教養教育科目
情報工学概論

Webと情報検索

森　辰則
mori@forest.eis.ynu.ac.jp

この講義の目的

- ネットワークを利用した処理システムである、分散処理システムとそのモデルについて学ぶ。
- インターネット利用において、最も重要なサービスであるWWW (World Wide Web) について、その検索サービスの仕組みを学ぶ。

目次

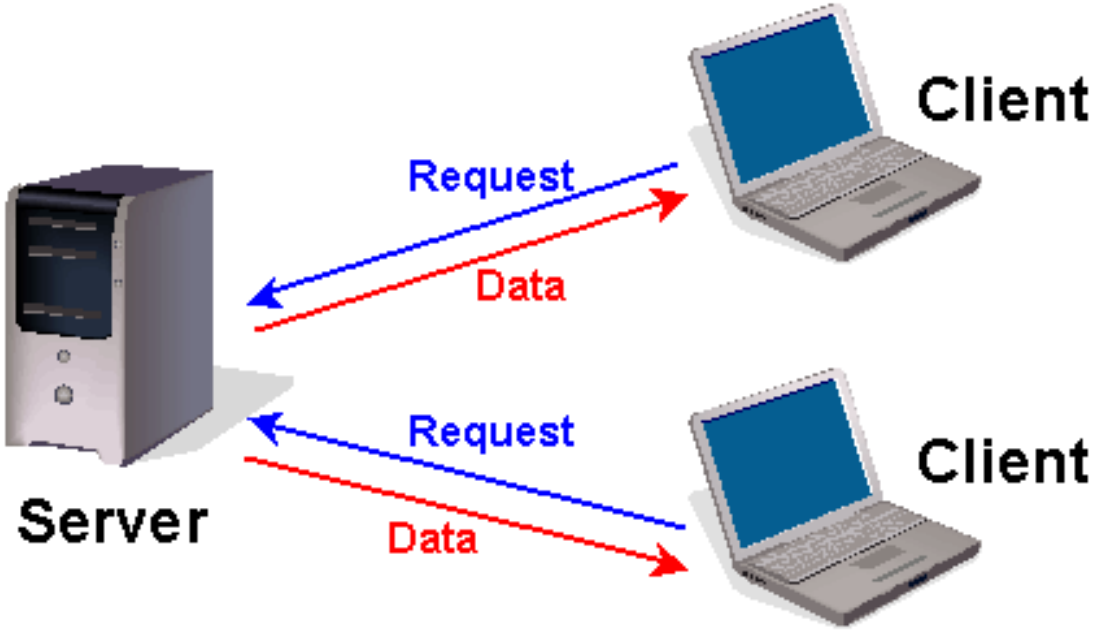
- 分散処理システムとそのモデル
- World Wide Web
- インターネットの「入り口」
- 検索型Webページ情報サービス
  - どうやって情報を集めるか
  - どうやって利用者の要求に適したページを探すか

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

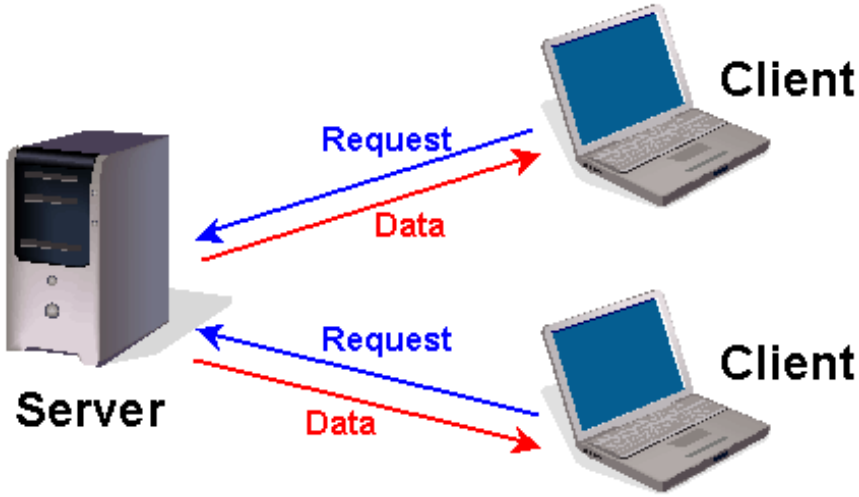
サーバ・クライアントモデル(1)

- 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



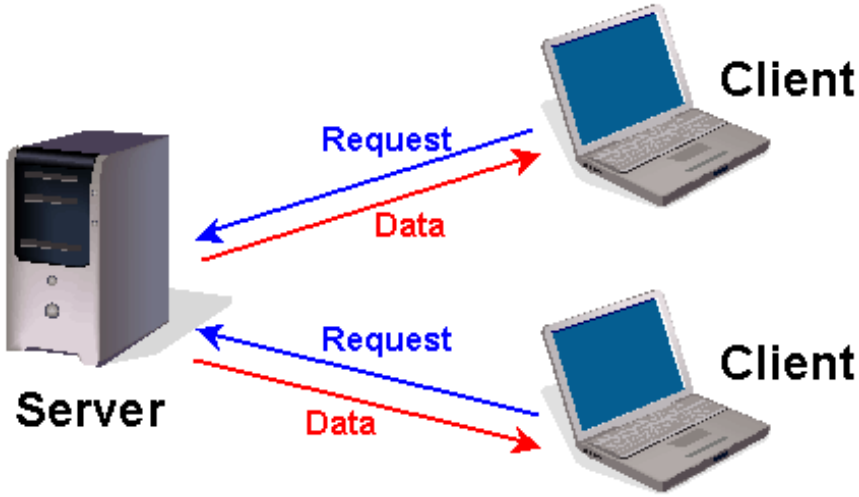
サーバ・クライアントモデル(2)

- サーバ(server)
  - 元の意味: 何かサービスをしてくれる人/物
  - 電話なら: 電話局の設備
  - ATMなら: 残高管理システム
  - ネットワークでは: ある特定のサービスを提供する計算機(とそのプログラム)



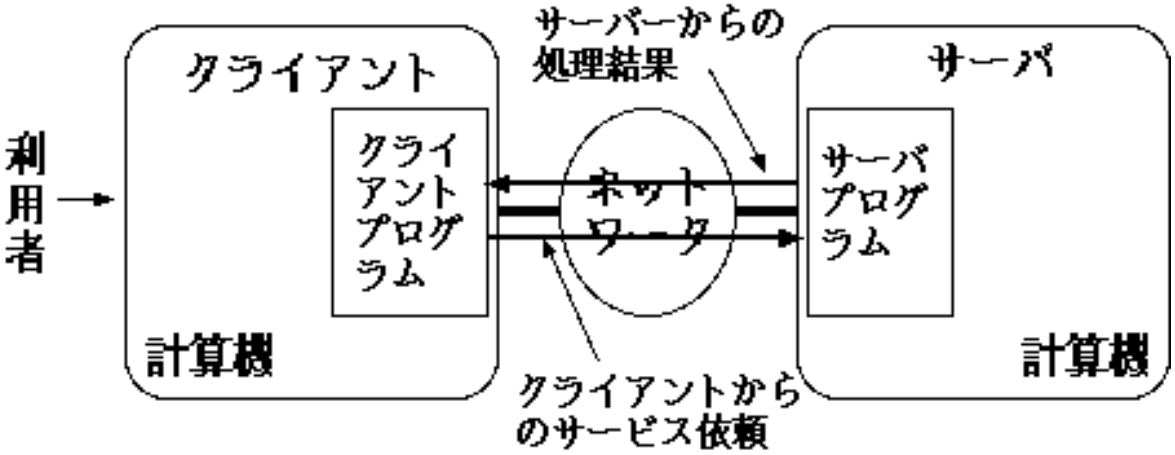
サーバ・クライアントモデル(3)

- クライアント(client)
  - 元の意味: あるサービスを利用する人．つまり，顧客
  - 電話なら: 電話器
  - ATMなら: ATMの端末(お金を出し入れするところ)
  - ネットワークでは: ある特定のサービスを利用する計算機(とそのプログラム)



