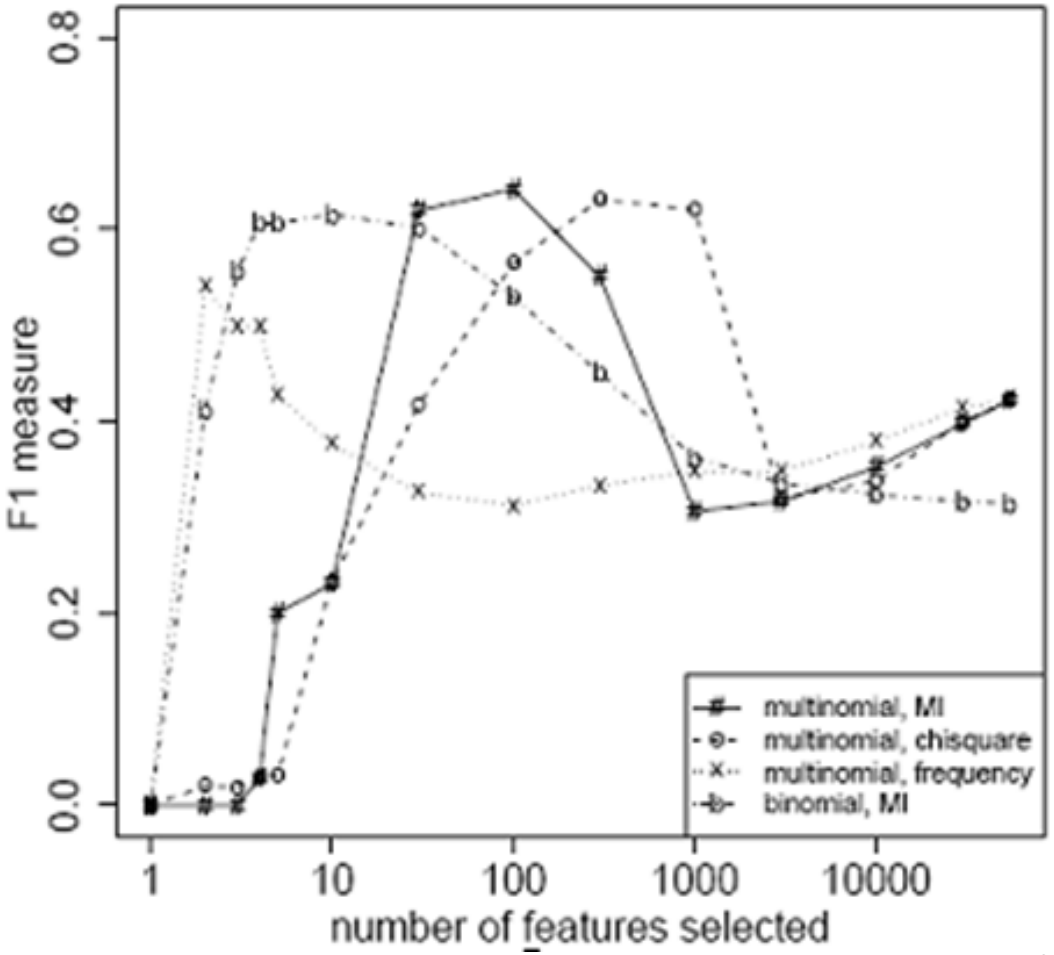
Bộ dữ liệu Reuter-RCV1

Thống kê khi-bình phương và tần số

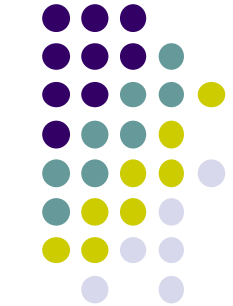
$$X^2(\mathbb{D},t,c) = \sum_{e_t\in\{0,1\}}\sum_{e_c\in\{0,1\}} \frac{(N_{e_te_c} - E_{e_te_c})^2}{E_{e_te_c}}$$

$$X^2(\mathbb{D},t,c) = \frac{(N_{11}+N_{10}+N_{01}+N_{00}) \times (N_{11}N_{00} - N_{10}N_{01})^2}{(N_{11}+N_{01}) \times (N_{11}+N_{10}) \times (N_{10}+N_{00}) \times (N_{01}+N_{00})}$$

