UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

TRUY VẤN THÔNG TIN ĐA PHƯƠNG TIỆN
INFORMATION RETRIEVAL

CRAWLER và MỘT SỐ MÔ HÌNH TRONG INFORMATION RETRIEVAL

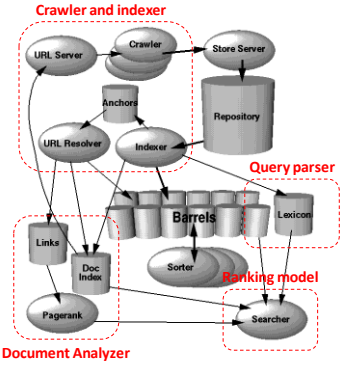
1

Recap: Information Retrieval – IR là gì?

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Many universities and public libraries use IR syst…

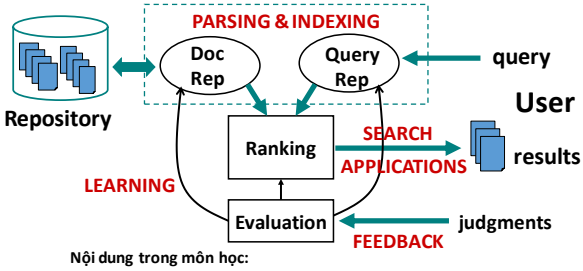
2

Recap: Kiến trúc hệ thống IR và Search



CS@UVa 3

Recap: Kiến trúc hệ thống IR và Search

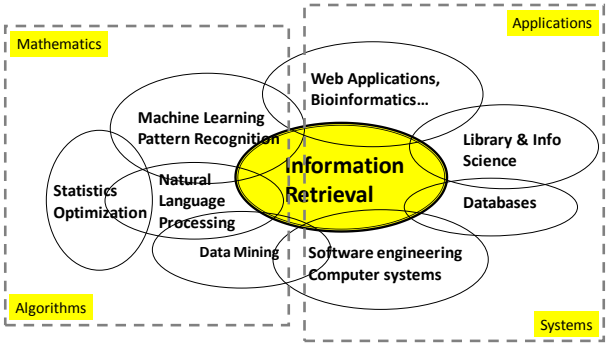


Nội dung trong môn học:

1) Search engine architecture; 2)Retrieval models;
3) Retrieval evaluation; 4) Relevance feedback;
5) Link analysis; 6) Search applications.

CS@UVa 4

Recap: Lĩnh vực IR

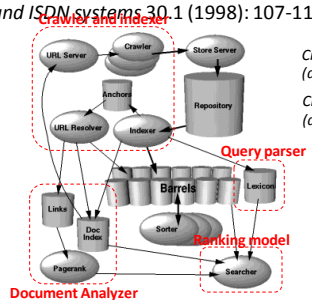


Nội dung

1. Search và các thành phần của IR
2. Một số mô hình trong IR
   - 2.1 Boolean model
   - 2.2 Vector space model
   - 2.3 Probabilistic model

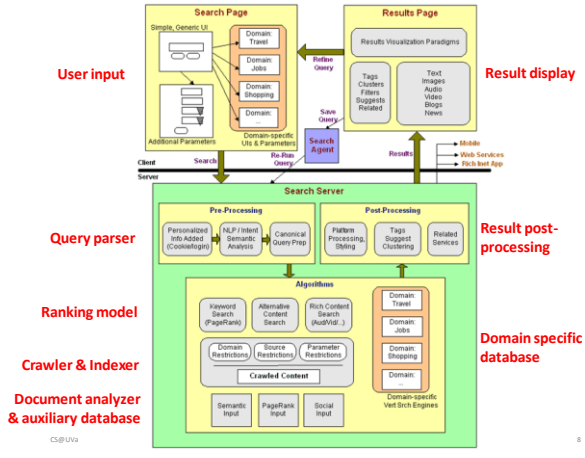
Kiến trúc của một Search engine

- ***"The Anatomy of a Large-Scale Hypertextual Web Search Engine"*** - *Sergey Brin and* Lawrence Page, *Computer networks and ISDN systems* 30.1 (1998): 107-117.

