

平成24年度学術ポータル担当者研修 2012年8月1日/22日

# 学術情報流通を実現する技術(1)

--要素技術(検索、DB等の基盤技術に  
特化した話を中心に)--

国立教育政策研究所  
教育研究情報センター

総括研究官

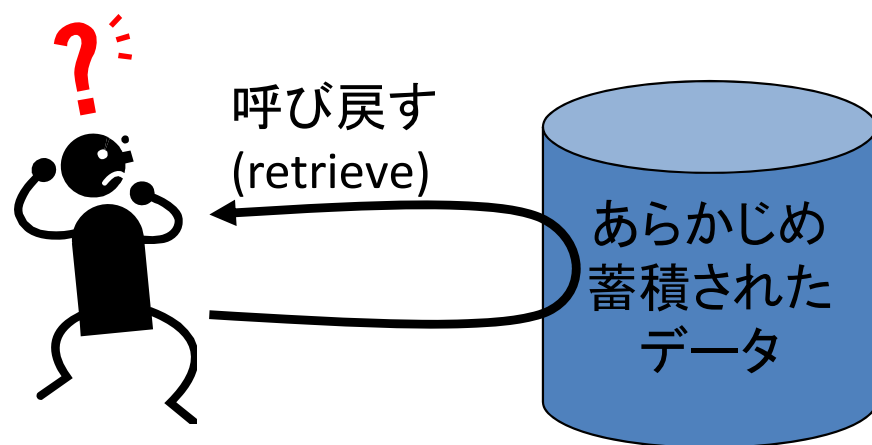
江草由佳

Twitter: @yegusa

# 情報検索とは（由来）

## ● 情報検索

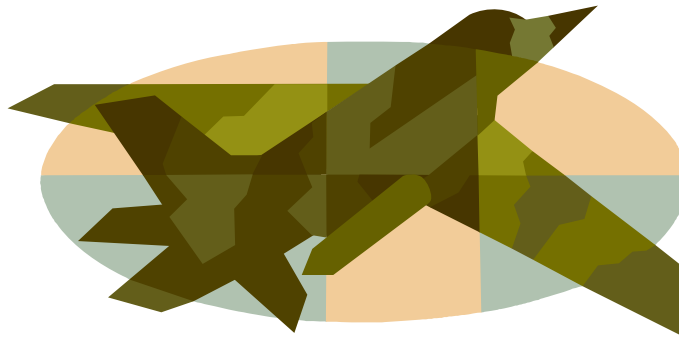
- IR: information (storage and ) retrieval
  - 情報(information) を呼び戻すこと(retrieval)
  - 元はinformation storage and retrieval 情報の蓄積と検索
- 1950年にムーアーズ(Calvin N. Mooers)が初めて定義
- 1960年代に広く使われるようになる
- (search: これも「検索」と訳すが。。。)



retriever(レトリバー):  
獲物をくわえて戻って  
くるように訓練された猟犬

# データベースの起源

- 1950年代
- 米国国防総省が戦力に関する**情報を保管、集中管理**するためコンピュータを使ったライブラリーを開発
- **データの基地**(data base)から由来



# データベースの定義(1)

- 著作権法二条十の三
  - 論文、数値、図形その他の情報の集合物であって、それらの情報を電子計算機を用いて検索することができるように体系的に構成したもの
- 日本工業規格(JIS)
  - 適用業務分野で使用するデータの集まりであって、データの特性とそれに対応する実態の間の関係とを記述した概念的な構造によって編成されたもの(X0017)
  - 特定の規則に従って電子的な形式で、一か所に蓄積されたデータの集合であって、コンピュータでアクセス可能なもの(X0807)

# データベースの定義(2)

## --データベースの特徴--

- コンピュータを用いて検索できることが重要。情報が電子メディアに蓄積され、コンピュータを使用して検索できる状態になっている。
- データや情報がコンピュータ処理できるように体系的に整理され、統合化・構造化されて蓄積・保存されており、必要な情報だけを部分的に取り出せる。
- 蓄積情報の検索や更新が容易に行えるよう、効率化を図ったものである。

