情報技術演習
第2回 「情報の構造化と形式化」

2006/10/10

久保田秀和
文学部／情報学研究科
kubota@ii.ist.i.kyoto-u.ac.jp
http://www.ii.ist.i.kyoto-u.ac.jp/~kubota/

講義Webページ

- 一般向け公開ページ
  - http://cresta.kuis.kyoto-u.ac.jp/positlog/itp.html
- 受講者向け内部ページ
  - http://cresta.kuis.kyoto-u.ac.jp/positlog/061010ld.html
  - 閲覧するには別途メールで配布するアカウントが必要です.

演習の日程（変更）

- 端末室のシステム更新予定が講義日程と重ならないことが判ったので，後半の予定を変更します．
  - 11/14　第7回 「情報の統合と共有」
    - Wiki, Blog, Ajax
  - 11/21　第8回 「情報基盤」
    - OSI参照モデル，TCP/IP, IRC
  - 11/28　第9回 「情報流通」
    - SemanticWeb, FOAF, フォークソノミー
  - 12/5　第10回 「コミュニティコンピューティング」
  - 12/12　第11回　最終報告会
  - 2007/1/9　第12回 「インタラクション技術」

本日の講義・演習

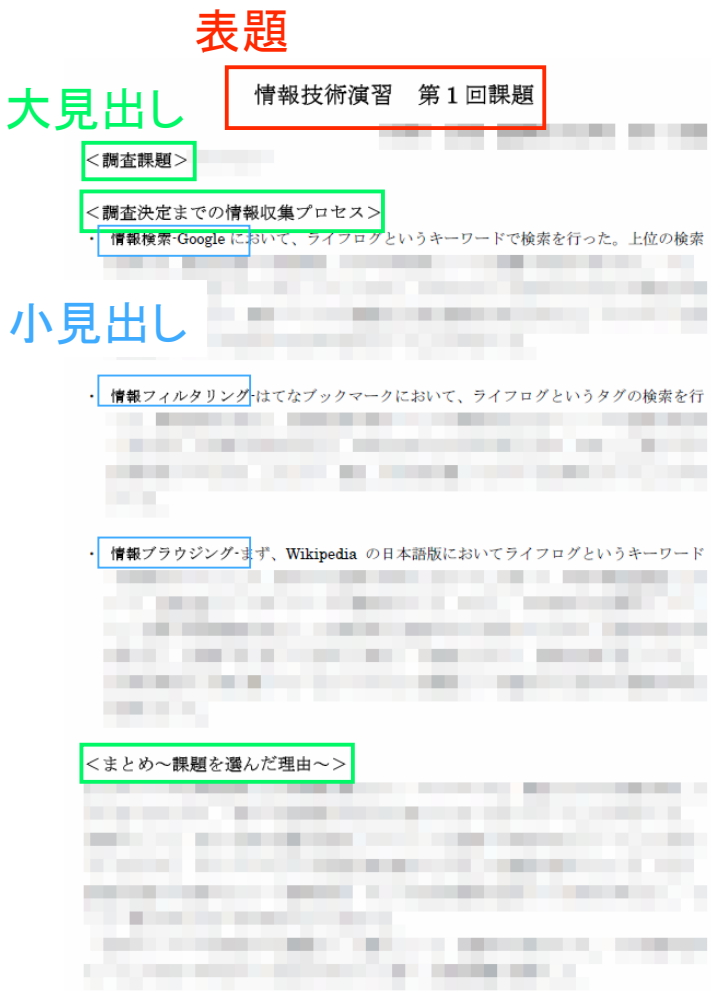
- 情報の構造化
  - 前回提出されたレポートを題材に
- 情報の形式化
  - 形式化の効用
  - HTMLおよびCSSを用いた演習

情報の構造化

- 構造化
  - 情報を構成する要素とその相互関係の明示化
    - 受け手が着目すべき箇所を判りやすくする
  - 情報を効率良くやりとりするための技術
- 構造をどのように定めるか？
  - 一般的な作法，定型に従う
    - 手紙の作法
    - 論文の投稿規定，作法
      - 要素：表題，著者，内容梗概，章，節
      - 親子関係：1章（1.1節，1.2節，1.3節…）
      - 順序：序論→背景→提案手法→実験→議論→結論
  - 送り手（自分）の意図，受け手の意図に沿って考える

