資料3

ネットワークデータの有効活用

平成17年4月12日

情報フロンティア研究会事務局

全体構成

背景：Web上の情報の氾濫

ネットワークデータの有効活用が課題

◆方策１：データの収集・分析・統合の推進

情報検索・セマンティックｗｅｂ

◆方策２：データからの知見発見・知見の応用

ネットワークデータマイニング

Webの誕生

インターネット：TCP／IPプロトコルでコンピュータネットワークを相互に接続した世界規模の分散型ネットワーク。1969年に米国国防省で開発されたアーパネットが起源

インターネット上のアプリケーションの一つとして、Webが誕生

WWW(World Wide Web)：インターネット上に分散している様々な情報をコンピュータの機種やOSに関係なく多くの人々が共有し、簡単に見られることを目標に作られたシステム。1989年にティム・バーナーズ・リーが考案

《Webの3大構成要素》

URI（URL）：データ（テキスト、画像、音声ファイル等）の場所を指定する書式
HTTP：データの転送規約（プロトコル）
HTML：Webページ上での文書表示書式

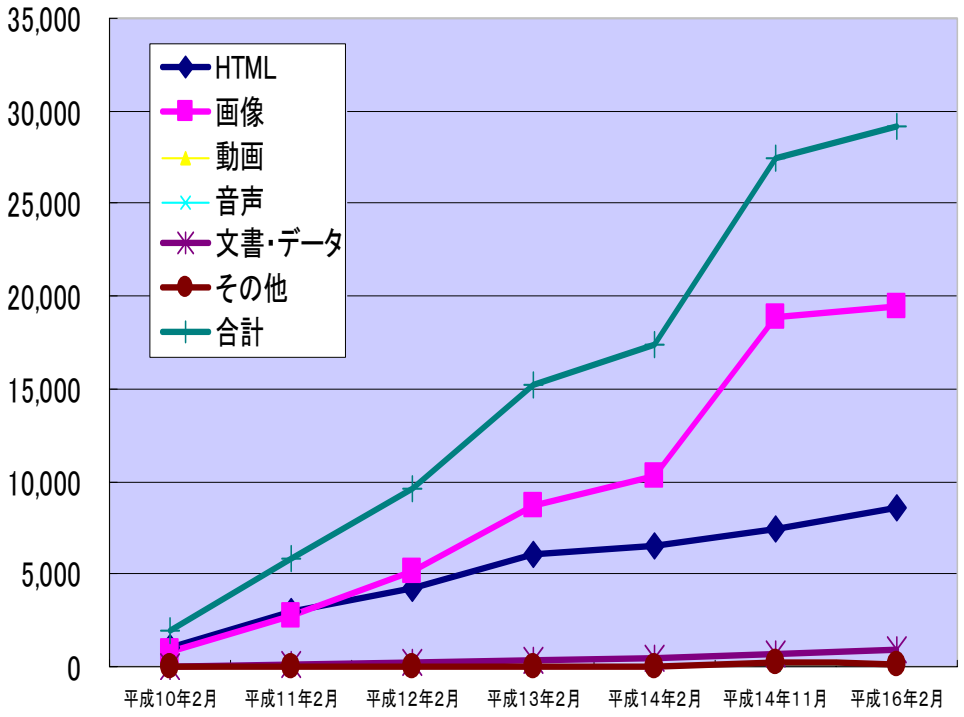
閲覧・作成の容易さにより爆発的に普及
同時に、Web上のデータも爆発的に増大

ネットワークデータの現状

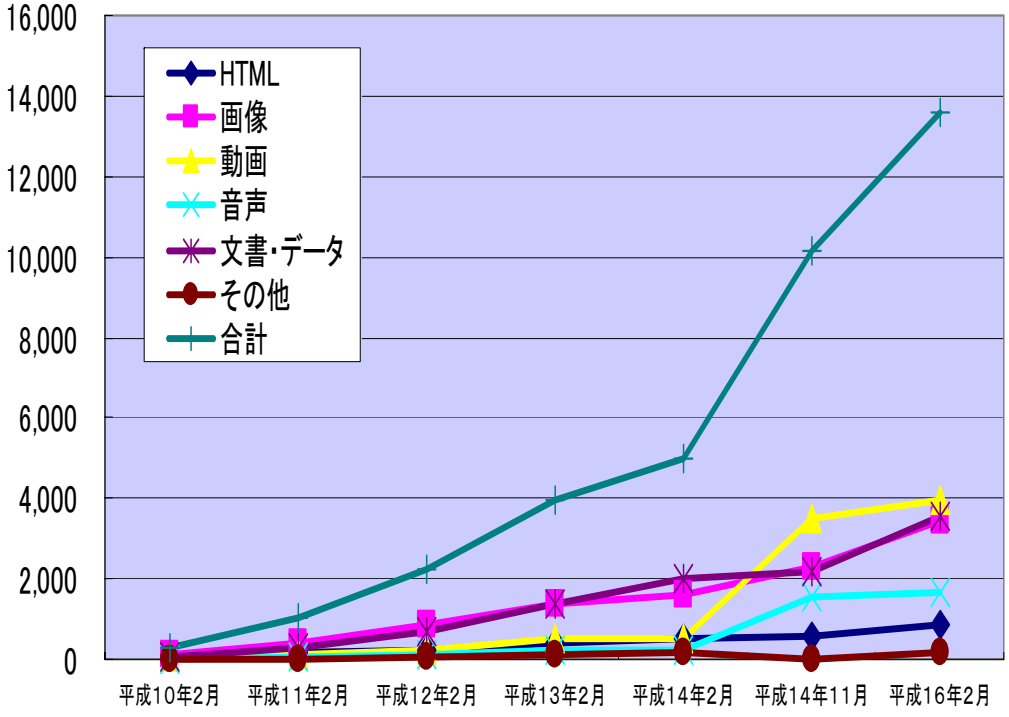
Webから得られる情報は飛躍的に増加

【JPドメインインターネットコンテンツ量のファイルタイプ別の推移】

◆種類別ファイル数の推移



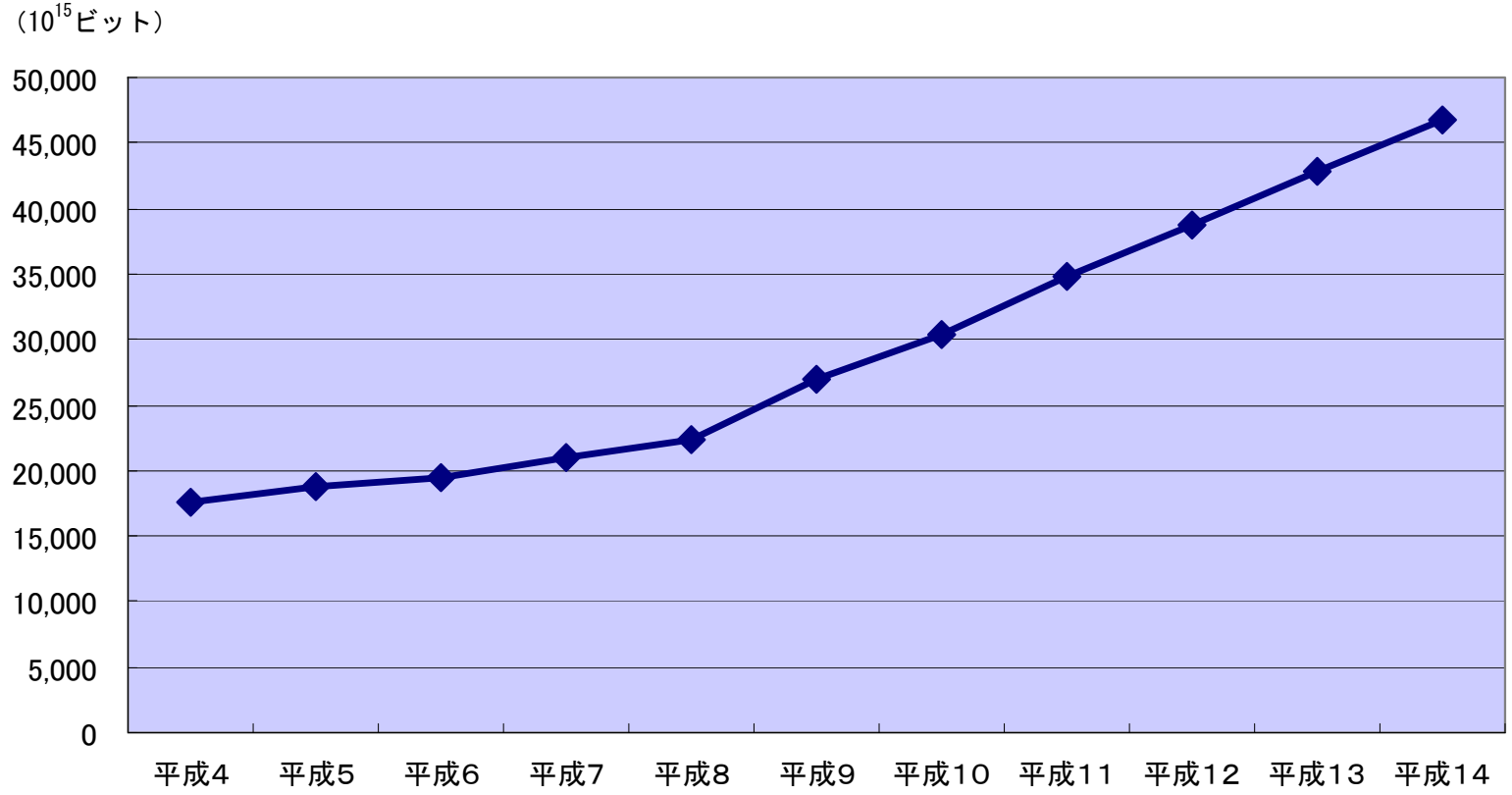
◆種類別データ量の推移



出典：平成16年版情報通信白書