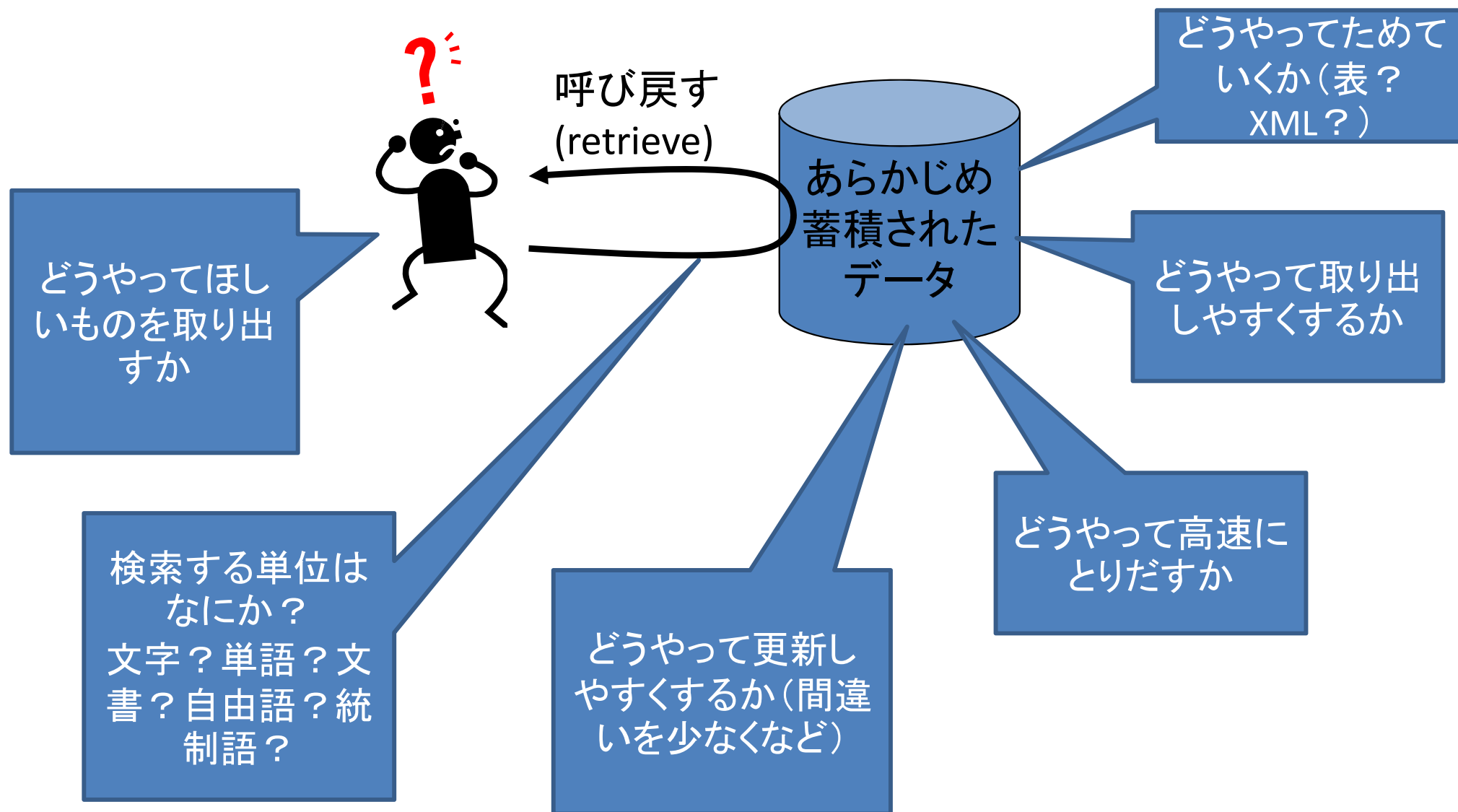
※ ちなみに、ヨーロッパにおけるデータベースの定義では、コンピュータを使用するかしないか、電子的であるかどうかについては特に限定していない

