数物･電子情報系学科専門科目
情報工学概論　Webと情報検索

森　辰則

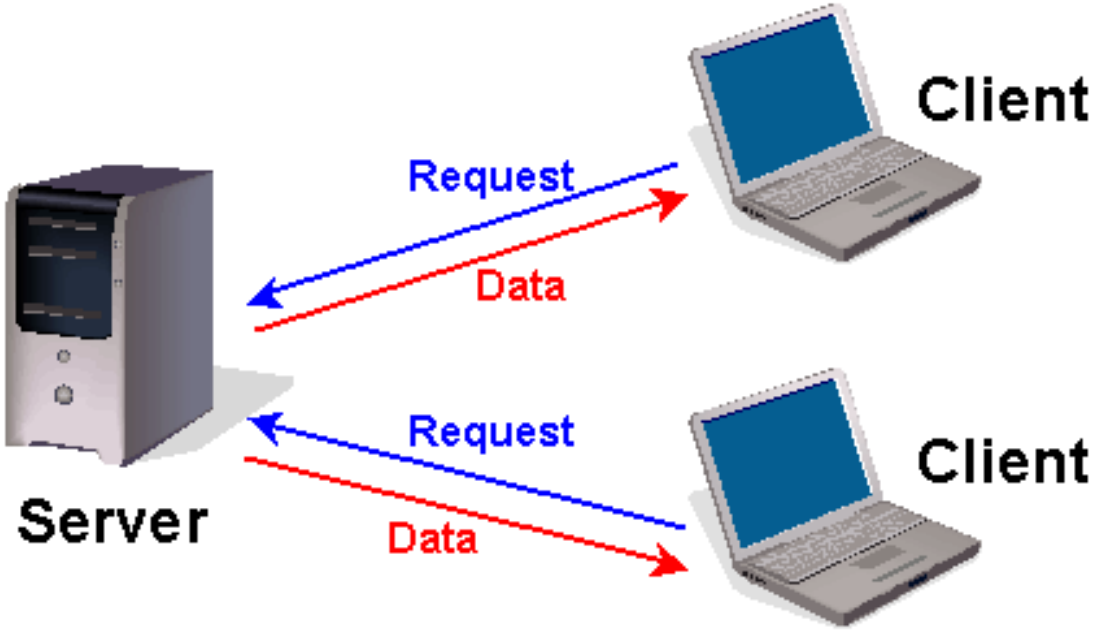
mori@forest.eis.ynu.ac.jp

分散処理システム

- 分散処理システム
  - ネットワークによって接続されている
  - 一群の計算機が
  - 協調して情報処理を行なう
- 透過性
  - 分散処理システムにおいて計算資源が分散していることを利用者に意識させないこと
    - どこにいても同じように使える
    - 分散処理システムの目標

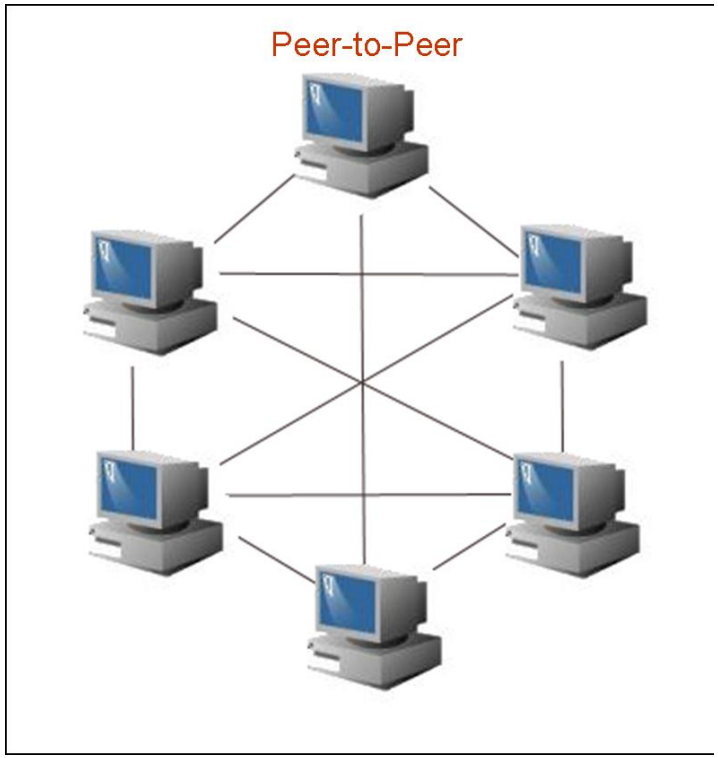
サーバ・クライアントモデル(1)