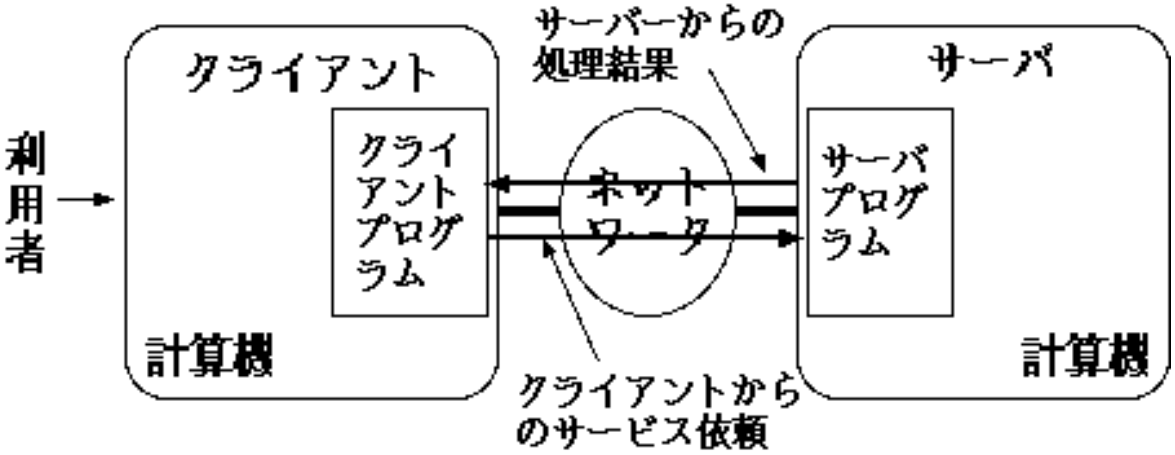
サーバクライアントモデルの動き (1)

- サーバ，クライアントともにサービスに対応する共通のプロトコルを使う.
  1. クライアントがサーバに向けて「サービスを利用したい」というメッセージを送る.
  2. サーバはそのメッセージをうけて，実際の処理を行ない，処理結果をクライアントに送る.
  3. 1 に戻る.



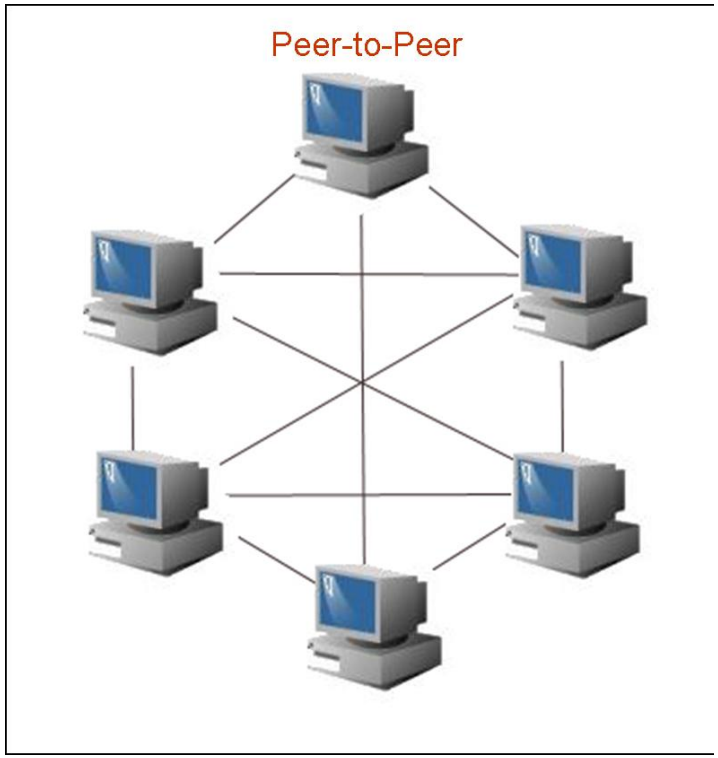
サーバクライアントモデルの動き (2)

- 利用者からみたサーバクライアントモデル
  1. 利用者は，ある特定のサービス(例えば，電子メール)を利用するために，そのサービスに対応するクライアントプログラムを用意する.
  2. そのクライアントプログラムを，サーバとなる計算機を指定して動かし，サービスを利用する.



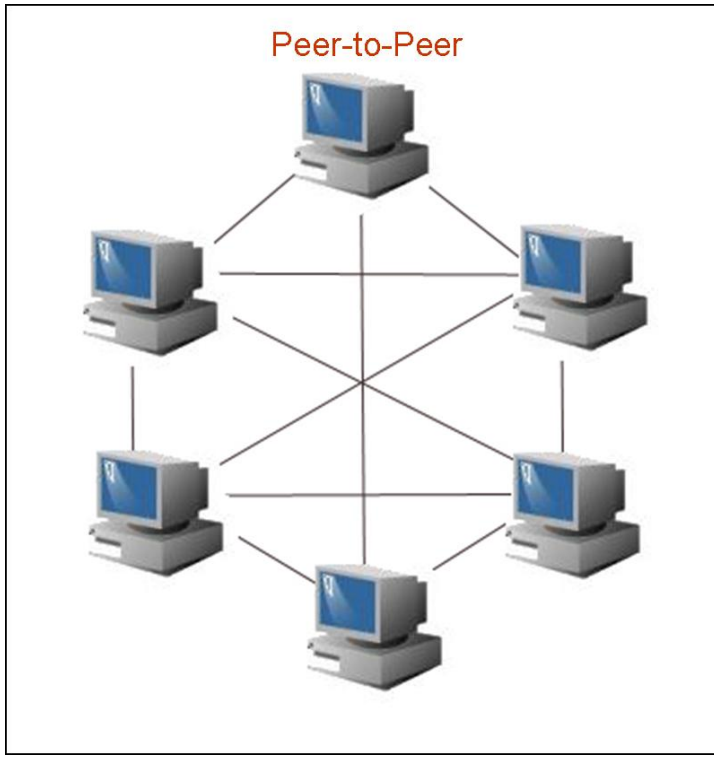
P2Pモデル (1)

- 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．



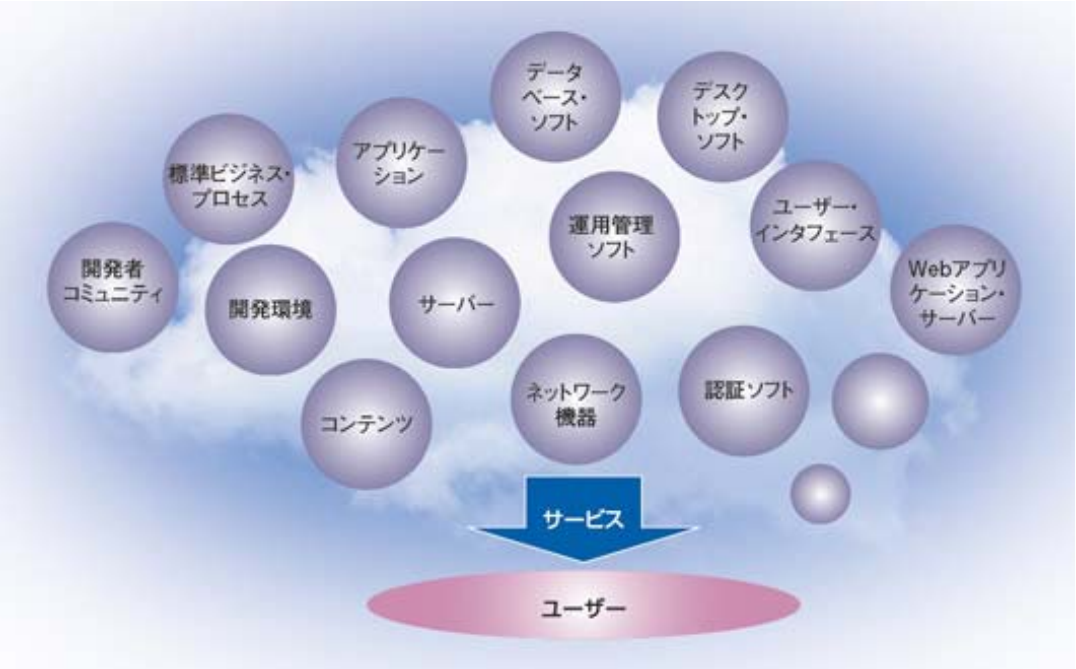
P2Pモデル (2)

- 情報の在処を調べるために中央サーバが必要な方式(中央サーバ型)と，バケツリレー式にデータを転送し中央サーバが不要である方式(純粋型)がある.
- ネットワーク上でファイルを共有する「ファイル交換ソフトウェア(file exchange software)」などで用いられている.
  - P2Pモデル自身は中立であるが、ファイル交換ソフトには、問題が存在することが多い。
    - 著作権保護の問題
    - ウィルスソフトウェアの標的になりやすい



クラウドコンピューティング (1)