# (要素技術の話の前に)

## 前提：計算機は意外と・・・おぼか

- 計算機(コンピューター、プログラム、システム)は魔法じゃない
- 計算機が得意なこと→(単純な)同じ作業の繰り返しを正確に・人間より速く・必ずやってくれる
  - 途中で投げ出したり、飽きたり、疲れたりしない(ただし、電気は食う)
- 計算機はかなり細かいレベルで手順まで含めて命令しないと動かない
  - 適当にやっなど命令はできない
  - なにか「適当に」やっているように見えることは裏で細かいレベルまでの命令を人間が仕込んでいる
  - 逆に言うと、かなり細かいレベルまでの命令を人間がかんがえつければ、「適当な」動きを計算機を使って実現はできない

# 検索にまつわる様々な観点(一部)



# どうやってためていくか？表？XML？

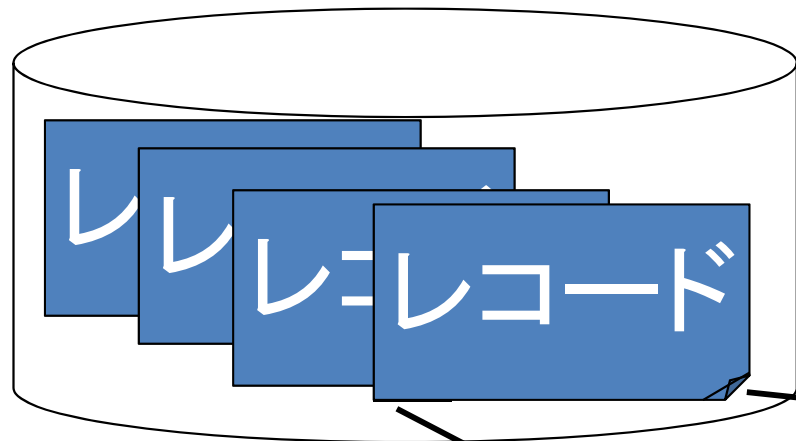
## -- データベースの種類 --

- リレーショナルデータベース (RDB)
  - 表としてデータを扱う
- オブジェクトデータベース
  - オブジェクトとしてデータを扱う
- XMLデータベース
  - XMLとしてデータを扱う
- 全文データベース
  - いろいろある、なんでもRDBにすればよいというものではない
  - 書誌データ、全文データはRDBには向いていない