- インターネットプログラムを理解するキーポイント!!
- ネットワーク上の計算機の間に主従関係があるモデル



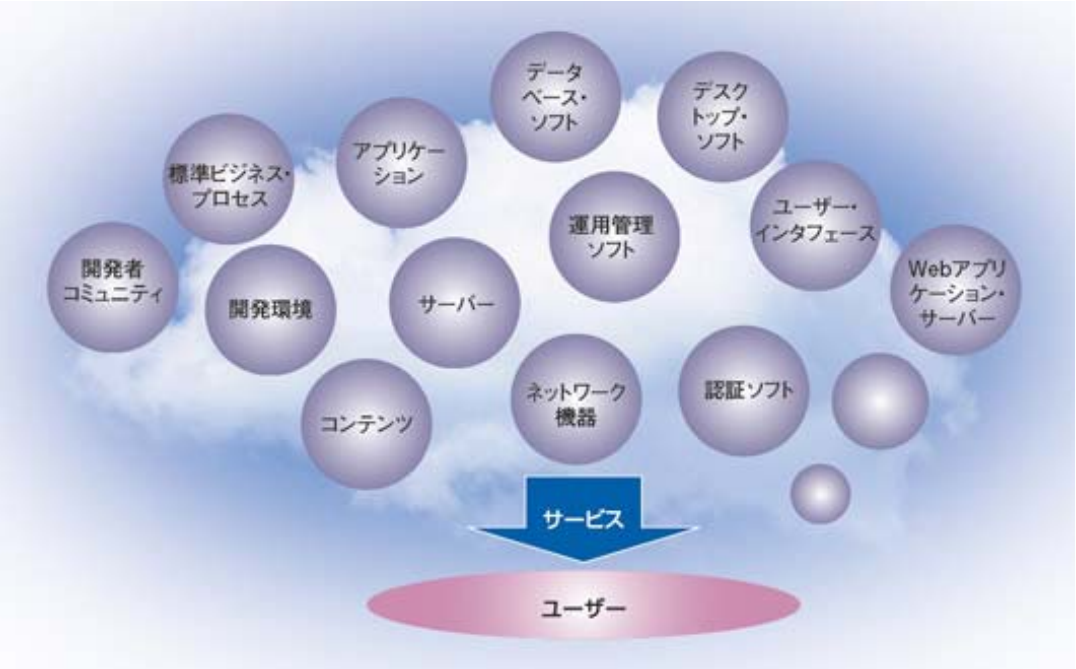
P2Pモデル (1)

- Peer to Peer の略．Peer = 仲間，同等のもの．
- コンピュータ同士が直接通信をして，お互いの持つ情報をやりとりする通 信形式．通信するコンピュータ間には主/従関係がないことが特長．不特定多数のコンピュータ間で直接情報のやりとりが行なえる方式がほとんど．



