災害情報学

収集（ネット上の情報収集技術）

村川猛彦

自己紹介

- 氏名：村川猛彦
- 所属：ネットワーク情報学メジャー／知能科学クラスタ
- 担当科目：情報処理Ⅰ・Ⅱ，ネットワーク演習，ネットワークセキュリティ，Webアプリケーション構築演習など
- 研究：データ工学，ディジタルアーカイブ，全文検索，災害ブログ記事収集など
- 個人的なこと：貴志川線沿線在住，双子を含む4人の娘の父親

この授業でお話しすること

- ネット上の情報収集
- 検索にまつわる概念・用語
- 情報の収集・蓄積を，自分でも
- 自然言語処理と全文検索

配付資料について
- 授業内容の理解を支援するためのものです．
- 12枚のスライドを選び，両面に印刷しました．試験では参照できません．
- 一部，意図的に空欄にしています．講義を聴き，記入してください．

検索の第一歩

- PCでもスマートフォンでもいいので,「ぐぐる」
- 音声検索も可
- 行きつ戻りつ

- 「検索」の英単語はsearchだが,「情報検索」はinformation retrievalと書くことが多い.

災害関連で見ておきたい情報

- 東日本大震災アーカイブ
- NHK東日本大震災アーカイブス
- 国立国会図書館東日本大震災アーカイブ（ひなぎく）
- 3.11 震災・復興 - ニュース特集：朝日新聞デジタル

アーカイブとは

- 定義1：重要記録を保存・活用し，未来に伝達すること
- 定義2：歴史的な記録の集積．個人や組織がその生涯や存続期間を通じて生み出した一時的な記録の総体であって，意識や公認されないもの（怪文書など）も含む．
- archiveと綴り，「書庫」「保存記録」と訳される．archivesは「公文書館」を意味する．
- コンピュータを活用したアーカイブは，その用途などに応じて「ウェブアーカイブ」「ディジタルアーカイブ」と呼ばれる．

他に保存・伝達するところは？

- 公的に運営されている施設のうち、文化・自然遺産に関わる品を収集しているもの
- 施設の例：博物館（Museum），図書館（Library），美術館（Gallery），公民館（Kominkan）
- 頭字語の例：GLAM，MLAK
- コンテンツの電子化や，インターネットでの情報公開も，さかんに行われている．「図書館情報学」と呼ばれる学問分野が関係する．
- 和歌山大学では図書館のほか，紀州経済史文化史研究所がこの役割を担う．

検索にまつわる概念・用語

- 全文検索：複数の文書（の中身）から特定の文字列を検索すること.
- 検索エンジン（サーチエンジン）
  - インターネットに存在する情報を検索する機能やサービス
  - 全文検索機能を備えたソフトウェア
- メタデータ：あるデータへの付随的なデータ．検索対象となる．人手で付与する（アノテーション）ほか，デジタルカメラなどでの撮影時に作られるExif情報などもメタデータと呼ばれる.

検索にまつわる概念・用語

- ワイルドカード・正規表現：少ない字数で条件（パターン）を指定する様式・文法.
  「.txtで終わる文字列」の指定は
  - ワイルドカードだと，*.txt
  - 正規表現だと，^.*¥.txt$
- 完全一致：データベースなどの検索で，検索語と完全に一致する状況や検索条件をいう.
  - 全文検索は一般に「部分一致」である.
  - 「前方一致」「後方一致」などが利用可能な検索システムもある.

情報の収集・蓄積を，自分でも

- インターネットで利用可能なサービスは
  - データ量が多く，内容チェックもなされている.
  - 効率よく検索する仕組みが備わっている.
- しかし
  - URLや利用方法の変更，サービスの終了も起こり得る.
  - 望む情報が得られない場合もある.
- そこで
  - 自発的に収集・蓄積すれば，自分（の範囲）で情報をコントロールできる．良いものを公開して，収益を得たり，注目を集めたりすることも.

収集した情報を記録するには

- その都度，保存
- ブラウザのブックマーク
  - FirefoxやChromeでは，異なるPC間（同一アカウント）のブックマーク共有も可能.
- ソーシャルブックマーク
  - はてなブックマークなど．誰がいつどのページにブックマークしたかや，ページに付けられたブックマーク数・コメントなども知ることができる.
- クラウドストレージ
  - Dropbox，Evernoteなど．共有や検索がしやすい．基本無料とし，保存量や月間アップロード量などに制約があるものも.