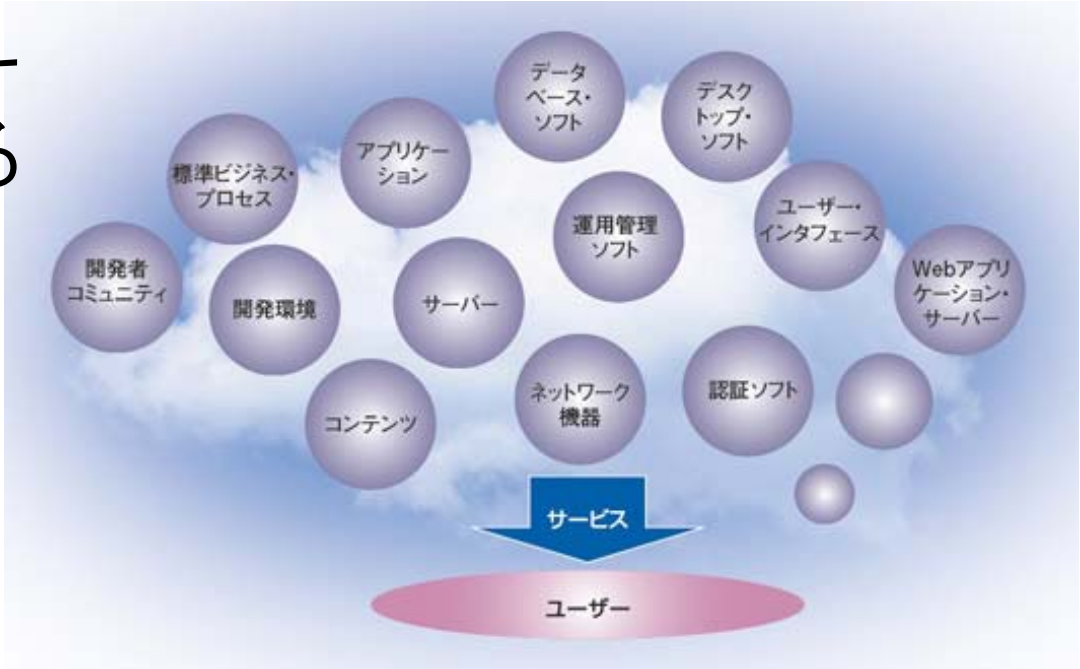
- クラウド=雲。ネットワークを図示するときに雲形の図形を描いていた。
- ネットワーク上に分散して存在する計算資源を利用して、利用者に情報サービスを提供するコンピュータ処理の方式



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

クラウドコンピューティング (2)

- 特長
  - ネットワーク上に存在する計算資源を意識することなしに利用できる仕組み
  - 計算資源: 計算能力、記憶能力、通信能力 etc.
- 利用者が用意するものは
  - ネットワークに接続できる計算機
  - Webブラウザ(後述)のような表示システム



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

クラウドコンピューティングの例(1)

- Amazon Web Services
  - http://aws.amazon.com/jp/
  - Amazon Elastic Compute Cloud（EC2）
    - コンピューターの処理能力を提供
  - Amazon Simple Storage Service（S3）
    - ストレージサービス
  - Amazon SimpleDB
    - データベースサービス
  - 他

クラウドコンピューティングの例(2)

- Google App Engine
  - https://developers.google.com/appengine/?hl=ja Webアプリケーションを Googleの計算資源上で実行できる。
- Microsoft Windows Azure Services Platform
  - Windows Azure
    - クラウドOS
  - Building Block Service
    - アプリケーション用のサービス

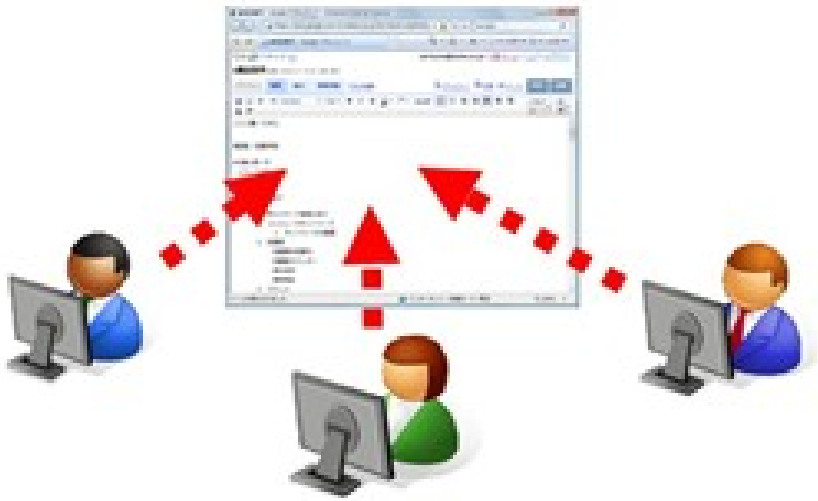
クラウドコンピューティングに基づくアプリケーションの例

- アプリケーション
  - Googleの一連のオンラインアプリケーション
    - Gmail
    - Googleマップ
    - Googleドキュメント
  - Microsoft Office Web Apps
- オンラインストレージ
  - ネットワーク上に存在する外部記憶装置が使える。
- オンラインデスクトップ (デスクトップクラウド)
  - どのコンピュータをつかっていても、Webブラウザ上で同じデスクトップ環境を使える。
  - アプリケーション、オンラインストレージのサービスと組み合わされることが多い。

Googleドキュメント

- ネットワーク上のドキュメントファイルを複数の利用者が共同で編集できる

Google Document でのリアルタイム共同編集



複数のユーザで、同時に、かつリアルタイムに、
一つのファイルを編集することが可能

http://kimc.kaetsu.ac.jp/helpdesk/index.php?Google%A5%C9%A5%AD%A5%E5%A5%E1%A5%F3%A5%C8
より引用

オンラインデスクトップ(デスクトップクラウド)の例

- Glide OS
  - Webブラウザの画面上に、デスクトップ環境が表示されマウス操作等で利用できる。

World Wide Web

World Wide Web (WWW) (1/2)

- インターネットの代名詞

World Wide Web (WWW) (2/2)

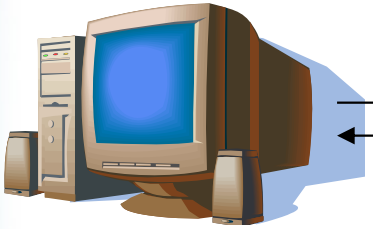
- わーるどわいどうぇっぶ，うぇっぶ，だぶりゅだぶりゅだぶりゅ，だぶりゅすりー
- 基本は、不特定多数に情報を公開する「広告型」の情報サービス
  - 本のような「読み物」。
  - 通常は「置き場所」がわかれば、無料で読める。
    - 世界中に広がった図書館のようなもの
  - 利用における2つの立場
    - 読者の立場: 他人が書いた「読み物」を読み、自分の知りたい情報を集める立場
    - 著者の立場: 自分の知っていること，あるいは，PRしたいことを書いて，広く皆に読んでもらう立場
- 近年、WWWの仕組みを基盤としてさまざまなサービスが展開されている 。

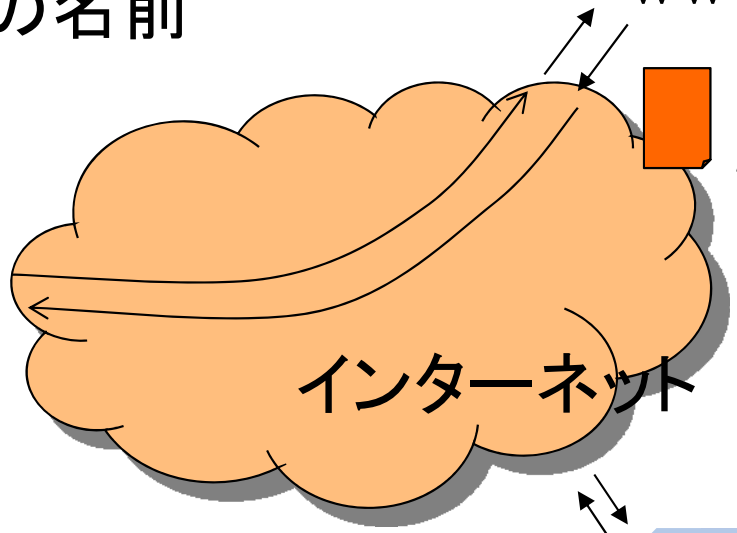
WWWとホームページ

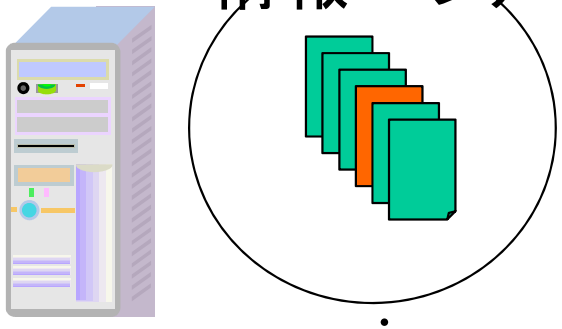
- WWW (World Wide Web)
  - 全世界に広がる蜘蛛の巣状の情報網
  - インターネット上の（独立した）多数のコンピュータが情報提供を行う
- ホームページ
  - ある組織や団体が、
  - WWWの仕組みにより、
  - 広く一般に公開している情報。
  - 特に、その「表紙」に当たる情報を指すこともある。
  - 幾つかの個別の情報単位の集合で構成されることもあるので、その各々を「Webページ」と呼ぶことも