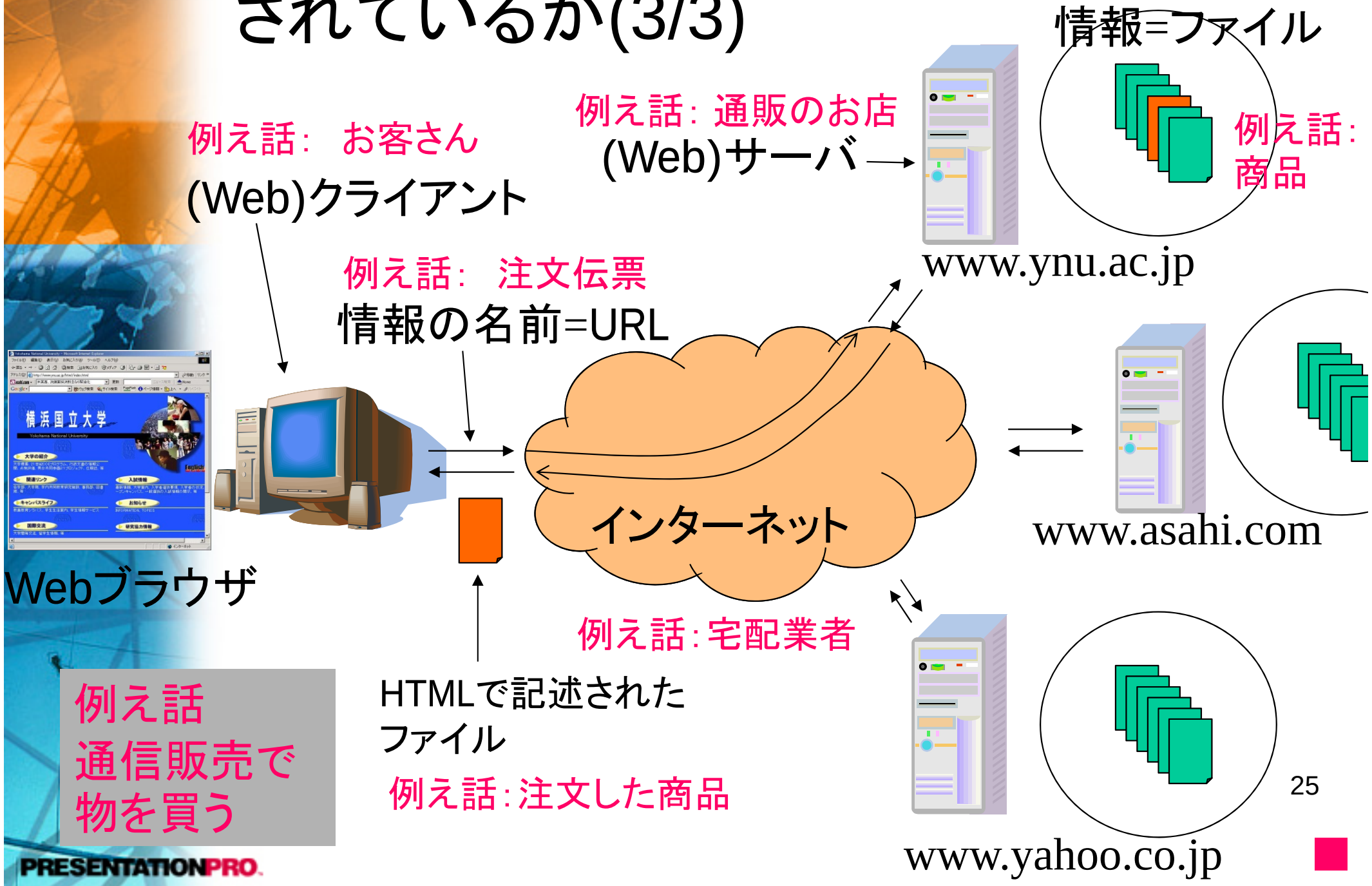
クラウドコンピューティング (1)

- クラウド=雲。ネットワークを図示するときに雲形の図形を描いていた。
- ネットワーク上に分散して存在する計算資源を利用して、利用者に情報サービスを提供するコンピュータ処理の方式



日経BP IT pro より引用
http://itpro.nikkeibp.co.jp/article/COLUMN/20080410/298616/

WWWはどのようにして全世界に公開されているか(3/3)



例え話
通信販売で
物を買う

URL (Uniform Resource Locator)