収集した情報を検索するには

- メタデータを検索 vs 本文を検索
- 本文検索の方法（例）
  - アプリケーションの検索機能
  - grep型（非インデックス型）検索
  - インデックス型の全文検索エンジンを導入
- 多様なファイル形式への対応
  - 検索対象は基本的にテキスト（txt，html，xml，csv）
  - doc(x)，xls(x)，pdfには，テキスト抽出ツール（xdoc2txt，ChupaTextなど）を使用するとよい.

自然言語処理

- 言語の分類
  - 自然言語：人間どうしの意思疎通のための言語．日本語・英語など．
  - 人工言語：コンピュータ言語（プログラミング言語を含む）など
- 日本語は単語ごとに区切られていない．
  - 分かち書き
  - 形態素解析（MeCabなど）
  - 係り受け解析も（CaboChaなど）

通常の文	雨が降れば傘をさす
文節区切り	雨が　降れば　傘を　さす
単語（形態素）区切り	雨　が　降れ　ば　傘　を　さす

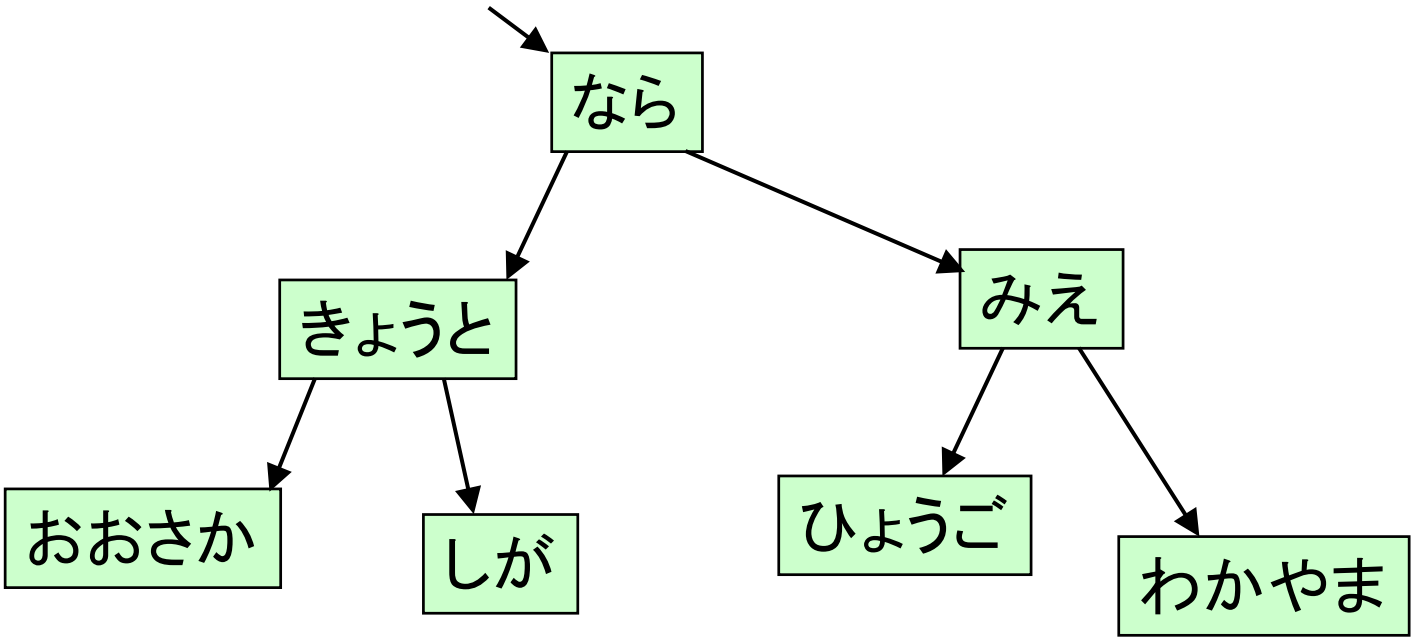
形態素解析ソフトウェアMeCabの実行例

```
$ echo "雨が降れば傘をさす" | mecab
雨　　名詞,一般,*,*,*,*,雨,アメ,アメ
が　　助詞,格助詞,一般,*,*,*,が,ガ,ガ
降れ　動詞,自立,*,*,五段・ラ行,仮定形,降る,フレ,フレ
ば　　助詞,接続助詞,*,*,*,*,ば,バ,バ
傘　　名詞,一般,*,*,*,*,傘,カサ,カサ
を　　助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
さす　動詞,自立,*,*,五段・サ行,基本形,さす,サス,サス
EOS
$
```