ユビキタスネット社会の実現でこの傾向は加速：情報家電はもとより各種センサーやICタグ等から得られるデータにより、ネットワーク上では膨大なデータが流通

【情報流通量（選択可能情報量）の推移】



注）選択可能情報量：各メディアの情報受信点において、１年間に情報消費者が選択可能な形で提供された情報の総量

出典：平成16年版情報通信白書

検索技術の現状

Web上の情報の増大、Web構造の複雑化に伴い、情報検索機能の高度化に向けた取組みが活発化

【サーチエンジン高度化に向けた主な取組み】

キーワード検索ツールの高度化・高速化

検索インタフェースの向上、検索結果の索引付け・表示手法（ディレクトリ）の開発、エージェントの機能強化、メタサーチエンジンの構築、フィルタリングツールの高度化

複雑な検索システムの開発

構造化データ、半構造化データ、無構造データが混在するWebにおいて高次の問い合わせを行うための言語、システムの開発

マルチメディアサーチエンジンの開発

検索対象データをWeb上のテキストだけでなく、映像等にも拡大するためのシステムの開発

情報検索技術の限界

Web情報の急増

サーチエンジンの機能改善・データベース容量増加が間に合わず、サーチエンジンの再現率（カバー率）や最新情報への更新頻度が低下

収集情報の統一性の欠如

数字データで漢数字とアラビア数字が混在するなど、項目・情報分類が統一されておらず、情報統合が困難

Web構造の複雑化

Webが互いにリンクをはることのできるハイパーリンク構造となっており、情報の所在が動的・分散的であるため、既存のデータベース管理技術の応用に限界

検索者の記憶力の限界

静止的・安定的な利用環境（同じ画面、同じ姿勢、同じ部屋の場所等）により記憶が曖昧になる中で、記憶に基づく情報を検索することは困難

セマンティックWebの登場