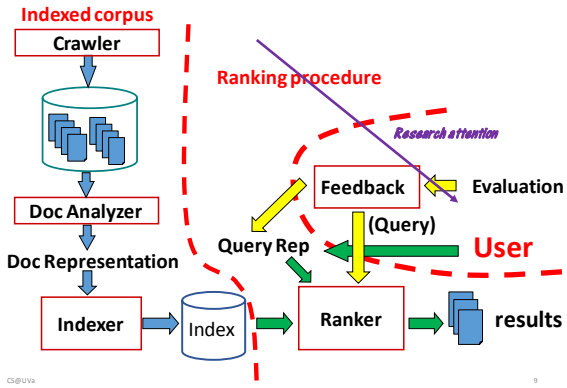


Citation count: 12197 (as of Aug 27, 2014)

Citation count: 13727 (as of Aug 30, 2015)



Luồng xử lý của Search Engine



Các thành phần cơ bản của IR

- Thông tin - **Information need**
  - "*an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need*" – wiki
  - Một hệ thống IR cố gắng "satisfy" một users' **information need**
- Câu truy vấn - Query
  - Một cách **biểu diễn** users' information need
  - Có nhiều cách: bằng ngôn ngữ tự nhiên, ….

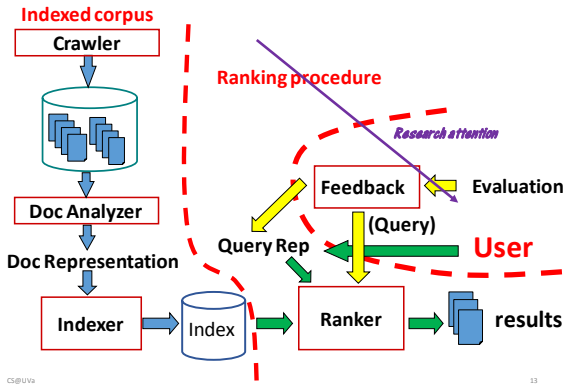
Các thành phần cơ bản của IR

- Dữ liệu - Document
  - Biểu diễn các thông tin có thể là câu trả lời cho user's information need
  - Dạn o, audi
- Liên qu
  - Sự l on need
  - Dựa trên nhiều khía cạnh: topical, semantic, temporal, spatial,…..

One sentence about IR - *"rank documents by their relevance to the information need"*

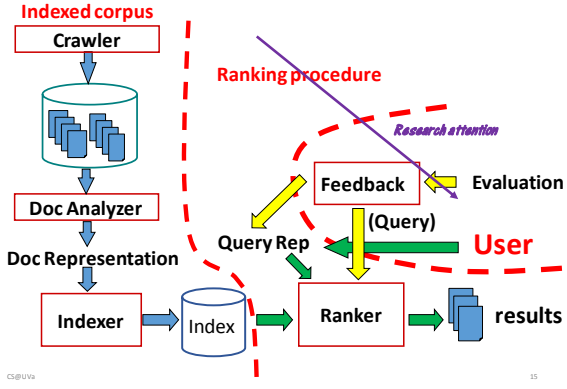
Một số thành phần của a search engine

- **Web crawler**
  - A **automatic program that systematically browses the web for the purpose of Web content** indexing and updating
- **Document analyzer & indexer**
  - **Manage the crawled web** content and provide **efficient access** of web documents



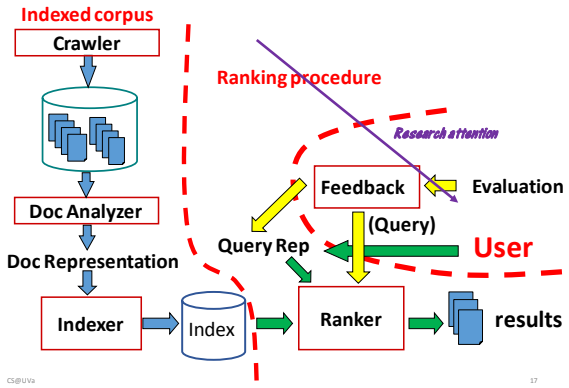
Một số thành phần của a search engine

- **Query parser**
  - **Compile user-input keyword queries** into managed system representation
- **Ranking model**
  - **Sort candidate documents according to it relevance** to the given query
- **Result display**
  - **Present the retrieved results** to users for satisfying their information need