全文検索の分類とインデックス

- 非インデックス型検索
  - grepなど．自作も容易．
  - インデックスを使用しない．
  - 全探索となるため，検索に要する時間は，データ量に比例する（遅い）．
- インデックス型検索
  - フリーではElasticsearchやGroonga，商用ではAccela BizSearchなど．
  - 事前にインデックス（索引）を作成する．その分，データ領域も必要．
  - 検索はインデックスの一部の参照により瞬時に行われる（速い）．

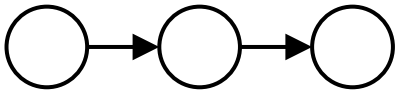
2府5県のインデックス

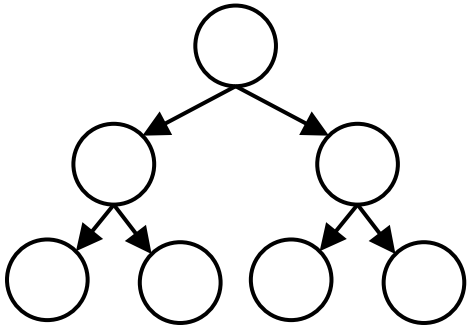


- 「しが」は？
- 「ふくい」は？

検索時間の比較

- 10億件の情報から検索するとき
  - 非インデックス型では，
    1件の照合に要する時間×10億
  - 二分木に基づくインデックス型では，
    1件の照合に要する時間×30
- n件の情報から検索するとき
  - 非インデックス型ではO(n)…線形時間
  - インデックス型ではO(log n)…対数時間





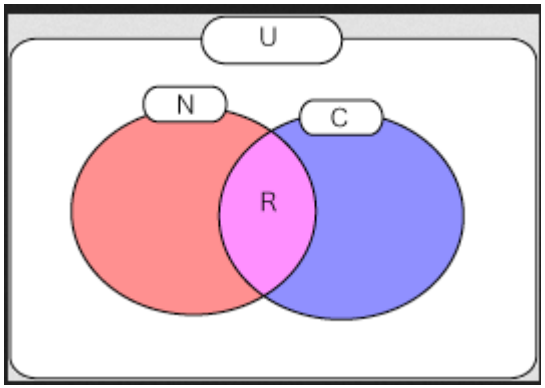
なぜか？
- 高さhの二分木には最大2^h-1のノード（情報）を持つことができる.
- $n=2^h-1$ とし，これをhについて解くと，$h=\log_2(n+1)=O(\log n)$

10億$=1000^3 \fallingdotseq (2^{10})^3=2^{30}$

欲しい情報が，検索できる？

- 望ましくない検索結果
  - 検索漏れ：「チョコ」で検索しても，「チョコレート」がヒットしなかった．
  - 検索ノイズ：「京都」で検索すると，「東京都」が含まれた．
- 検索性能の指標
  - 再現率：R / N
  - 適合率：R / C
  - F値：$\dfrac{R}{\frac{1}{2}(N+C)}$

（再現率と適合率の調和平均）

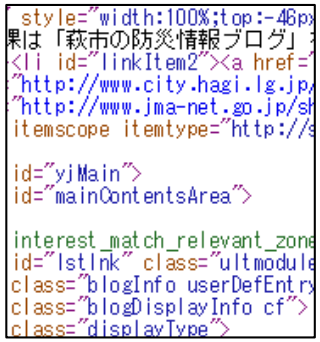


N：検索結果の文書の数
C：全対象文書中の正解文書の数
R：検索された適合文書の数

Webスクレイピング

- あるトピックのインターネット上の情報を何万件も収集したい…となったら，スクレイピング
- スクレイピングとは：Webページのデータ（HTMLなど）を取得し，そこから特定の情報を抽出すること.
- 基本的な手順：URLよりコンテンツ獲得，メタデータ・本文抽出，リンク抽出，データ管理

ブラウザ表示



ソース(HTML)

本文抽出

円滑な情報収集・管理・公開のために

- 著作権，肖像権，個人情報およびプライバシー，robots.txtに配慮する.
- 規格やガイドラインに従う.
  - ひなぎくに関連して，総務省が「震災関連デジタルアーカイブ構築・運用のためのガイドライン」を公表している.
  - XMLで管理する際には「ダブリンコア」（Dublin Core, DC）も.
- 元の情報が変更・削除された場合の対応を決めておく.
  - 「2017-12-27参照」などの表示も一つのやり方