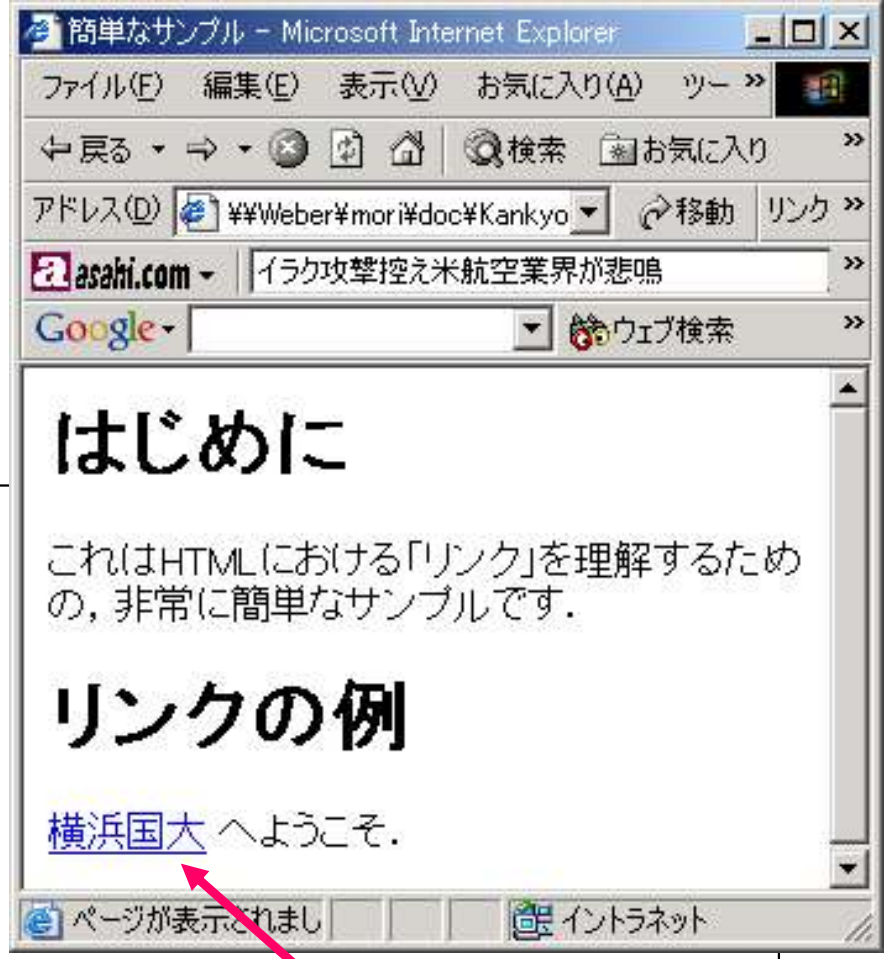
- 情報の在処を示す記号列
- 「授受の方式」＋「サーバの名前」＋「サーバ内での情報の名前」
- 例：横浜国立大学のWebページのURL

http://www.ynu.ac.jp/index.html

- http://　：http という方式　　例え話：輸送の方式
- www.ynu.ac.jp　：www.ynu.ac.jp と言う名前のコンピュータ　　例え話：通販業者の名前
- /index.html　：/index.htmlという名前の情報（ファイル）　　例え話：商品の名前

HTML (2/2)

リンクの箇所をマウスでクリックすると対応するURLのWebページが表示される.