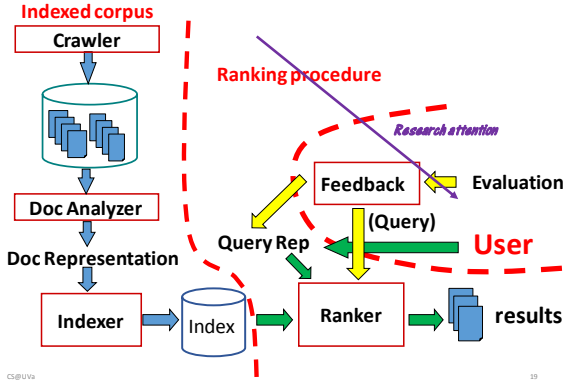
Một số thành phần của a search engine

- **Retrieval evaluation**
  - Assess the **quality** of the return results
- **Relevance feedback**
  - Propagate the **quality judgment back to the system** for search result refinement



Một số thành phần của search engine

- **Search query logs**
  - **Record users' interaction history** with search engine
- **User modeling**
  - Understand users' longitudinal information need
  - Assess users' satisfaction towards search engine output

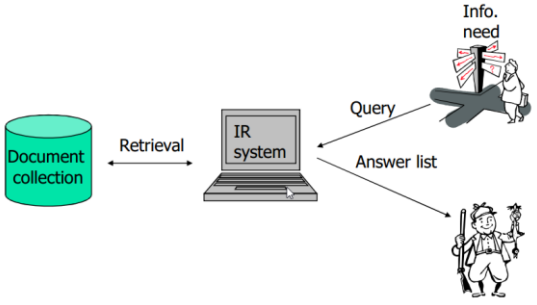


Browsing v.s. Querying

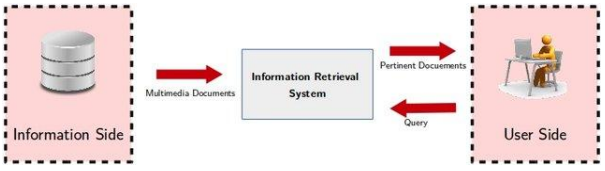
- Browsing – what Yahoo did before
  - The sys
  - with st
  - navigat
  - inform
  - enable
  - Works
  - to expl
  - know w
  - can't c
  - (e.g., w
- Querying – what Google does
  - eyword) query,
  - turns a set of
  - nts
  - the user knows
  - y to use for
  - ormation need

2. Hệ thống IR

- **Mục tiêu** = tìm tập tài liệu phù hợp từ tập rất lớn các tài liệu để có được thông tin thích hợp.