# どうやって取り出しやすくするか？

## --レコードと検索フィールド--



検索フィールド名

検索フィールド値

検索フィールド

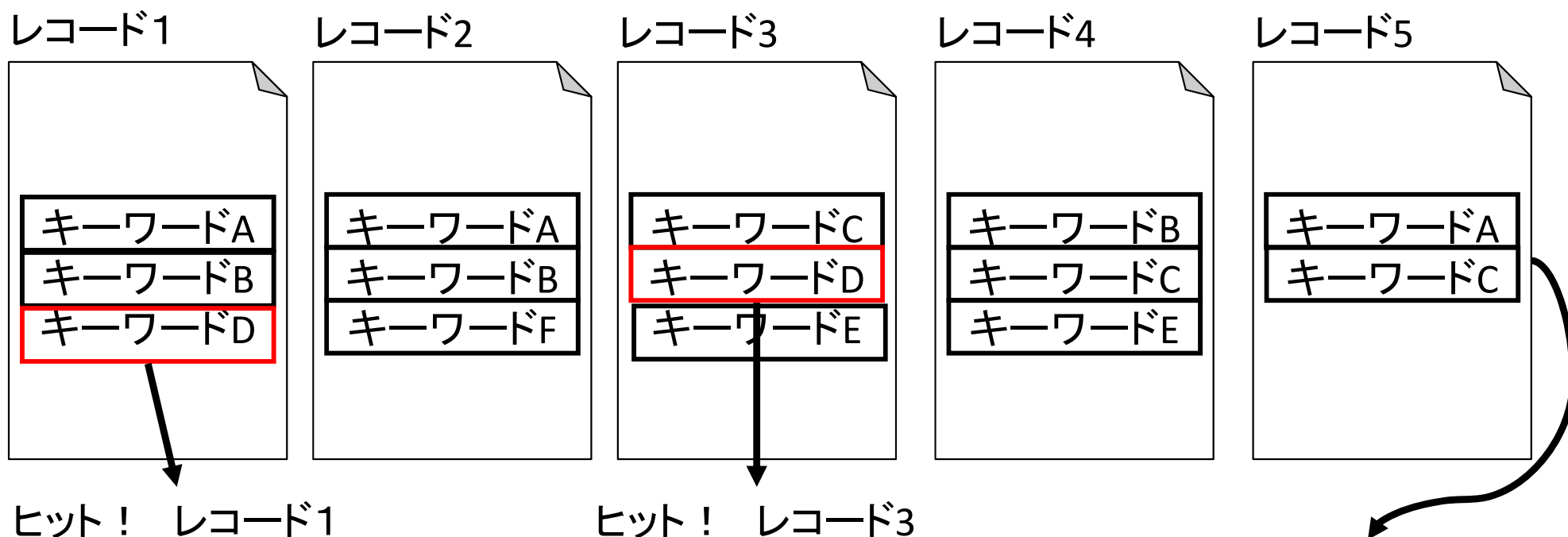
(仮想的な)  
検索フィールド

|        |                          |
|--------|--------------------------|
| 論題:    | Reading—速読・多読<br>について考える |
| 著者名:   | 清水由理子                    |
| 請求記号:  | P343-5C2-14              |
| 掲載誌名:  | 獨協大学外国語教育研究14            |
| 発行年月:  | 1995.12                  |
| 掲載ページ: | p.273～282                |
| 登録日:   | 19970930                 |
| 本文:    | (682CAD-11.pdf)          |

# どうやって高速にとりだすか(1)

--なんにも仕掛けがないと。。。--

- レコードを最初から最後まで順番に検索
- レコードが多くなると時間がかかってしまう



例:「キーワードD」を検索する

ヒットした件数は2件、  
ヒットしたレコードは  
レコード1とレコード3

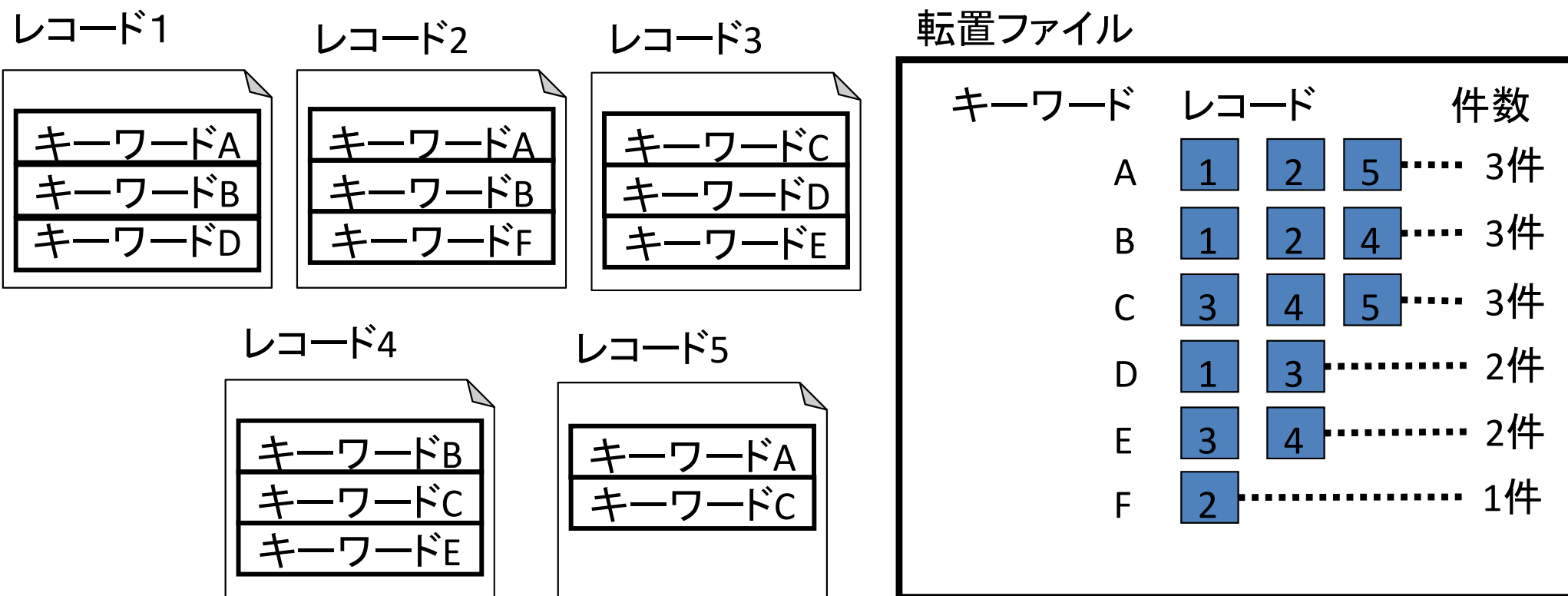
# どうやって高速にとりだすか(2)