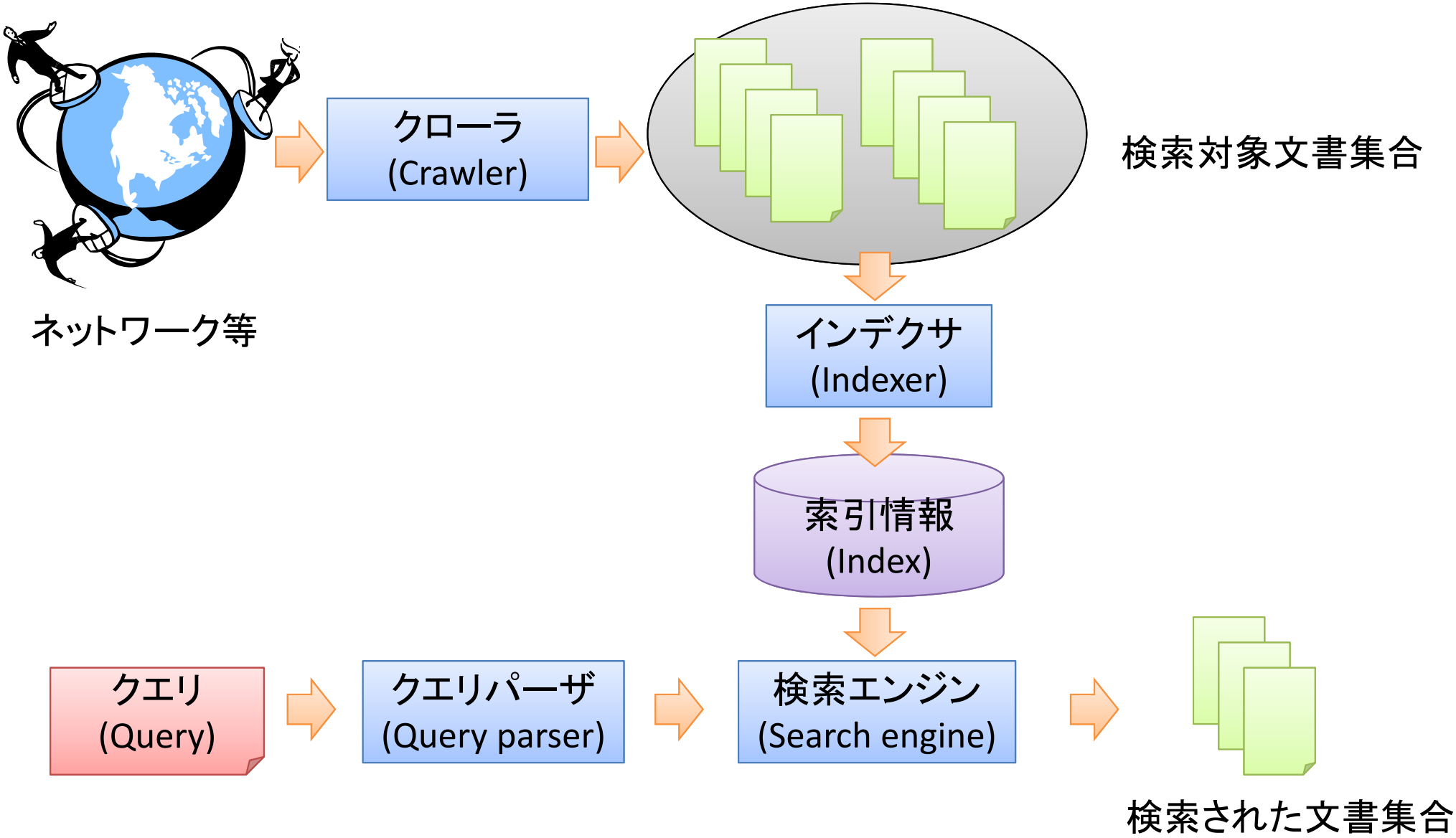

```
<html>
<head>
<title>簡単なサンプル</title>
</head>
<body>
<h1>はじめに</h1>
これはHTMLにおける「リンク」を理解するための，
非常に簡単なサンプルです．

<h1>リンクの例</h1>
<a href="http://www.ynu.ac.jp/index.html">横浜国大</a>
へようこそ．
```