WWWはどのようにして全世界に公開されているか(1/3)

情報の名前=
コンピュータの名前＋
そのコンピュータの中の
「ファイル」の名前

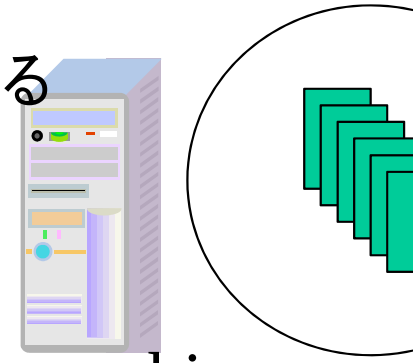




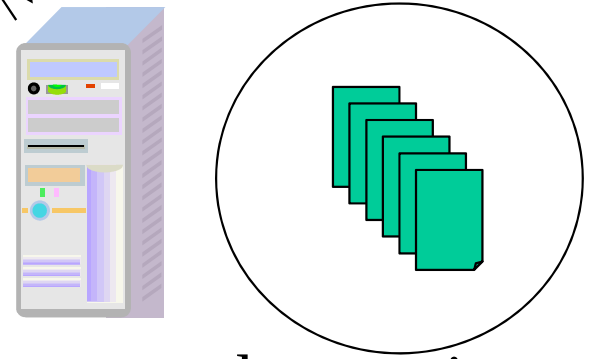


www.ynu.ac.jp

該当する
情報



www.asahi.com



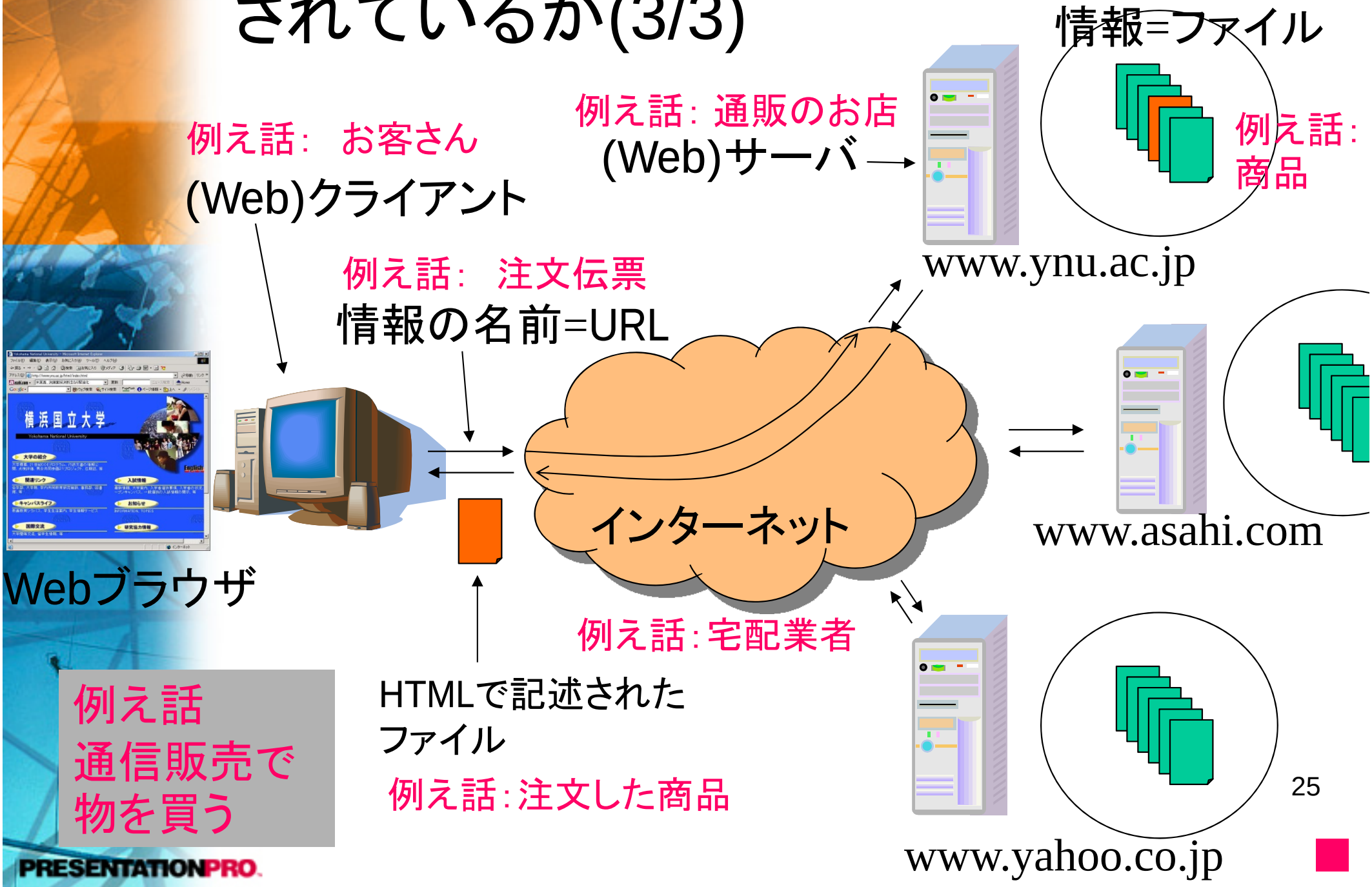
www.yahoo.co.jp

指定したコンピュータに
特定のファイルを要求すると、
そのコンピュータがそのファイルの内容を転送してくれる。

WWWはどのようにして全世界に公開されているか(2/3)

- 通信路はインターネット➔全世界に広がる
- サーバ･クライアント形式の情報伝達
  - サーバ(server) ：情報を提供するシステム（インターネット上に存在するどこかのコンピュータ）
  - クライアント(client)：情報を受け取るシステム（利用者のコンピュータ）．この場合、Webブラウザというソフトウェア．
- 伝達される情報
  - クライアント➔サーバ
    - どの情報が欲しいか？
    - URL (uniform resource locator)とよばれる記号列で指定
  - サーバ➔クライアント
    - 提供される情報=「ファイル」=「文書」
    - HTMLという形式に従った内容を持つファイル
    - HTML=HyperText Markup Language

WWWはどのようにして全世界に公開されているか(3/3)