## -- インデックス を用意する --

- 高速にデータにアクセスするために必要
- インデックス方式
  - 転置ファイル (Inverted file; インバーテッドファイル)
    - もっともよく使われる方式
  - TRI木
  - Suffix Array
  - …いろいろある

# 転置ファイル(1)

- 検索キーごとにレコードを集計したファイルを作成する方法
- Inverted file(転置ファイル、倒置ファイル)ともいう



# 転置ファイル(2)

- 例:「キーワードD」を検索する
- 欠点:あらかじめ転置ファイルを作らなければならない、ファイルの容量増

レコード1

|        |
|--------|
| キーワードA |
| キーワードB |
| キーワードD |

レコード2

|        |
|--------|
| キーワードA |
| キーワードB |
| キーワードF |

レコード3

|        |
|--------|
| キーワードC |
| キーワードD |
| キーワードE |

レコード4

|        |
|--------|
| キーワードB |
| キーワードC |
| キーワードE |

レコード5

|        |
|--------|
| キーワードA |
| キーワードC |

転置ファイル

| キーワード | レコード  | 件数 |
|-------|-------|----|
| A     | 1 2 5 | 3件 |
| B     | 1 2 4 | 3件 |
| C     | 3 4 5 | 3件 |
| D     | 1 3   | 2件 |
| E     | 3 4   | 2件 |
| F     | 2     | 1件 |

# 転置ファイル (3)

## レコード例(1)

http://nyti.ms/po6J1Z  
 うーむ  
 Aaron Swartz  
 が  
 昨年  
 9  
 月  
 の  
 JSTOR  
 から  
 の  
 大量  
 ダウンロード  
 容疑  
 で  
 逮捕  
 ・  
 起訴  
 とは  
 絶句  
 #librahack

## レコード例(2)

要は  
 ライブラリアン  
 に  
 必要  
 な  
 の  
 は  
 、  
 国内  
 文学  
 読ん  
 だり  
 、  
 月  
 9  
 の  
 ドラマ  
 を  
 見  
 たり



ひとまとめにして、  
 辞書順に並べ替え

#librahack  
 9  
 Aaron  
 JSTOR  
 Swartz  
 http://nyti.ms/po6J1Z  
 、  
 。  
 ・  
 いい  
 から  
 が  
 こと  
 し  
 じゃ  
 すぎ  
 する  
 た  
 たり  
 だり  
 て  
 で