横浜国大ホームページへのリンク

横浜国大ホームページのURL

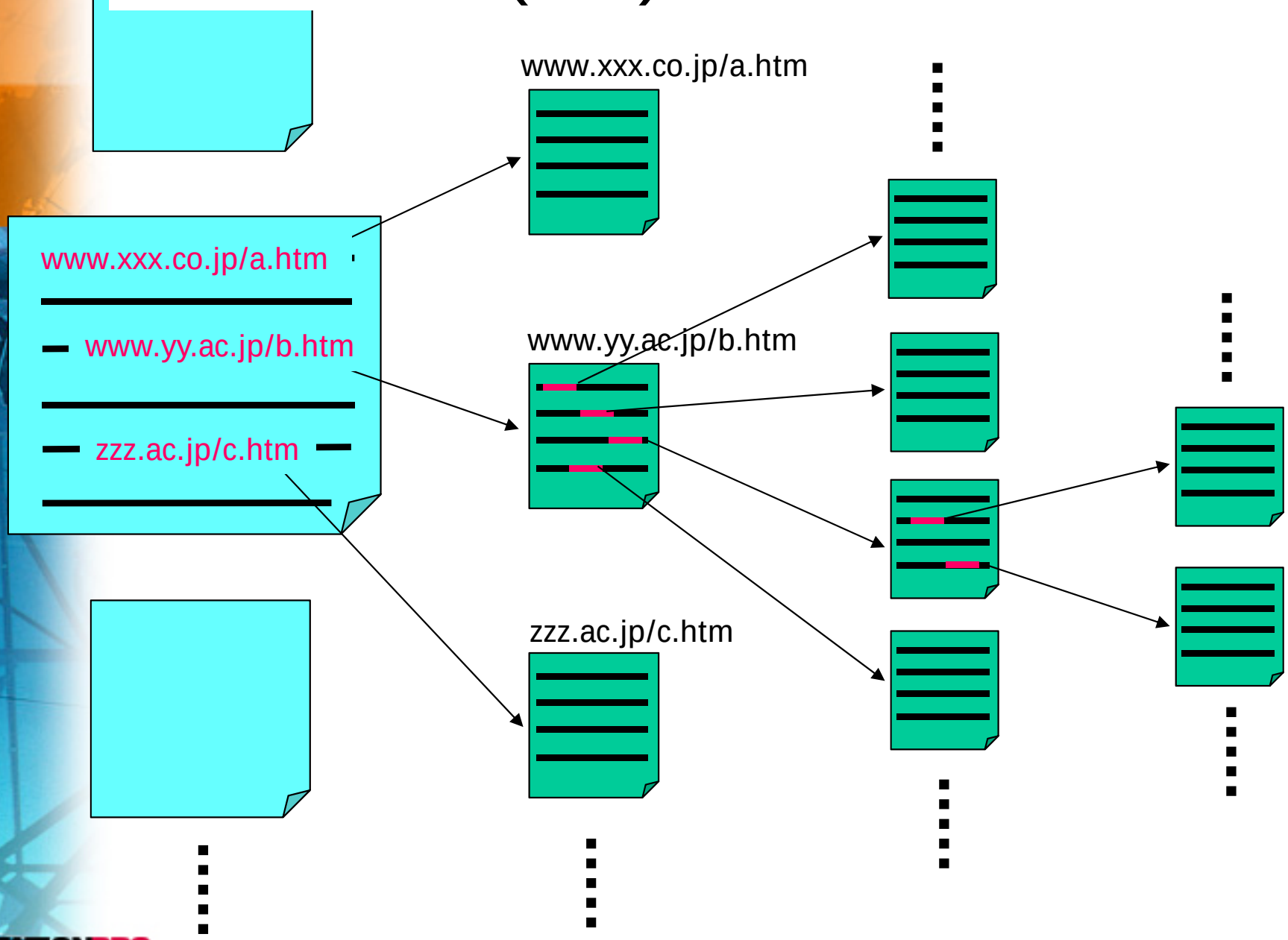
例え話：関連商品の注文伝票

検索型Webページ情報サービスの仕組み



インターネット上のWebページを網羅的に集める(2/3)

初期URLリストに登録されているWebページ



利用者の要求を満足するWebページの検索

1. 検索要求として与えられた単語を含むWebページを見つける。
2. 見つかったWebページに対して、重要度を計算し、その値の大きいものから順に提示する。
   - 検索要求との適合度による重要度計算
   - Webページの人気度に基づく重要度計算

与えられた単語を含むWebページを見つける方法

検索用辞書による方法

- 辞書1: 語➔語の番号
- 辞書2: 文書の番号➔語の番号(のリスト)
- 辞書3: 語の番号➔文書の番号(のリスト)

- クローラがWebページを訪れる度に、
  1. 辞書1によりそのページに入っている語が既に登録されているかを調べる。未登録ならば新しい語の番号を与えて登録。
  2. 辞書2を更新。