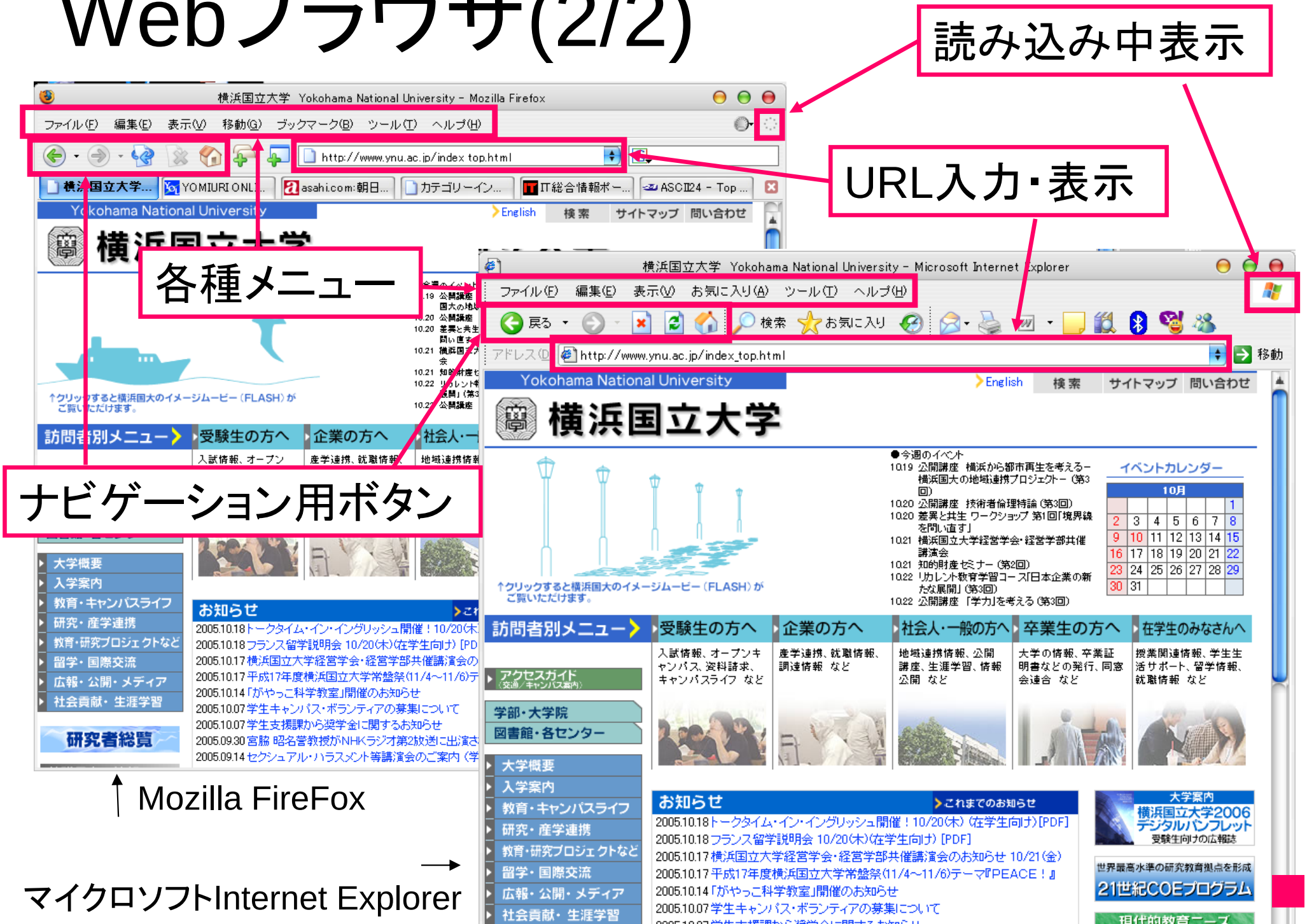


例え話
通信販売で
物を買う

Webブラウザ(1/2)

- 利用者が用いるクライアントソフトウェア
  - URLで指定した情報を該当サーバに要求
  - 該当サーバから受け取った情報を解釈して、綺麗に表示
- 例え話：注文受付と商品利用

Webブラウザ(2/2)


読み込み中表示
URL入力・表示
各種メニュー
ナビゲーション用ボタン
横浜国立大学 Yokohama National University - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 移動(G) ブックマーク(B) ツール(T) ヘルプ(H)
http://www.ynu.ac.jp/index_top.html
横浜国立大学 Yokohama National University - Microsoft Internet Explorer
ファイル(F) 編集(E) 表示(V) お気に入り(A) ツール(T) ヘルプ(H)
戻る 検索 お気に入り
アドレス(D) http://www.ynu.ac.jp/index_top.html
移動
Yokohama National University
English 検索 サイトマップ 問い合わせ
横浜国立大学
訪問者別メニュー
受験生の方へ
企業の方へ
社会人・一般の方へ
卒業生の方へ
在学生のみなさんへ
入試情報、オープンキャンパス、資料請求、キャンパスライフ など
産学連携、就職情報、調達情報 など
地域連携情報、公開講座、生涯学習、情報公開 など
大学の情報、卒業証明書などの発行、同窓会連合 など
授業関連情報、学生生活サポート、留学情報、就職情報 など
アクセスガイド
学部・大学院
図書館・各センター
大学概要
入学案内
教育・キャンパスライフ
研究・産学連携
教育・研究プロジェクトなど
留学・国際交流
広報・公開・メディア
社会貢献・生涯学習
研究者総覧
お知らせ
これまでのお知らせ
2005.10.18 トークタイム・イン・イングリッシュ開催！10/20(木)(在学生向け)[PDF]
2005.10.18 フランス留学説明会 10/20(木)(在学生向け) [PDF]
2005.10.17 横浜国立大学経営学会・経営学部共催講演会のお知らせ 10/21(金)
2005.10.17 平成17年度横浜国立大学常盤祭(11/4～11/6)テーマ『PEACE！』
2005.10.14 「がやっこ科学教室」開催のお知らせ
2005.10.07 学生キャンパス・ボランティアの募集について
イベントカレンダー
10月
大学案内
横浜国立大学2006 デジタルパンフレット
受験生向けの広報誌
世界最高水準の研究教育拠点を形成
21世紀COEプログラム
Mozilla FireFox
マイクロソフトInternet Explorer

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

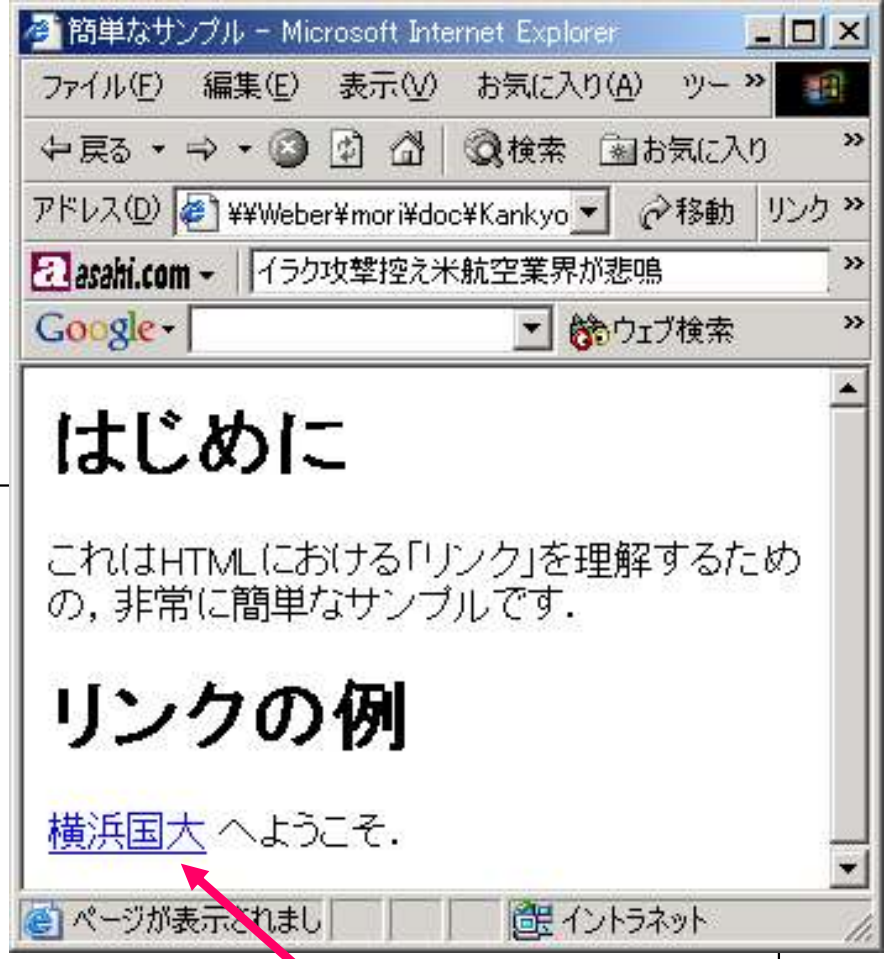
- http://　：http という方式　　例え話：輸送の方式
- www.ynu.ac.jp　：www.ynu.ac.jp と言う名前のコンピュータ　　例え話：通販業者の名前
- /index.html　：/index.htmlという名前の情報（ファイル）　　例え話：商品の名前

HTML (Hyper-Text Markup Language)(1/2)

- ハイパーテキスト（Hyper-Text）
  ＝「超文書」
- 文書と文書の間の参照関係を記述できる枠組み
- 参照関係は「リンク」と呼ばれる.
- リンクの実体はURL
- 例え話：購入した商品に、関連商品の注文伝票がついている。ついでに購入するのが楽。

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

インターネットの「入り口」

Webブラウザに最初に表示させるWebページは？

例:Yahoo! Japan (http://www.yahoo.co.jp/)

Webページ検索

最新ニュース

よく使う情報
(オークション，天気，TV， etc.)