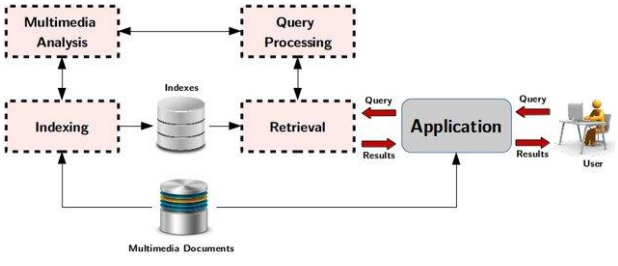


2. Hệ thống IR



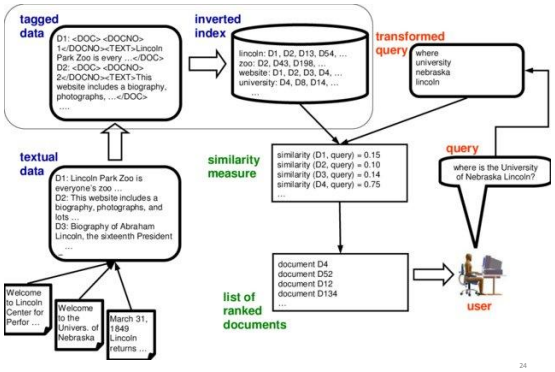
22

2. Hệ thống IR



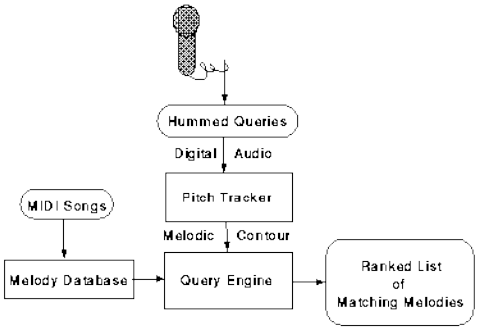
23

2. Text



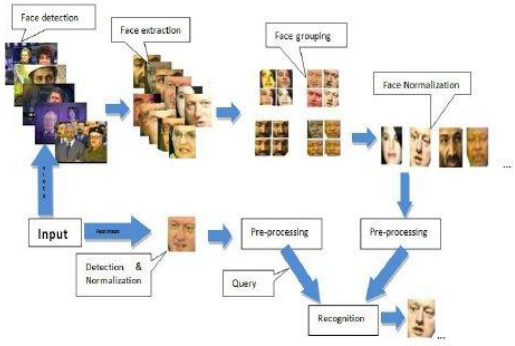
24

2. Audio



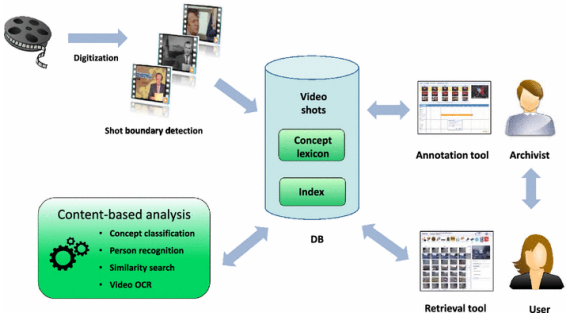
25

2. Face retrieval

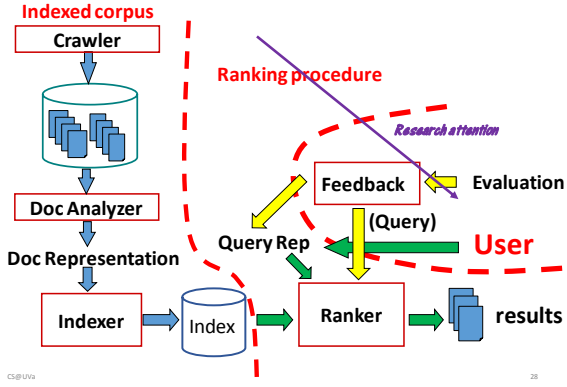


26

2. Video



27



CS@UVa 28

2. Crawler dữ liệu

Web Crawler - A automatic program that systematically browses the web for the purpose of Web content indexing and updating

- Synonyms: spider, robot, bot

2. Crawler - cách thức hoạt động

Mã giả:

```
Def Crawler(entry_point) {
    URL_list = [entry_point]
    while (len(URL_list)>0) {
        URL = URL_list.pop();
        if (isVisited(URL) or !isLegal(URL) or !checkRobotsTxt(URL))
            continue;
        HTML = URL.open();
        for (anchor in HTML.listOfAnchors()) {
            URL_list .append(anchor);
        }
        setVisited(URL);
        insertToIndex(HTML);
    }
}
```

Which page to visit next?

Is it visited already? Or shall we visit it again?

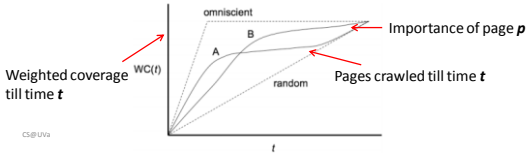
Is the access granted?