構造化の例

- 前回レポート課題のキーワードは
  - 「調査課題の決定」
  - 「選んだ理由」
  - 情報収集における
    - 「情報検索」
    - 「情報フィルタリング」
    - 「情報ブラウジング」
- 右のレポートはこれらのキーワードに基づいた構造が与えられているため，出題者である私にとって読みやすい．



機械可読

- 人にとって理解の助けとなるような構造も，そのままの形では計算機(*1)にとって理解の難しいことが多い
  - (*1) いわゆるコンピュータのこと．情報科学の分野で歴史的にしばしばあてられる語．

情報の空間的な重ね合わせや二次元的な広がりは人にとってアピールする構造である．

ただし，計算機にとっては取り扱いが難しい．

- 計算機にとって処理可能な情報を，機械可読（machine-readable）な情報と呼ぶ

情報の形式化

- 情報を機械可読とするためには，形式化が必要
  - 形式化（formalize）・・・意味や構造を定められた形に沿って表現する
- 形式的な表現の例（機械可読でない場合も含む）
  - 数式
  - 論理式
    - 命題論理　A⇒B　　・・・AならばBである
    - 述語論理　∀x (cat (x)⇒ animal (x))　　・・・猫は動物である
  - データベース
    - 京大図書館OPAC（ http://kensaku.libnet.kulib.kyoto-u.ac.jp/ ）の書誌情報
  - UML（Universal Modeling Language）
    - システム開発プロセスにおいて参与者間の議論を容易にするための標準モデリング言語
    - システムの要件（自然言語）とシステムの詳細（プログラミング言語）の隔たりを埋める
  - マークアップ言語（XML, HTMLなど）
    - 形式の明示的でない情報（主に自然言語文書）に対して，「タグ」と呼ばれる特別な文字列を埋め込むことにより，形式化された意味や構造に関する情報を追加するための言語

本演習シリーズの扱う範囲

形式化の効用

- 人間にとっての効用
  - 抽象的思考や情報共有，役割分担の助けとなる
- 計算機にとっての効用
  - プログラミング言語やデータ構造など実装に依存した処理が可能となる
    - ただし，計算機向けの形式は人間の知識の本質的な内容を理解したり伝達したりするためには適さない

形式化のアプローチ(1)

- 人の知識を機械可読な形式で記述するには？
- 知識ベース（knowledge-base）システム（1960年代～）
  - プロダクション規則に基づく推論を中心とした，エキスパートシステム
    - 専門家から引き出した，「こういうときはこうする」という行動規則に関する知識
  - 例）
    - 有機化合物の分子構造推定
      - Edward A. Feigenbaumら, Dendralシステム, 1965-
    - 感染症の診断・治療支援
      - Edward H . Shortliffe, “MYCIN: Computer-Based Medical Consultations”, 1976.
      - 問診形式のインタラクション技術を含む

練習

- Webで触れることのできるエキスパートシステム
  - **かぜ症候群診断支援「Dr.Kaze」**
    - http://www.wind.ne.jp/hassii/expert/index.htm
    - ハイパーリンクによる分岐がそのまま素朴に知識を表現している
  - **金沢市観光計画支援システム「まわるまっし金沢」**
    - http://www.fuji.ne.jp/~kissy/kanazawa/index.html
- ほかにもいろいろ探してみましょう

形式化のアプローチ(2)

- 事例ベース推論（Case-Based　Reasoning）システム（1980年代～）
  - Roger C. Schank, Dynamic Memory, 1982
  - 過去における類似する事例の検索と修正による推論
    - 規則として抽出することが難しい，経験的な事例
- 知識メディア
  - Mark Stefik, “The Next Knowledge Medium”,1986
    - 知識の生成，流通，利用のライフサイクルを促進するメディアを構築することによって，人海戦術的に知識を獲得
- WorldWideWebの公開（Tim　Berners-Lee,　1991-）を経て，SemanticWeb（第9回演習で詳説）へ
  - Webが知識メディアの性格を担うようになってきた
    - 「20Q」（トゥエンティーキュー） http://www.asovision.com/20q/top.html
      - 知識ベースの概念に基づくおもちゃ．Webを介した人海戦術的な推論精度向上の身近な例として．
      - http://www.itmedia.co.jp/news/articles/0511/25/news038.html
  - Web上の大量のコンテンツを知識源として利用可能にする

本演習シリーズの扱う範囲

形式化のアプローチ(3)