おすすめ情報

「インターネット」の入り口

- 「ポータルサイト(portal site)」と呼ばれる。
  - ポータル(portal)＝玄関，入り口
  - サイト(site)=場所，敷地
- 良く使う情報（への参照情報）を簡潔に配置
- インターネット上のサービスを利用するための最初の一歩に位置づけられる。
- 例え話：様々な通販業者を整理して提示した総合カタログ

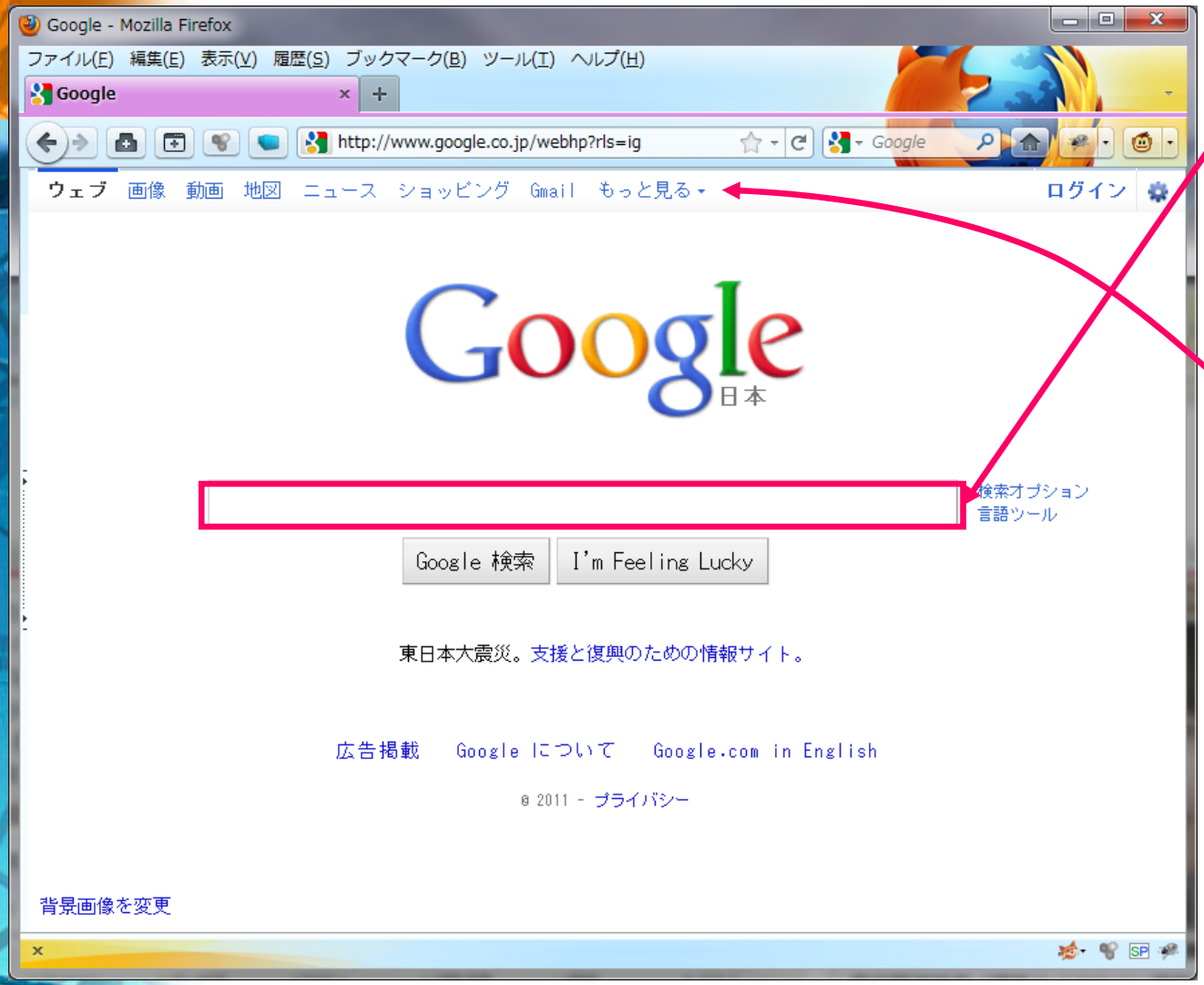
ポータルサイトの実例

- Yahoo!　Japan (Yahoo! Japan)
  - http://www.yahoo.co.jp/
- goo (NTTレゾナント)
  - http://www.goo.ne.jp/
- msn (Microsoft)
  - http://jp.msn.com/

- 参考：　Google(Google,Webページ検索のみ)
  - http://www.google.co.jp/

検索型Webページ情報サービス(1/3)

- 例：Google (グーグル)
  (http://www.google.co.jp/)



- 検索したい事柄を表す語句を入力すると，関連するWebページのリストを得られる.
- Web文書だけではなく，画像の検索もできる.
- 「検索エンジン」と呼ばれることが多い。

検索型Webページ情報サービス(2/3)

「横浜国立大学」を検索

画像

Web経由での配布資料用画像非表示

文書

PRESENTATIONPRO.

検索型Webページ情報サービス(3/3)

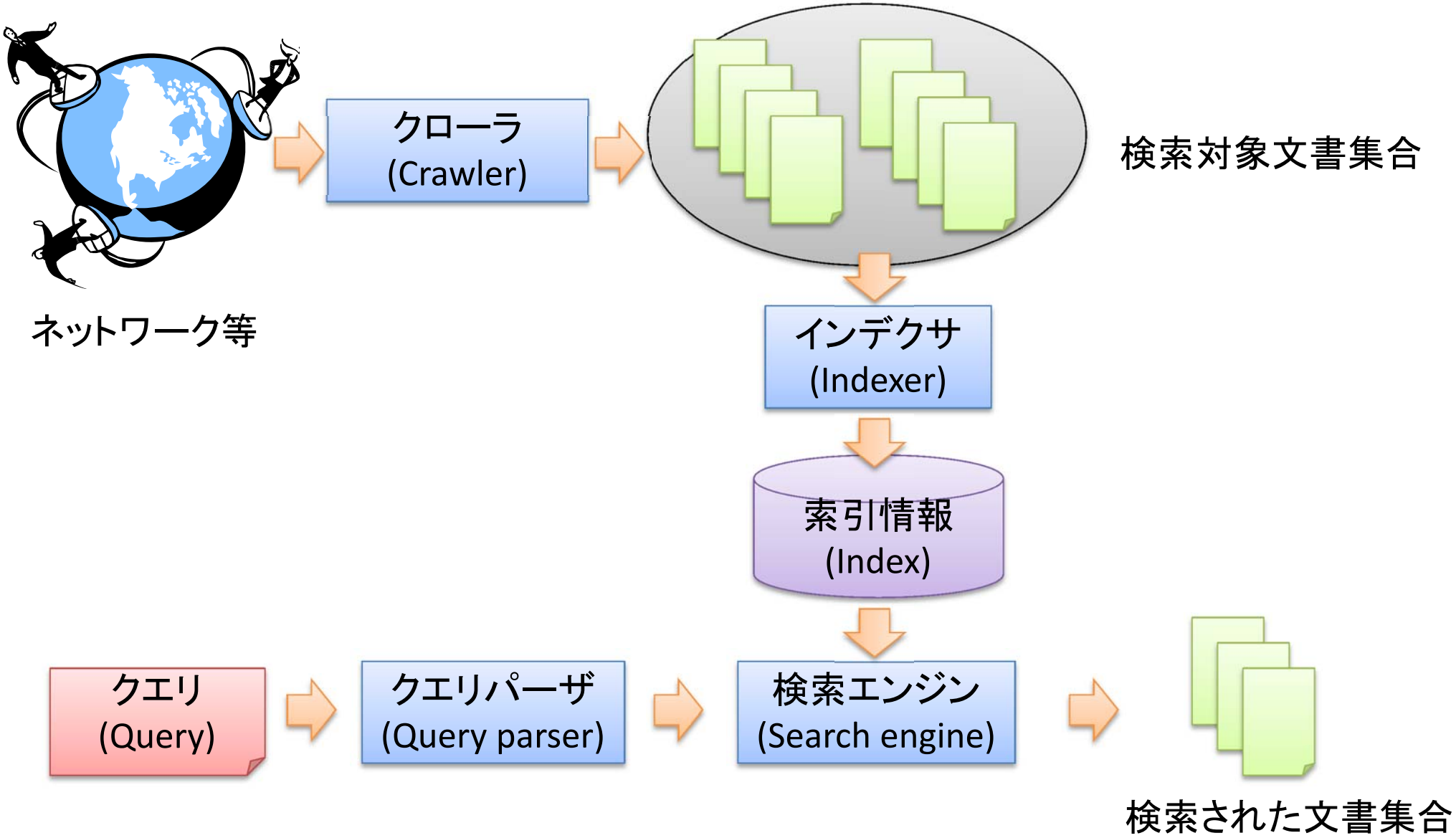
- インターネット上のWebページの情報を集めるのはソフトウェアが自動的に行う.
  - 大規模で網羅的なWebページの収集
  - 検索の精度は，検索方式による
- 登録されるWebページは膨大な量になるが，分野等で分類されているわけではない.
- 知りたい事柄が何であるかが明らかであるときに使用可能。（語句の列で表現できる必要がある）
- 例え話:うまい例えがないが、強いて言えば、商品購入の相談窓口