2. Crawler - Một số chiến thuật thu thập

- Duyệt theo chiều rộng - Breadth first
  - Uniformly explore from the entry page
  - Memorize all nodes on the previous level
  - As shown in pseudo code
- Duyệt theo chiều sâu - Depth first
  - Explore the web by branch
  - Biased crawling given the web is not a tree structure
- Duyệt theo chủ đề - ưu tiên - Focused crawling
  - Prioritize the new links by predefined strategies

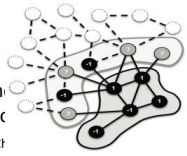
2. Crawler - Duyệt ưu tiên

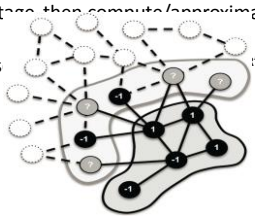
- Prioritize the visiting sequence of the web
  - The size of Web is too large for a crawler (even Google) to completely cover
  - Not all documents are equally important
  - Emphasize more on the high-quality documents
    - Maximize weighted coverage
      - In 1999, no search engine indexed more than 16% of the Web
      - In 2005, large-scale search engines index no more than 40-70% of the indexable Web



2. Crawler - Duyệt ưu tiên

- Prioritize by in-degree [Cho et al. WWW'98]
  - The page with the highest number of in
    from previously downloaded pages is d
- Prioritize by PageRank [Abiteboul et al. WWW'07, Cl
  - Breadth-first in early st
    PageRank periodically
  - More consistent with s





2. Crawler - Duyệt ưu tiên

- Prioritize by topical relevance
  - In vertical search, only crawl relevant pages [De et al. WWW'94]
    - E.g., restaurant search engine should only crawl restaurant pages
  - Estimate the similarity to current page by anchortext or text near anchor [Hersovici et al. WWW'98]
  - User given taxonomy or topical classifier [Chakrabarti et al. WWW'98]

2. Crawler - Tránh trùng lặp

- Given web is a graph rather than a tree, avoid loop in crawling is important
- What to check
  - URL: must be normalized, not necessarily can avoid all duplication
    - http://dl.acm.org/event.cfm?id=RE160&CFID=516168213&CFTOKEN=99036335
    - http://dl.acm.org/event.cfm?id=RE160
  - Page: minor change might cause misfire
    - Timestamp, data center ID change in HTML
- How to check
  - trie or hash table

2. Crawler - Một số quy định khi lấy thông tin

Crawlers can retrieve data much quicker and in greater depth than human searchers

- Costs of using Web crawlers
  - Network resources
  - Server overload
- Robots exclusion protocol
  - Examples: CNN, UVa

2. Crawler - Một số config của web

- Exclude specific directories:

```
User-agent: *
Disallow: /tmp/
Disallow: /cgi-bin/
Disallow: /users/paranoid/
```

- Exclude a specific robot:

```
User-agent: GoogleBot
Disallow: /
```

- Allow a specific robot:

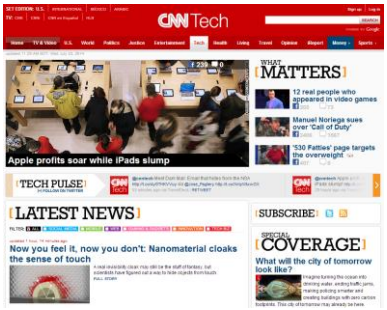
```
User-agent: GoogleBot
Disallow:

User-agent: *
Disallow: /
```


2. Crawler - Re-visit web

- The Web is very dynamic; by the time a Web crawler has finished its crawling, many events could have happened, including creations, updates and deletions
  - **Keep re-visiting the crawled pages**
  - **Maximize freshness and minimize age of documents** in the collection
- Strategy
  - Uniform re-visiting
  - Proportional re-visiting
    - Visiting frequency is proportional to the page's update frequency

2. Crawler - Cách phân tích một webpage

- What you care from the crawled web pages



2. Crawler - Cách phân tích một webpage

- What machine knows from the crawled web pages

2. Crawler - Cách phân tích một webpage

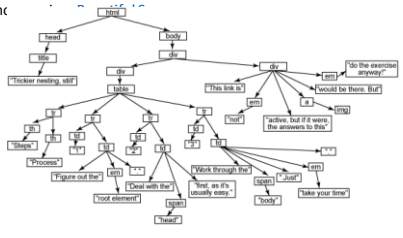
- Needs to analyze and index the crawled web pages
  - Extract informative content from HTML
  - Build machine accessible data representation

