# 検索の仕組み
# 索引検索

2015
時実 象一
tokizane@aichi-u.ac.jp

## 検索の仕組み

- 検索方法の種類
  - 逐次検索 (Sequential Search)
  - 索引検索 (Index Search)



## 逐次検索

- 文書 (レコード) を先頭から順に見て、検索語との一致を探す
- 例
  - 本を最初のページから調べる
  - ブラウザのページ内の検索
  - Word, PDF の文章内の検索
- 簡単だがデータの量が多いと時間がかかる



第一章　新人賞を射止めるには

そのジャンルの王道を行くオーソドックスで面白い物語は、「新人賞を受賞してプロ作家としてデビューして以降に一般読者向けに刊行すれば、それはヒットしてファンも付いてくれる可能性があります。

この辺りが、選者と一般読者の感性が大きく異なる点の一つと言えるでしょうね。

**（定理２）選者を感心させるには、選者の知らないことを書くべし）**

**（定理３）いくら面白い物語を作っても、選者に既視感を抱かせたら落とされる）**

**【高橋】**　そういう意味では、ものすごく見当違いの努力を一生懸命にできた気がします。受講前は、自分が読みたいものを「一般読者に向けて」書けばいいと思っていました。つまり、先鋭的すぎる知識を避けて、より一般的な方向へと、受賞から遠ざかる努力を一生懸命にしてきた、というわけです。

**【若桜木】**　なるほど、きもちありなん、ですね。小松左京賞の努力賞を受賞された高橋さんですから、そうなんですから、他のアマチュアの書き手は、推して知るべし、ですね。

**【高橋】**　逆に、受賞した八年前は文章もすごく下手くそで、それから何作も書いて、いろんな意味で上手くなったはずなのに、不思議なことに書けば書くほど公募の成績が悪くなっていった



んです。先生の講座を受講するようになって、その理由がやっと分かりました。

もちろん、「分かる」からすぐ「できる」わけじゃないので、いまだに既視感のない作品を作れないでいるわけなんですけれども。

**私が新人賞を受賞できた理由**

**【若桜木】**　ライトノベル新人賞の応募作など、既視感どころか、単にそれまでの既存作の「良いとこ盗り」の寄せ集めで、パロディにしか思えない作品も非常に多い、「こんなの俺にも書ける・私にも書ける」という調子で寄せ集めるんでしょうが。

とにかく、パロディは基本的に面白さを目指しますから、パロディの新人賞受賞作となると、第八回日本ファンタジーノベル大賞優秀賞の「月便」くらいのものでしょう。

この作品は「ロビンソン・クルーソー」（ダニエル・デフォー作）のパロディで、主人公はダニエル・デフォーとフライデー、主要登場人物として、ロビンソン・クルーソーのモデルとされるスコットランドの航海長アレキサンダー・セルカークまで出てくるという念の入れようで、そこの厳密に選者は「自分の知らないことに感心」する傾向が顕著なだけですが、



## 逐次検索の例

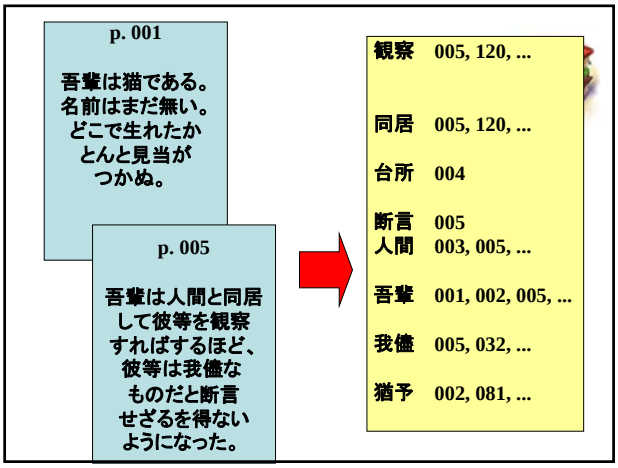
## 索引検索

- あらかじめキーワードの索引を作っておく
- 例
  - 本の巻末の索引を調べる
  - OPAC
  - Google などの検索エンジン
- データ量が多くても検索時間は変わらない