セマンティックWeb：コンピュータがWebの情報を機械的に処理することのできる枠組み

■WWWで扱われる情報はコンピュータが自動的に処理できるものではないことから、コンピュータが処理できるデータ（メタデータ）をWeb上の情報に付加

■Webページの内容や関連情報をメタデータとして記述し、さらにメタデータ間の関係を定義辞書（オントロジー）を用いて意味づけ

■メタデータは単純で統一的な記述手法を採用（RDFやOWL)

■コンピュータはハイパーリンクをたどり、用語やルールの定義に基づきデータの意味を論理的に解釈

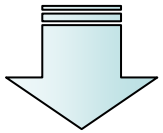
■高度な機能を持つ「知的エージェント」が実現可能

セマンティックWebの将来性

ブログのタイトル・概略をメタデータ化（RSS）し、ポータルサイトの表示内容を自動更新することが可能

Web上における個人情報の管理手法として、個人情報をメタデータ化（FOAF）し、ソーシャルネットワーク（SNS）の組成や円滑な運営に活用可能

Webサービスや情報機器にもメタデータをつけることで、動的なサービス発見や合成といった高度な利用が可能（セマンティックWebサービス）



これらメタデータ間の連携を図れば、更なる新サービスが生まれる可能性が高い

データマイニングの発展

データマイニング：意志決定を目的としてデータ集合から情報や知識を抽出する処理

【データマイニングの発展形態】

データベース管理システム

⇩ 多次元分析

オンライン分析処理（ＯＬＡＰ）

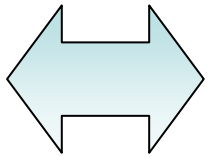
顧客データや購買データをデータベースに格納し、これを様々な角度から検索・集計して問題点や解決策を発見。専門家ではなく経営者層や一般担当者が活用