- Thống kê khi-bình phương
  - Công thức xác suất: e_t, e_c : như MI, các biến E là kỳ vọng, N là tần số quan sát được từ tập tài liệu D
  - Ước lượng cho MI: các giá trị N như MI
- Tần số
  - Một ước lượng xác suất

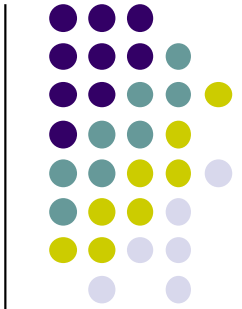


Thu gọn đặc trưng phân lớp text đa lớp



- Bài toán phân lớp đa lớp
  - Tập C = {c_1, c_2, …, c_n)
  - Cần chọ đặc trưng tốt nhất cho bộ phân lớp đa lớp
- Phương pháp thống kê khi-bình phương
  - Mỗi từ khóa
    - Lập bảng xuất hiện/không xuất hiện các đặc trưng trong lớp văn bản
    - Tính giá trị thống kê khi-bình phương
  - Chọn k đặc trưng (từ khóa)
- Phương pháp lựa chon từng lớp
  - Tính bộ đặc trưng tốt cho từng phân lớp thành phần
  - Kết hợp các bộ đặc trưng tốt
    - Tính toán giá trị kết hợp: trung bình (có trọng số xuất hiện) khi kết hợp
    - Chọn k-đặc trưng tốt nhất sau khi tính toán kết hợp