|     |         |     |     |   |
|-----|---------|-----|-----|---|
| 1   | とは      | 1   | 昨年  | 1 |
| 1,2 | な       | 2   | 絶句  | 1 |
| 1   | なあ      | 2   | 専門  | 2 |
| 1   | ない      | 2   | 存在  | 2 |
| 1   | なく      | 2   | 逮捕  | 1 |
| 1   | なら      | 2   | 大量  | 1 |
| 1,2 | に       | 2   | 読ん  | 2 |
| 2   | の       | 1,2 | 必要  | 2 |
| 1   | は       | 2   | 文学  | 2 |
| 2   | ば       | 2   | 勉強  | 2 |
| 1   | ほう      | 2   | 本音  | 2 |
| 2   | みんな     | 2   | 名乗る | 2 |
| 2   | もっと     | 2   | 遊び  | 2 |
| 1   | を       | 2   | 容疑  | 1 |
| 1,2 | ん       | 2   | 要は  | 2 |
| 2   | ダウンロード  | 1   |     |   |
| 2   | ドラマ     | 2   |     |   |
| 2   | ライブラリアン | 2   |     |   |
| 2   | 悪い      | 2   |     |   |
| 2   | 家       | 2   |     |   |
| 2   | 起訴      | 1   |     |   |
| 2   | 月       | 1,2 |     |   |
| 2   | 見       | 2   |     |   |
| 2   | 言え      | 2   |     |   |
| 1   | 国内      | 2   |     |   |

# 転置ファイル (4)

|                       |     |         |     |     |   |
|-----------------------|-----|---------|-----|-----|---|
| #librahack            | 1   | とは      | 1   | 昨年  | 1 |
| 9                     | 1,2 | な       | 2   | 絶句  | 1 |
| Aaron                 | 1   | なあ      | 2   | 専門  | 2 |
| JSTOR                 | 1   | ない      | 2   | 存在  | 2 |
| Swartz                | 1   | なく      | 2   | 逮捕  | 1 |
| http://nyti.ms/po6J1Z | 1   | なら      | 2   | 大量  | 1 |
| 、                     | 1,2 | に       | 2   | 読ん  | 2 |
| 。                     | 2   | の       | 1,2 | 必要  | 2 |
| ・                     | 1   | は       | 2   | 文学  | 2 |
| いい                    | 2   | ば       | 2   | 勉強  | 2 |
| うーむ                   | 1   | ほう      | 2   | 本音  | 2 |
| か                     | 2   | みんな     | 2   | 名乗る | 2 |
| かけ                    | 2   | もっと     | 2   | 遊び  | 2 |
| から                    | 1   | を       | 2   | 容疑  | 1 |
| が                     | 1,2 | ん       | 2   | 要は  | 2 |
| こと                    | 2   | ダウンロード  | 1   |     |   |
| し                     | 2   | ドラマ     | 2   |     |   |
| じゃ                    | 2   | ライブラリアン | 2   |     |   |
| すぎ                    | 2   | 悪い      | 2   |     |   |
| する                    | 2   | 家       | 2   |     |   |
| た                     | 2   | 起訴      | 1   |     |   |
| たり                    | 2   | 月       | 1,2 |     |   |
| だり                    | 2   | 見       | 2   |     |   |
| て                     | 2   | 言え      | 2   |     |   |
| で                     | 1   | 国内      | 2   |     |   |

- 検索する際には、ひとまとめにした単語リストに対して照合する。
- 単語リストの長さを  $T$  とした場合、平均で  $\log_2(T)$  回の照合で検索できる(=2分探索)

⇒ 対数をとるので、飛躍的な速度が実現できる。

⇒ 100語で 平均3.3回

⇒ 100万語で平均20回

「ダウンロード」で検索する場合

# 転置ファイル(5)

--どうやってキーワードを取り出す？--

レコード例(1): 「<http://nyti.ms/po6J1Z> うーむ、Aaron Swartzが昨年9月のJSTORからの大量ダウンロード容疑で逮捕・起訴とは、絶句。  
#librahack」

<http://twitter.com/tmasao/status/93607169886396416>

- <http://nyti.ms/po6J1Z>
- うーむ
- Aaron
- Swa
- が
- 昨年
- 9
- 月
- の
- JSTOR
- から
- の
- 大量
- 逮捕
- ・
- 起訴
- とは
- 絶句
- #librahack

意味のある語単位で切る＝形態素解析

(日本語の文章の意味(かかり受けや品詞など)を解析して意味のある語で区切るようにする)