7

## 本の






## 索引検索の例

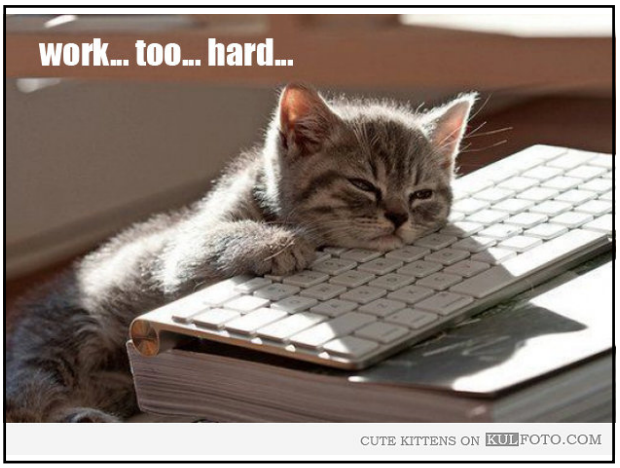

10

## 索引検索の例


## 索引ファイル

- 索引語 (キーワード) の抽出
  - 対象レコード (文書、ページ) から検索の手がかり (語、句) を取り出す
  - レコード (ページ) 番号をつける
  - 索引語を 50 音順、ABC 順に並べる
- 検索
  - 語句が一致するものを探す
  - 該当レコード (ページ) を取り出す

12




## 索引語の抽出

- 昔は人手で抽出
- 今はコンピュータが自動抽出
  - 英語は単語がスペースで区切られているので簡単
  - 日本語は単語の区切りがない
    - 形態素解析法
    - バイグラム法

## 索引語の抽出 (英語)

- 単語がスペースで区切られている

資料番号: 201833
標題: Information Retrieval in Chemistry and Chemical Patent Law

「Information」「Retrieval」「in」「Chemistry」「and」
「Chemical」「Patent」「Law」



## 索引語の抽出 (日本語)

- 形態素解析法
  - 文法解析と辞書を使って単語を区切る

資料番号: 98170
標題: インターネット時代の化学文献とデータベースの活用法

  - まず助詞などで語句を切り出す
    「インターネット時代」「化学文献」「データベース」「活用法」
  - カタカナの部分は単語として切り出す
    - 「インターネット」「時代」
  - 次に語句から辞書を使って単語を取り出す
    「化学文献」→「化学」「文献」



## 索引語の抽出 (日本語)

- バイグラム (bigram) 法
  - 1 文字ずつずらしながら、2 文字のペアを機械的に索引

資料番号: 98170
標題: インターネット時代の化学文献とデータベースの活用法

「イン」「ンタ」「ター」「ーネ」「ネッ」「ット」
「ト時」「時代」「代化」「化学」「学文」「文献」
「献と」「とデ」「デー」「ータ」「タベ」「ベー」
「ース」「スの」「の活」「活用」「用法」



## 形態素解析とバイグラム

- 形態素解析
  - 検索結果にノイズが少ない
  - 切り出しのための辞書の整備が必要
  - 検索エンジンなどで用いられる
    - Google, Yahoo!