2. Crawler - HTML parsing

- Generally difficult due to the free style of HTML
- Solutions
  - Shallow parsing
    - Remove all HTML tags
    - Only keep text between <title></title> and
  - Automatic wrapper generation [Crescenzi et al. VLDB'01]
    - Wrapper: regular expression for HTML tags' combination
    - Inductive reasoning from examples
  - Visual parsing [Yang and Zhang DAR'01]
    - Frequent pattern mining of visually similar HTML blocks

2. Crawler - HTML parsing

- jsoup
  - Java-based HTML parser
    - scrape and parse HTML from a URL, file, or string to DOM tree
    - Find and extract data, using DOM traversal or CSS selectors
      - children(), parent(), siblingElements()
      - getElementsByClass(), getElementsByAttributeValue()
  - Pytho



2. Crawler - Biểu diễn thông tin tài liệu

- Represent by a string?
  - No semantic meaning
- Represent by a list of sentences?
  - Sentence is just like a short document (recursive definition)
- <HEAD>Crowds in Liverpool to Mark 10th Anniversary of John Lennon's Death</HEAD>
  <DATELINE>LIVERPOOL, England (AP) </DATELINE>
  <TEXT>
  Dozens of fans of rock legend and former Beatle John Lennon gathered in the snow on a windy Saturday for a ceremony marking the 10th anniversary of his death. Liverpool's mayor, Dorothy Gavin, led Lennon devotees who laid wreaths at the foot of a bronze statue of The Beatles in the city's Cavern Walks shopping center. The center was built on the original site of the Cavern Club, made famous when The Beatles played there in the 1960s, and has become a place of pilgrimage. ``Give peace a chance,'' the title of one of singer-songwriter Lennon's greatest hits, was the theme for the day.
  ...
  Lennon and his wife, Yoko Ono, were returning to their apartment in New York's Dakota apartment building after a recording session on Dec. 8, 1980, when Lennon was shot to death by Mark David Chapman, a deranged fan to whom Lennon had given his autograph only hours before. Lennon was 40. A spokesman for the Lennon family said Ms. Ono and the couple's son, Sean, were in Europe and would spend the anniversary privately.
  ...
  Peebles said late in 1980 that Lennon had just recovered from a period when he had ``gone off the rails'' and his relationship with Ms. Ono had suffered. ``But (when I saw him) they'd had the baby, Sean had been born, and everything was great.''
  </TEXT>

2. Crawler - Biểu diễn thông tin tài liệu

Tách từ - Tokenization

- Break a stream of text into meaningful units
  - Tokens: words, phrases, symbols
    - **Input:** It's not straight-forward to perform so-called "tokenization."
    - **Output(1):** 'It's', 'not', 'straight-forward', 'to', 'perform', 'so-called', '"tokenization."'
    - **Output(2):** 'It', "'", 's', 'not', 'straight', '-', 'forward, 'to', 'perform', 'so', '-', 'called', "", 'tokenization', '.', '"'
  - Definition depends on language, corpus, or even context

2. Crawler - Biểu diễn thông tin tài liệu

Giải pháp Tách từ - Tokenization

- Regular expression
  - [\w]+: so-called -> 'so', 'called'
  - [\S]+: It's -> 'It's' instead of 'It', "s'
- Statistical methods
  - Explore rich features to decide where is the boundary of a word
    - Apache OpenNLP (http://opennlp.apache.org/)
    - Stanford NLP Parser (http://nlp.stanford.edu/software/lex-parser.shtml)
  - Online Demo
    - Stanford (http://nlp.stanford.edu:8080/parser/index.jsp)
    - UIUC (http://cogcomp.cs.illinois.edu/curator/demo/index.html)

2. Crawler - Biểu diễn thông tin tài liệu

- Bag-of-Words representation
  - Doc1: Information retrieval is helpful for everyone.
  - Doc2: Helpful information is retrieved for you.

	information	retrieval	retrieved	is	helpful	for	you	everyone
Doc1	1	1	0	1	1	1	0	1
Doc2	1	0	1	1	1	1	1	0

Word-document adjacency matrix

2. Crawler - Biểu diễn thông tin tài liệu

- Bag-of-Words representation
  - Assumption: word is independent from each other
  - Pros: simple
  - Cons: grammar and order are missing
  - ***The most frequently used document representation***
    - ***Image, speech, gene sequence***