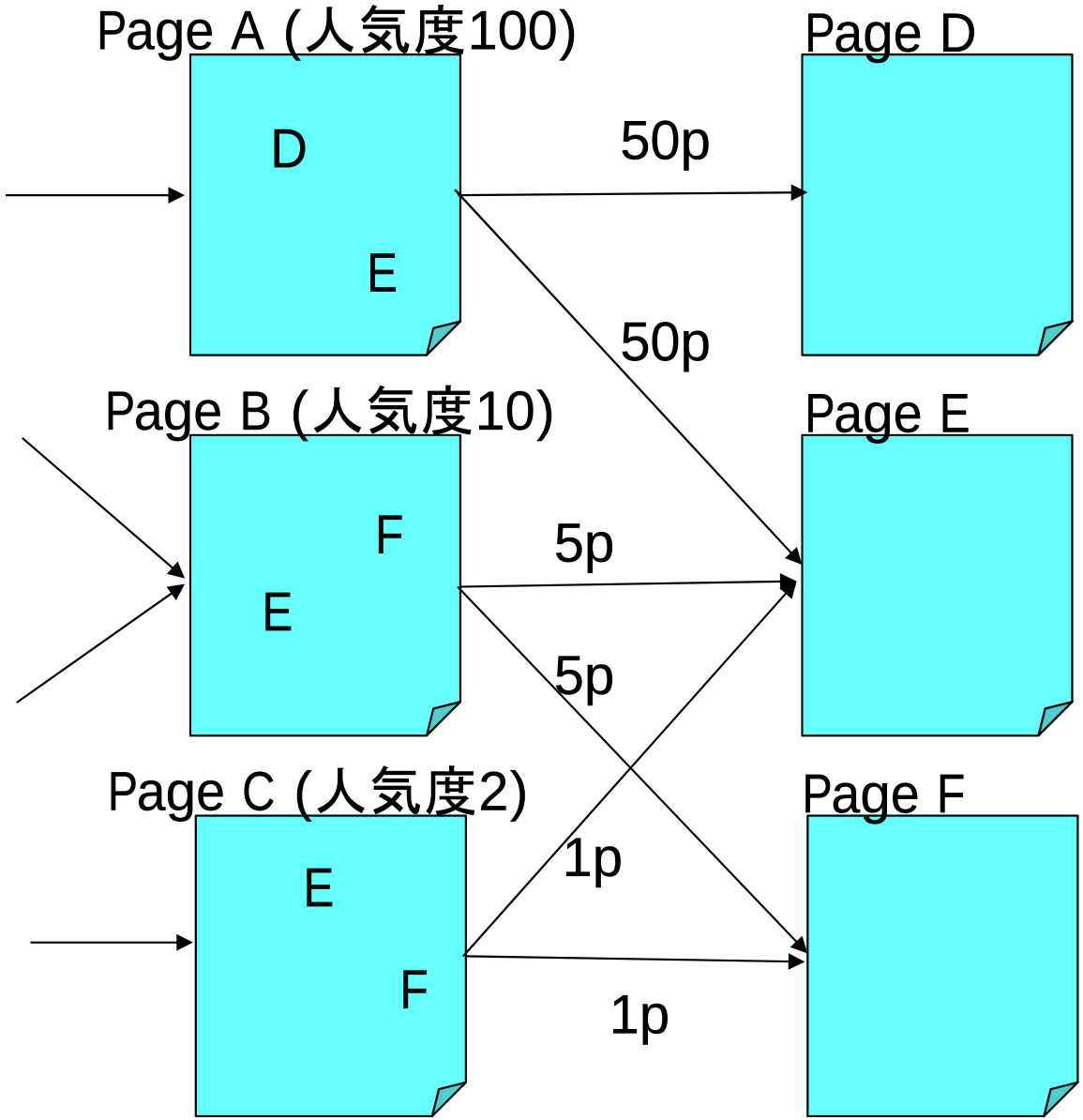
  全てWebページを訪れたあと、辞書2から辞書3を生成。
- ある語が与えられると、
  1. 辞書1により、語の番号に変換される。
  2. 辞書3により、その語が含まれる文書のリストを得る。

各Webページの重要度

- 検索質問に含まれる語が全て含まれるWebページでもその重要度は異なる。
  - 主要な話題なのか、枝葉の話題なのか
  - そのページ自身の信頼度

1. 語の分布に基づく方法
   - TF (Term Frequency)
   - IDF (Inverse Document Frequency)
2. Webページ中での取扱に基づく方法
   - タイトルや見出し語になっているか、太文字や大きな字体になっているかなど。
3. リンクの参照関係に基づく方法
   - GoogleにおけるPageRank手法

PageRank法(2/3)



人気度➔50

人気度➔50+5+1
➔ 56

人気度➔5+1
➔ 6

PageRank法(3/3)

- より正確にはランダムサーファモデルによる、確率過程でモデル化される。
  - ある一定確率eでランダムに別のWebページを選択する。
  - 確率$(1-e)$で現在のWebページ内のリンクを等確率で辿る。
- WebページpのPageRank $R(p)$は次の式を繰り返し適用して求めることができる。

$$R(p) = \frac{e}{n} + (1-e) \cdot \sum_{(q,p) \in G} \frac{R(q)}{\mathrm{outdegree}(q)}$$

q: pへのリンクを持つページ、$\mathrm{outdegree}(q)$: qからのリンクの数

演習

1. WebブラウザにURLを入力すると表示されるが、検索エンジンでは検索されないようなWebページはどのようにしたら作れるか？
2. PageRankはWebページの「人気度」を測るための一手法である。PageRankとは異なる手法でWebページの「人気度」を測る方法を考えてみよ。