⇩ 「仮説検証」から「暗黙知の発見」へ

データマイニング

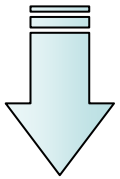
ネットワークデータマイニング

ユビキタスネット社会において、知の創発を育むためには、Web上のデータ等をマイニングすることが不可欠

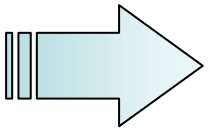


◆情報が構造化されているデータベースと異なり、Webは構造化データ、半構造化データ、無構造データが混在

◆Web上のデータの量は膨大・動的



既存のデータマイニング技術とは異なる技術・手法が発達



ネットワークデータマイニング

Webから価値のある情報を発見・抽出

ネットワークデータマイニングの類型

Webテキストマイニング

■Web上の無構造テキストデータを探索し、意味を導出

テキストドキュメントは量が膨大なうえ、個々のテキスト内にも個別情報が大量に存在することから、データ間の相関ルールの発見、流行傾向の発見・クラスタリング、時間軸を考慮したイベント発見を目的とした開発が主流

Webマイニング

■ユーザがWebにアクセスする記録やリンク情報を探索し、意味を導出

Web利用マイニング

アクセスログからユーザアクセスパターンを発見することにより、eコマースを展開するための顧客のブラウジングパターンの把握が可能
（カスタマイズ広告の作成、マーケティング戦略決定等）

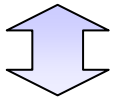
Web構造マイニング

Webはハイパーリンク構造をしていることから、Webドキュメントをノードとし、ハイパーリンクを枝として、節と枝との関係を分析して実用的情報を発見

Web構造マイニング（ネットワーク科学）の世界

基本的な考え方：還元主義との決別

還元主義：この世界の物質や現象は、基本的な物質の性質や運動に細かく分解して調査すれば、全体の性質や運動、変化等が全て説明できるという考え方



システムは、個々の要素と、それら要素間の相互作用に基づいて、初めて理解できる。多数の要素からなる系の性質や振舞いは、個々の要素を足し合わせたものとは全く異なる。組織全体が示すパターンは要素間の相互作用によって生じることが多く、その原因を個々の要素に帰することができない。 → 複雑系

Webの基本的構造である「スケールフリー・ネットワーク」が生物学、社会学、物理学、経済学の領域でも発見されており、Web構造マイニングの成果が学際的に応用可能

ネットワーク科学のICT分野での活用例

現行の活用事例：Web検索技術

Googleの検索エンジンの基礎技術であるPageRank(TM) は、ネットワーク分析技術を活用

PageRankTM は、Webの膨大なリンク構造を用いて、その特性を生かします。ページAからページBへのリンクをページAによるページBへの支持投票とみなし、Googleはこの投票数によりそのページの重要性を判断します。しかしGoogleは単に票数、つまりリンク数を見るだけではなく、票を投じたページについても分析します。「重要度」の高いページによって投じられた票はより高く評価されて、それを受け取ったページを「重要なもの」にしていくのです。
こうした分析によって高評価を得た重要なページには高いPageRankTM（ページ順位）が与えられ、検索結果内の順位も高くなります。PageRankTM はGoogleにおけるページの重要度を示す総合的な指標であり、各検索に影響されるものではありません。むしろ、PageRankTM は複雑なアルゴリズムにしたがったリンク構造の分析にもとづく、各Webページそのものの特性です。（Google：「Googleの人気の秘密」）

今後の活用見込み：アドホック・ネットワークの構成

ノードやリンクがダイナミックに生成や消滅を繰り返すアドホック・ネットワークを構築する際、社会学系で膨大な蓄積のある社会ネットワーク分析（ネットワークの解析技術の導入、様々なノード情報やフロー情報の多元性を加味したモデル化等）を導入することにより、より効率的なネットワーク構築手法の開発が可能

ネットワークデータの活用に係る課題（案）

- セマンティックWeb関連
  - メタデータの付与主体、簡易付与手法
  - メタデータの規格統一が不十分
  - メタデータの付与を促進するためのキラーアプリケーションの不在
  - オントロジーの記述・整備主体
  - メタデータ、オントロジーの信頼性確保

- ネットワークデータマイニング関連
  - 大規模なネットワークデータの解析手法
  - 多元な属性をもつデータ間の関係性の解析手法
  - 動的な関係性構造変化の把握手法
  - ICT以外の、社会ネットワーク分野等との分野横断・学際的な連携
  - セマンティックWebとの連携