2. Crawler - Biểu diễn thông tin tài liệu

- Improved Bag-of-Words representation
  - N-grams: a contiguous sequence of n items from a given sequence of text
    - E.g., Information retrieval is helpful for everyone
    - Bigrams: 'information_retrieval', 'retrieval_is', 'is_helpful', 'helpful_for', 'for_everyone'
  - Pros: capture local dependency and order
  - Cons: purely statistical view, increase vocabulary size $O(V^N)$

2. Crawler - Biểu diễn thông tin tài liệu

- Index document with all the occurring word
  - Pros
    - Preserve all information in the text (hopefully)
    - Fully automatic
  - Cons
    - Vocabulary gap: cars v.s., car
    - Large storage: e.g., in N-grams $O(V^N)$
  - Solution
    - Construct controlled vocabulary

2. Crawler - Biểu diễn thông tin tài liệu

Chuẩn hóa dữ liệu: Normalization

- Convert different forms of a word to normalized form in the vocabulary
  - U.S.A -> USA, St. Louis -> Saint Louis
- Solution
  - Rule-based
    - Delete periods and hyphens
    - All in lower case
  - Dictionary-based
    - Construct equivalent class
      - Car -> "automobile, vehicle"
      - Mobile phone -> "cellphone"

2. Crawler - Biểu diễn thông tin tài liệu

Stemming

- Reduce inflected or derived words to their root form
  - Plurals, adverbs, inflected word forms
    - E.g., ladies -> lady, referring -> refer, forgotten -> forget
  - Bridge the vocabulary gap
  - Risk: lose precise meaning of the word
    - E.g., lay -> lie (a false statement? or be in a horizontal position?)
  - Solutions (for English)
    - Porter stemmer: pattern of vowel-consonant sequence
    - Krovetz Stemmer: morphological rules

2. Crawler - Biểu diễn thông tin tài liệu

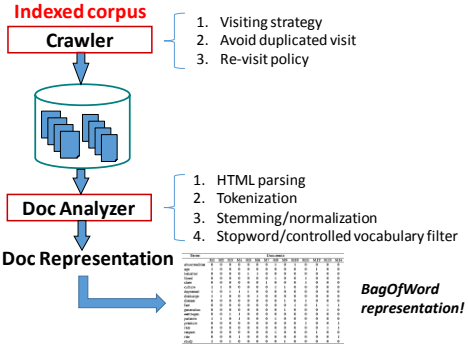
Stopwords

- Us
  -
  -
  -
  -

Nouns	Verbs	Adjectives	Prepositions	Others
1. time	1. be	1. good	1. to	1. the
2. person	2. have	2. new	2. of	2. and
3. year	3. do	3. first	3. in	3. a
4. way	4. say	4. last	4. for	4. that
5. day	5. get	5. long	5. on	5. I
6. thing	6. make	6. great	6. with	6. it
7. man	7. go	7. little	7. at	7. not
8. world	8. know	8. own	8. by	8. he
9. life	9. take	9. other	9. from	9. as
10. hand	10. see	10. old	10. up	10. you
11. part	11. come	11. right	11. about	11. this
12. child	12. think	12. big	12. into	12. but
13. eye	13. look	13. high	13. over	13. his
14. woman	14. want	14. different	14. after	14. they
15. place	15. give	15. small	15. beneath	15. her
16. work	16. use	16. large	16. under	16. she
17. week	17. find	17. next	17. above	17. or
18. case	18. tell	18. early		18. an
19. point	19. ask	19. young		19. will
20. government	20. work	20. important		20. my
21. company	21. seem	21. few		21. one
22. number	22. feel	22. public		22. all
23. group	23. try	23. bad		23. would
24. problem	24. leave	24. same		24. there
25. fact	25. call	25. able		25. their

The OEC: Facts about the language

Abstraction of search engine architecture



Automatic text indexing

Query: "to be or not to be"

- In modern search engine
  - **No** stemming or stopword removal, since computation and storage are no longer the major concern
  - More advanced NLP techniques are applied
    - Named entity recognition
      - E.g., people, location and organization
    - Dependency parsing