検索型Webページ
情報サービス

検索型Webページ情報サービスの仕組み

- 例え話
  1. 全世界の通販業者とそれが扱う商品のリストを作成する。
  2. そして、利用者の要望に応じて、良さそうな商品を勧める。
- 基本的な動作
  1. 全世界のホームページを集め、
  2. 集めた情報から利用者の求めるものを提示する
- どうやって情報を集めるか
  - たくさん
  - 高速に
- どうやって利用者の要求に適したページを探すか
  - どのようにして各ページの「良さ」を判断するか
  - 利用者の要求をどのように判断するか

検索型Webページ情報サービスの仕組み



Googleのデータセンタ (1/2)

- データセンター: 各種コンピュータや通信装置を設置・運用することに特化した建物

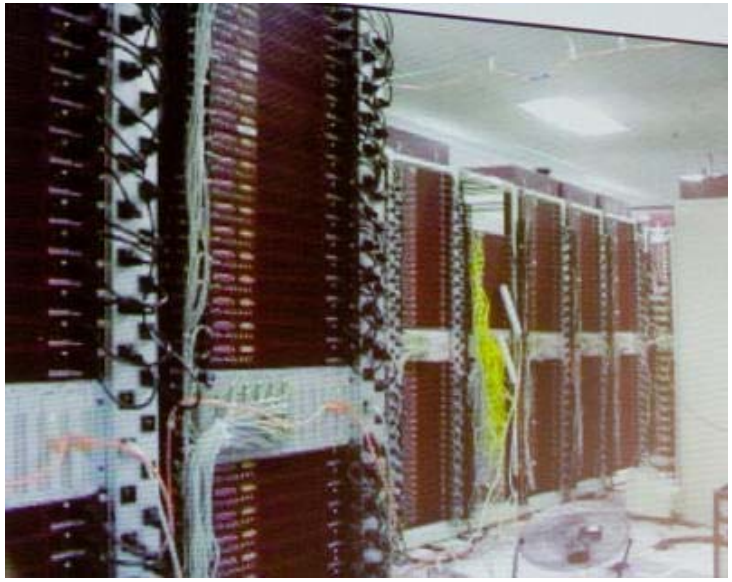
http://www.wayfaring.com/maps/show/48030
http://www.datacenterknowledge.com/archives/2008/Mar/27/google_data_center_faq.html
http://japan.cnet.com/marketing/20374847/

Googleのデータセンタ (2/2)

Lenoir, North Carolina

Zurich, Switzerland

Council Bluffs, Iowa



データセンタ内部

どうやって情報を集めるか

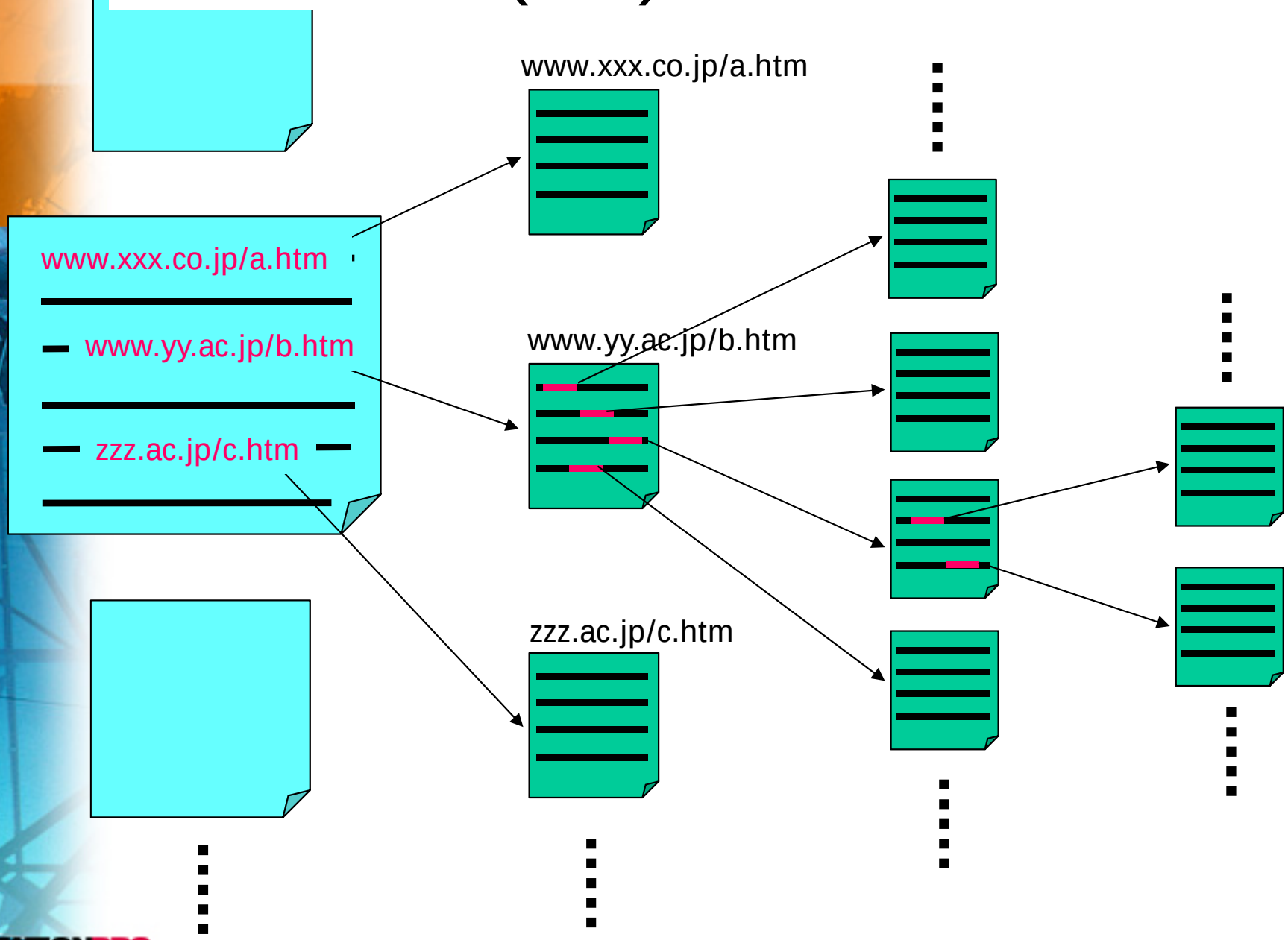
インターネット上のWebページを網羅的に集める(1/3)

- 全自動で収集するには…

1. 最初に幾つかのURLのリストを与える。
   - 多くのWebページへのリンクを持つページが適する。
2. リスト中のURLの内、まだ訪れていないものに対応するWebページ(HTMLファイル)を取得する。
3. 取得したWebページの中にある、URLを取得。URLのリストに追加する。2 へ進む。