## 形態素解析とバイグラム

- バイグラム
  - システムが簡単・安価
  - 辞書がいらない
  - 断片的な単語で検索できる
  - ノイズがある
    - 「京都」を検索しようとすると「東京都」もヒットする
    - 「タイ」を検索しようとすると「スタイル」もヒットする
  - 社内文書など小規模なシステムでよく用いられる
    - OPAC, NDL-OPAC、新聞記事データベース

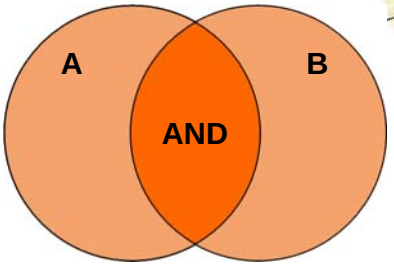
19


## ブール演算

- 論理演算、集合演算ともいう
- 集合間の
  - 積　　AND
  - 和　　OR
  - 差　　NOT
  - を求める
- 情報検索で用いる

21

## AND 検索

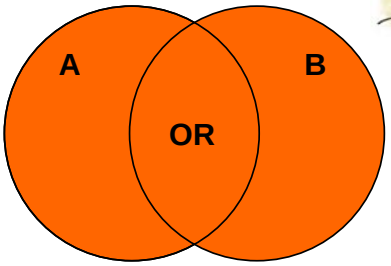



22

## Google の検索オプション


## Google の AND

## OR 検索

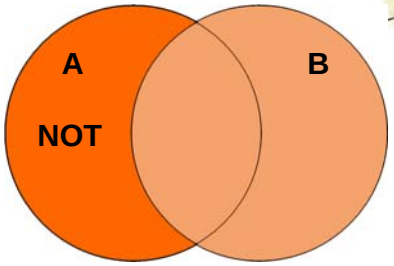



25

## Google の OR


26

## NOT 検索




27

## Google の NOT


## 検索の仕組み (AND)

資料番号: 98170
標題: インターネット時代の化学文献とデータベースの活用法

| 索引語 | 件数 | レコードリスト |
|---|---|---|
| | | |
| インターナル | 253 | ... 50405, 52183, 60031, 64521, ... |
| インターネット | 3862 | ... 50937, 64742, 68451, 69227, 69228, 85258, 89966, 92454, 94755, 98170, 103220 ... |
| インターン | 28 | ... 67940, 54697, 54336, ... |
| | | |
| | | |
| データ | 3301 | ... 84161, 84594, 85258, 85850, 90102, 93397, 95151, 89150, 98168, 98170, 98175, 98572, 101090, ... |
| データベース | 485 | ... 53219, 54083, 77305, 83210, 98170, 100351, 123087, ... |
| データマイニング | 71 | ... 83943, 88541, 88630, 100455, ... |
| | | |

29

## 近接演算

- 語と語のあいだの距離を調べて検索する
  - 距離が近いほど適合性が高い
- 通常、2 つの語が続いていることを指定
  - Google では「完全一致」と呼ぶ

## Google の近接演算 (フレーズ)


## 検索の仕組み (近接)

- 単語間の距離をはかって、適合性を見る

資料番号: 98170
標題: インターネット時代の化学文献とデータベースの活用法
1 2 3 4 5 6

| 索引語 | 件数 | レコードリスト |
|---|---|---|
| | | |
| インターナル | 253 | ... 50405-7, 52183-12, 60031-3, 64521-5, ... |
| インターネット | 3862 | ... 50937-1, 64742-8, 68451-8, 69227-17, 69228-6, 85258-2, 89966-10, 92454-7, 94755-8, 98170-1, 103220 ... |
| インターン | 28 | ... 67940-2, 54697-10, 54336-18, ... |
| | | |
| | | |
| データ | 3301 | ... 84161-9, 84594-1, 85258-7, 85850-6, 90102-13, 93397-15, 95151-18, 89150-4, 98168-4, 98170-5, 98175-8, 98572-13, 101090-2, ... |
| データベース | 485 | ... 53219-8, 54083-7, 77305-11, 83210-5, 98170-5, 100351-8, 123087-9, ... |
| データマイニング | 71 | ... 83943-17, 86541-3, 88630-2, 100455-9, ... |
| | | |

32

## 検索フィールド

- 検索する項目のこと
  - タイトル
  - 著者名
  - 件名 (統制語)
  - フリーワード (タイトル、著者名などすべて)

## 検索フィールド


34