# 転置ファイル(6)

--どうやってキーワードを取り出す？--

レコード例(1): 「<http://nyti.ms/po6J1Z> うーむ、Aaron Swartzが昨年9月のJSTORからの大量ダウンロード容疑で逮捕・起訴とは、絶句。  
#librahack」

<http://twitter.com/tmasao/status/93607169886396416>

- |   |         |      |      |
|---|---------|------|------|
| • <a href="http://nyti.ms/po6J1Z">http://nyti.ms/po6J1Z</a> | • JSTOR | • ロー | • 起訴 |
| • うー  | • から    | • ード | • 訴と |
| • ーむ  | • らの    | • ド容 | • とは |
| • Aaron   | • の大    | • 容疑 | • は絶 |
| • Swartz  |         |      | • 絶句 |
- 
- |      |   |  |  |
|------|---|--|--|
| • が昨 | <div data-bbox="421 1098 1966 1353" data-label="Text"> <p>N-gram: 強引に、2文字ずつなどで強制的に切る方法</p> </div> |  |  |
| • 年9 |   |  |  |
| • 9月 |   |  |  |
| • 月の |   |  |  |
- |      |      |
|------|------|
| • ウン | • 捕・ |
| • ンロ | • ・起 |

# 転置ファイル(7)

レコード例(2): 「要はライブラリアンに必要なのは、国内文学読んだり、月9のドラマを見たりするんじゃないくて、存在をかけて勉強することじゃないかなあ。本音を言えば、悪いが、みんな遊びすぎ。専門家を名乗るなら、もっと勉強したほうがいい。」

<http://twitter.com/arg/status/20007995359>

## 形態素解析

- |           |       |      |      |       |       |      |
|-----------|-------|------|------|-------|-------|------|
| • 要は      | • 読ん  | • する | • 勉強 | • 言え  | • 専門  | • ほう |
| • ライブラリアン | • だり  | • ん  | • する | • ば   | • 家   | • が  |
| • に       | • 、   | • じゃ | • こと | • 、   | • を   | • いい |
| • 必要      | • 月   | • なく | • じゃ | • 悪い  | • 名乗る | • 。  |
| • な       | • 9   | • て  | • ない | • が   | • なら  |      |
| • の       | • の   | • 、  | • か  | • 、   | • 、   |      |
| • は       | • ドラマ | • 存在 | • なあ | • みんな | • もっと |      |
| • 、       | • を   | • を  | • 。  | • 遊び  | • 勉強  |      |
| • 国内      | • 見   | • かけ | • 本音 | • すぎ  | • し   |      |
| • 文学      | • たり  | • て  | • を  | • 。   | • た   |      |

# 転置ファイル(8)

## --どうやって検索する？--

- レコード例:
  - (1) 「<http://nyti.ms/po6J1Z> うーむ、Aaron Swartzが昨年9月のJSTORからの大量ダウンロード容疑で逮捕・起訴とは、絶句。  
#librahack」
  - (2) 「要はライブラリアンに必要なのは、国内文学読んだり、月9のドラマを見たりするんじゃないかと、存在をかけて勉強することじゃないかなあ。本音を言えば、悪いが、みんな遊びすぎ。専門家を名乗るなら、もっと勉強したほうがいい。」
- Q: [Swartz] -> [Swartz] -> Hit! (1)
- Q: [専門家] -> [専門][家] -> Hit! (2)
- Q: [9月] -> [9][月] -> Hit? (1) (2)
- Q: [#librahack] -> Hit???
- Q: [アーロン・シュワルツ] -> No Hit?

インデックスを使った  
検索(仕組み/人)  
のところで、  
それぞれ工夫必要

# まとめ

- 情報検索～データベースとは？
- 検索にまつわる観点
- レコードの蓄積からインデックスの構築、検索まで
- 転置ファイルを例に
  
- 参考文献：
  - 情報アクセスの新たな展開－情報検索・利用の最新動向. 日本図書館情報学会編. 勉誠出版, 2009, 216p.  
ISBN: 978-4-585-00278-9
  - Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition). Ricardo Baeza-Yates, Berthier Ribeiro-Neto. ACM Press Books, 2011, 944p. ISBN: 978-0321416919