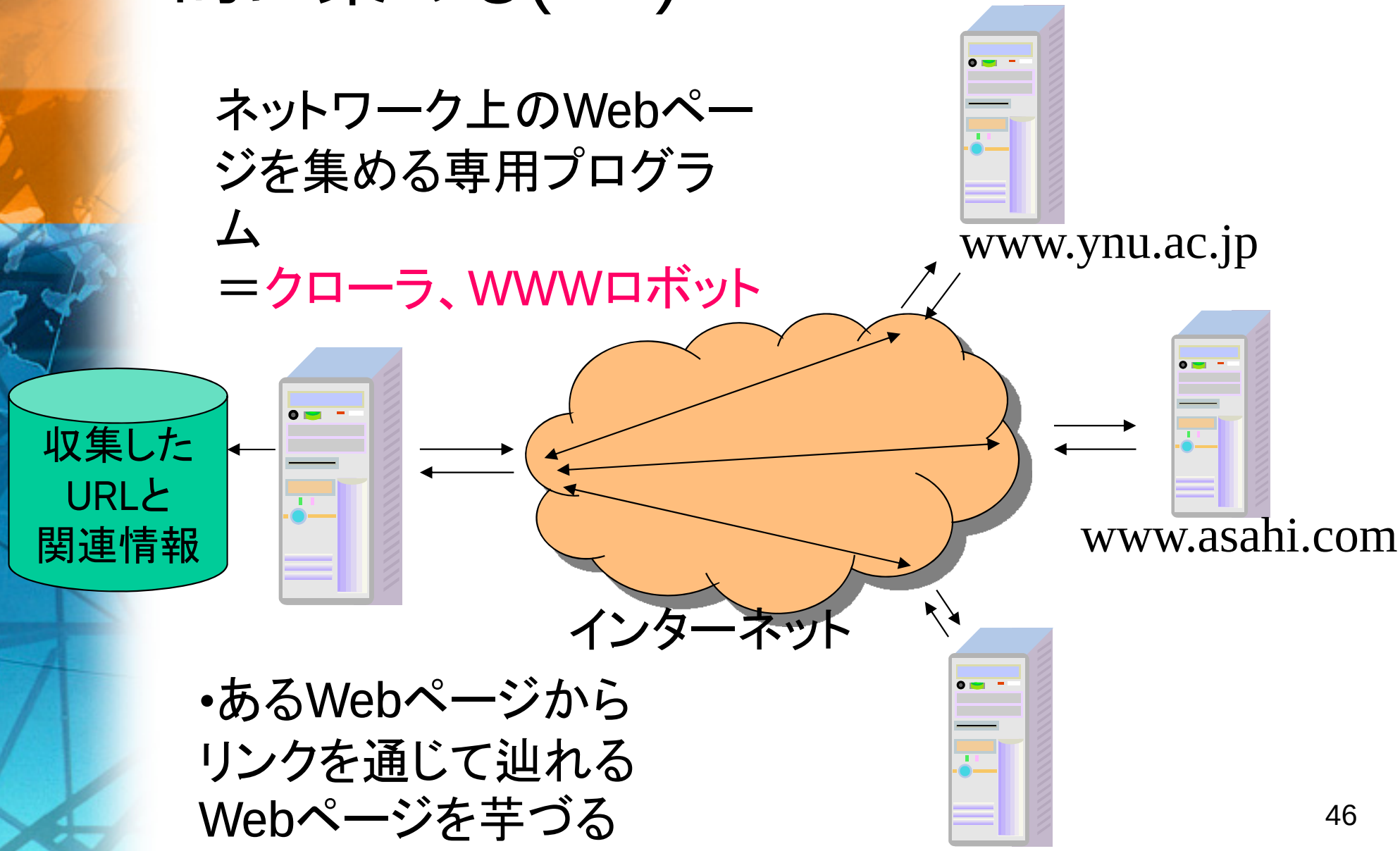
インターネット上のWebページを網羅的に集める(2/3)

初期URLリストに登録されているWebページ



インターネット上のWebページを網羅的に集める(3/3)

ネットワーク上のWebページを集める専用プログラム
＝クローラ、WWWロボット



- あるWebページからリンクを通じて辿れるWebページを芋づる式に端から集める。

クローラが集める情報

各Webページについて、そのURLとともに以下の情報を集める。(どれを使うかは、クローラによる)

- 本文に含まれる語、その語の位置
  - 日本語の場合、語の境界が明らかではないので、「形態素解析」が必要とされる。
  - 形態素解析：単語の区切りを見つけ、品詞や活用型・活用形を調べる技術
- アンカー
  - …で囲まれた箇所（アンカーテキスト）
  - 上記URL
- Title
  - <title>…</title>で囲まれた箇所（タイトル）
- keyword
  - <meta name="keywords" content="w1,w2,…">　における、w1,w2,…（キーワード）
- description
  - <meta name="description" content="紹介文">における、「紹介文」

どうやって
利用者の要求に適した
ページを探すか

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

Webページの重要度
語の分布に基づく方法(1/2)

- TF (Term Frequency)：語の出現回数
- 同じ文書の中で出現回数の多い語ほど重要であると考える。

自動車

鉄道

自動車

ロケット

自動車

- 左の例では、「鉄道」や「ロケット」よりも「自動車」が重要。
- 質問に「自動車」が含まれる場合には、「鉄道」や「ロケット」が質問に含まれる場合よりも、このページの重要度が高くなる。

Webページの重要度
語の分布に基づく方法(2/2)

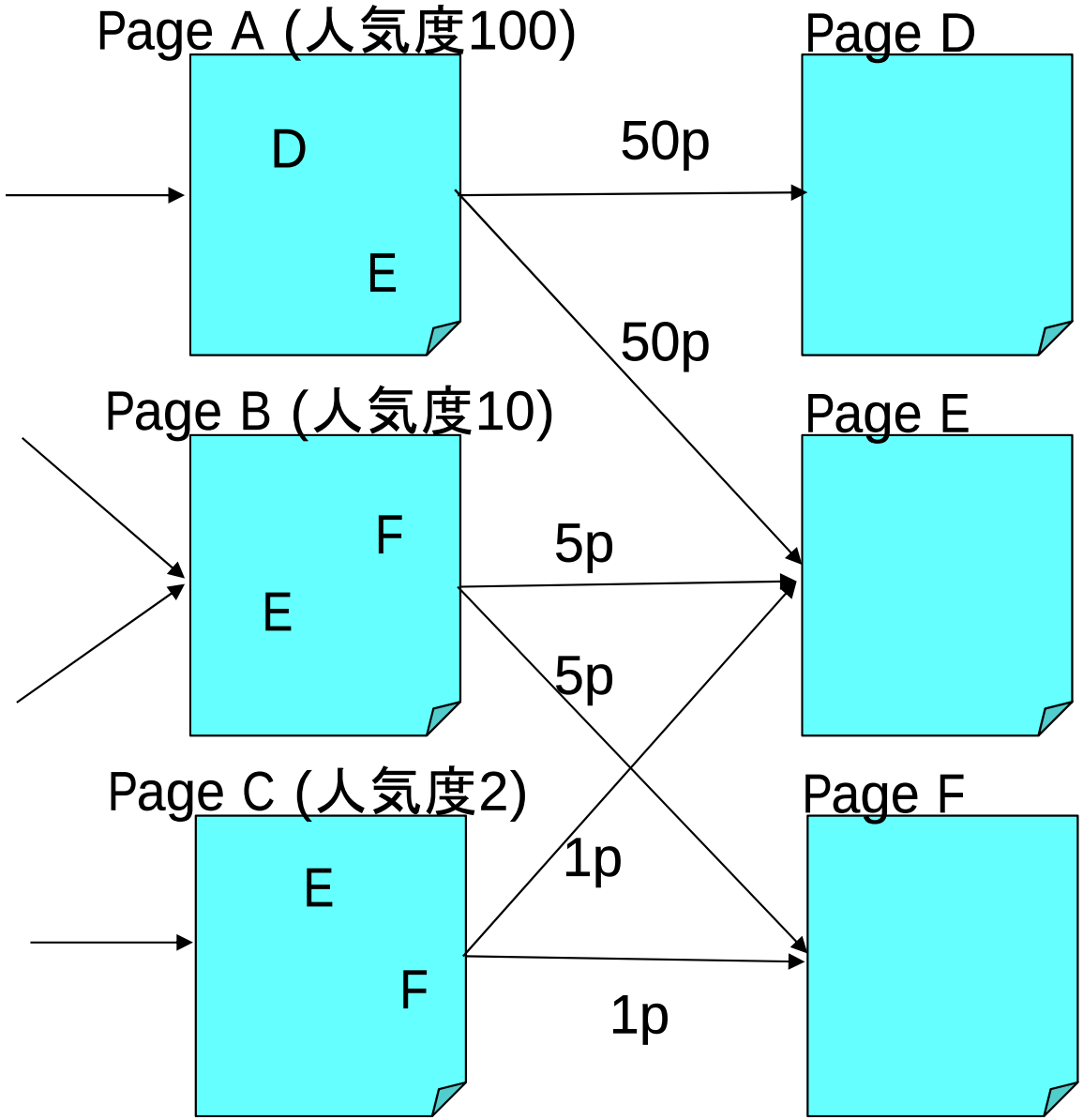
- IDF (Inverse Document Frequency): ある語が登場する文書の数の逆数
- ある文書に偏って出現する語ほど重要であると考える。

- 「果物」「野菜」よりも「レモン」「なす」「トマト」の方が偏って登場しているので、重要度が高い
- N/DF(W) で計算。DF(W)は単語Wの出現する文書数。Nは全文書数。

PageRank法(1/3)

- Googleにおいて、以下のように入力するとそのWebページに対してリンクが張られているWebページがわかる。
  - link:調べたいURL
  - 例：http://www.ynu.ac.jp/　へのリンク
- リンクされた数はそのWebページの「人気度」を表すと考えられる。
- また、人気度が高いWebページからのリンクは、他のページからのリンクよりも人気度に影響を与えると考えられる。

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

まとめ

- 分散処理システムとそのモデル
- World Wide Web
- インターネットの「入り口」
- 検索型サービスの実例　Google
- 検索型サービスの情報収集（クローラ）
- Googleはどうやって「良さそうな」情報を見つけるか？

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。