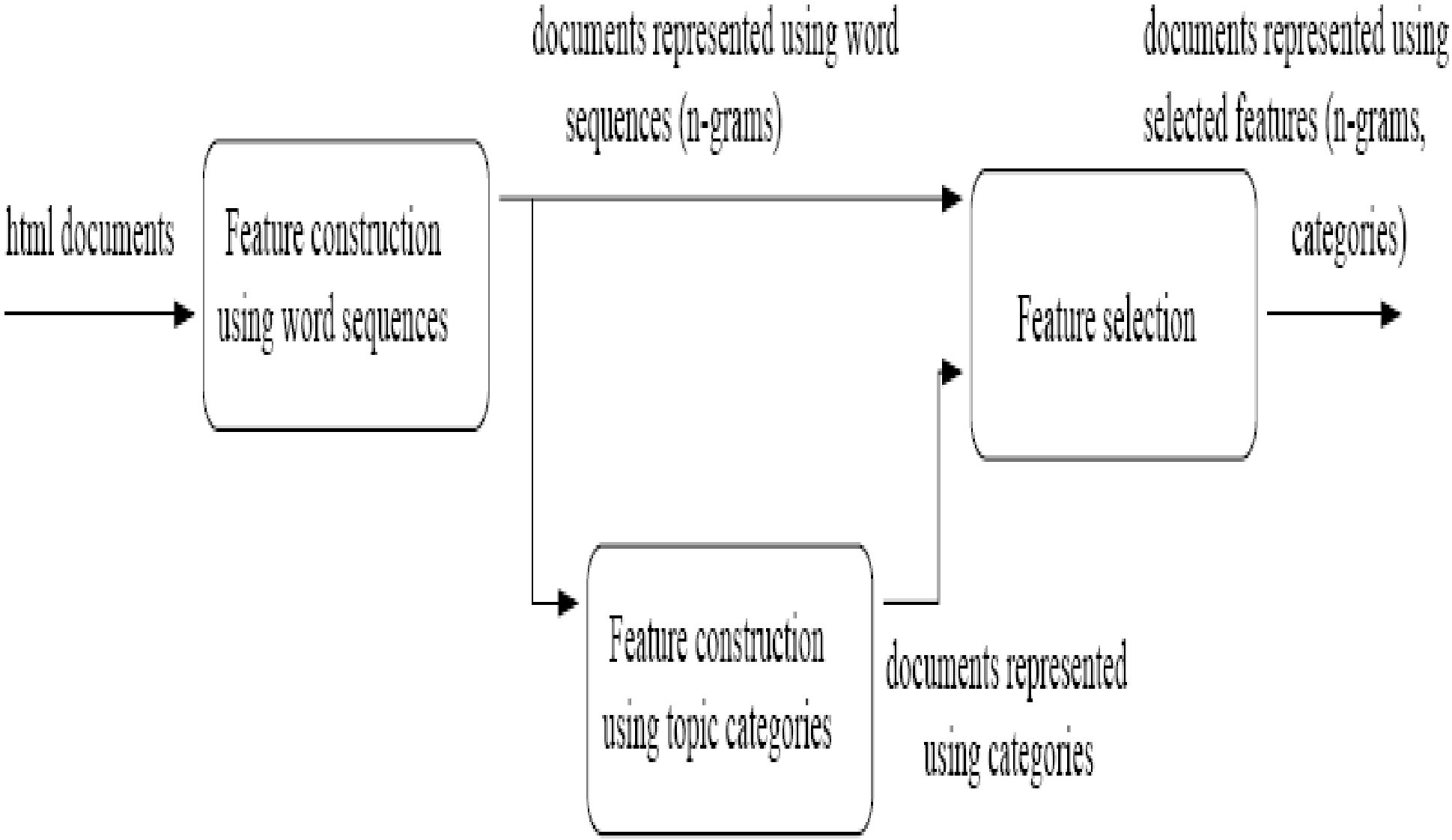
Biểu diễn Web



- Đồ thị Web
  - Web có cấu trúc đồ thị
    - Đồ thị Web: nút ⇔ trang Web, liên kết ngoài ⇔ cung (có hướng, vô hướng).
    - Bản thân trang Web cũng có tính cấu trúc cây (đồ thị)
  - Một vài bài toán đồ thị Web
    - Biểu diễn nội dung, cấu trúc
    - Tính hạng các đối tượng trong đồ thị Web: tính hạng trang, tính hạng cung..

  Nghiên cứu về đồ thị Web (xem trang sau)
- Đồ thị ngẫu nhiên
  - Tính ngẫu nhiên trong khai phá Web
    - WWW có tính ngẫu nhiên: mới, chỉnh sửa, loại bỏ
    - Hoạt động con người trên Web cũng có tính ngẫu nhiên
  - Là nội dung nghiên cứu thời sự

Một sơ đồ biểu diễn tài liệu Web



Dunja Mladenic' (1998). Machine Learning on Non-homogeneous, Distributed Text Data. *PhD. Thesis*, University of Ljubljana, Slovenia.

Một sơ đồ biểu diễn tài liệu Web

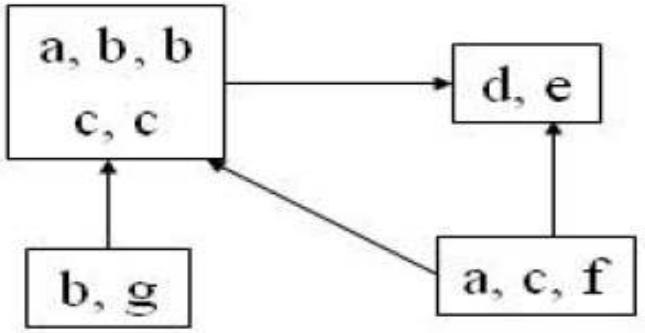
Các biểu diễn vector trang Web

Phương pháp 1:

a	b	c	d	e	f	g
1	2	2	0	0	0	0

Phương pháp 2:

a	b	c	d	e	f	g
2	3	3	1	1	1	1



Phương pháp 3:

Đoạn 1							Đoạn 2						
a	b	c	d	e	f	g	a	b	c	d	e	f	g
1	2	2	0	0	0	0	1	1	1	1	1	1	1

Phương pháp 4:

Đoạn1							Đoạn 2							Đoạn 3							Đoạn 4						
a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g	a	b	c	d	e	f	g
1	2	2	0	0	0	0	0	0	0	1	1	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1
1	2	2	0	0	0	0	1	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0

Dunja Mladenic' (1998). Machine Learning on Non-homogeneous, Distributed Text Data. *PhD. Thesis*, University of Ljubljana, Slovenia.

Một sơ đồ biểu diễn tài liệu Web

Máy tìm kiếm từ khóa nhanh: Hệ thống chỉ số ngược (Inverted Index)