- ただし，人の知識は形式化という方法では捉え切れない
  - マイケル・ポラニー「暗黙知の次元」(Michael Polanyi, “The Tacit Dimension”), 1966
    - 暗黙知（⇔形式知）
    - 「我々は語ることができるより多くのことを知ることができる」
  - Knowledge Management
    - 人の活動性も視野に含めた知識創造・知識管理支援
    - 野中郁次郎,竹内 弘高「知識創造企業」（1995，邦訳1996）
  - （関連）

    Community Computing，Community Support System
    - 人の活動支援のほうにより重点（第10回演習で詳説）

マークアップ言語

- HTML（HyperText Markup Language）
  - 仕様書（特徴，定義，形式について）は
    W3C, “HTML 4.01 Specification”（http://www.w3.org/TR/html401/）
  - **What is HTML?**
    - To publish information for global distribution, one needs a universally understood language, a kind of publishing mother tongue that all computers may potentially understand. The publishing language used by the World Wide Web is HTML (from HyperText Markup Language)
      - WWWにおいて情報を頒布のための共通言語
    - HTML gives authors the means to:
      - Publish online documents with headings, text, tables, lists, photos, etc.
        - » 情報流通
      - Retrieve online information via hypertext links, at the click of a button.
        - » 情報検索
      - Design forms for conducting transactions with remote services, for use in searching for information, making reservations, ordering products, etc.
        - » （人と人，人と計算機の間の）インタラクションの窓口
      - Include spread-sheets, video clips, sound clips, and other applications directly in their documents.
        - » 情報統合

HTML文書の例

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01
Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd">
<html lang="ja-JP">
<head>
<meta http-equiv="Content-Type"
content="text/html;charset=UTF-8">
<title>My first html</title>
</head>
<body>
 <h1>1章. 概要</h1>  ・・・
 <h1>2章. はじめに</h1>
  <p>段落1</p>
  <p>段落2</p>
 <h1>3章. 提案</h1>
  <h2>1節. 提案1</h2>
   <p>段落1</p>
   <h2>2節. 提案2</h2>  ・・・
   <p>段落1</p>
 <h1>4章. おわりに</h1>  ・・・
</body>
</html>
```

文書型宣言

表題

大見出し

段落

中見出し

タグと呼ばれる，<>記号で囲まれた特別な文字列を用いて，構造を形式的に記述する．その形式は文書型宣言において示されたDTD（後述）よって定められている．

本例では文書を構成する表題・大見出し・中見出し・段落などの要素が明示化されている．

IEならば「表示」→「ソース」で閲覧可能

HTMLの効用

- 計算機にとって，文書に込められた製作者の意図を汲むことが容易となる
  - よりHTMLに即して言い変えると，「多種多様なユーザエージェントにとって，タグの定める形式を手がかりとして文書を理解することが可能となる」
  - HTML User Agent（HTML 4.01 Specificationより）
    - HTML文書を解釈するすべての装置　User agents include visual browsers (text-only and graphical), non-visual browsers (audio, Braille), search robots, proxies, etc.

HTMLにおける形式化（1）

- HTMLはどの程度形式的に定義されているか?

(I)文書型定義（DTD：Document Type Definition）を用いた，要素と構造に関する厳密な形式化

例）順不同リスト（Unorderd lists）

・項目(a)
・項目(b)
・項目(c)

表示例

```
<ul>
<li>項目(a)</li>
<li>項目(b)</li>
<li>項目(c)</li>
</ul>
```

マークアップ例

```
<!ELEMENT UL - - (LI)+ -- unordered list -->
```

文書型定義から抜粋.
「UL要素はLI要素を1つ以上内容に持つ」（要素間の親子関係のあり方を形式的に定めている）

HTMLにおける形式化（2）

（II）自然言語を用いた解釈の余地を残した定義

- 例）「リンクされたリソース（データやソフトウェア）への訪問」
  - The default behavior associated with a link is the retrieval of another Web resource. This behavior is commonly and implicitly obtained by selecting the link (e.g., by clicking, through keyboard input, etc.).

- 形式化には妥当な程度がある

課題2-1：レポートの構造化，形式化

- 前回のレポートをHTMLを用いて構造化，形式化してください．
- HTMLの記法についてはWebを調査し，定められた仕様のいずれかに沿ってください．
- （理解できる人はXHTMLを用いても構いません）

問題の分割・分業

- 形式化による大きな効用の1つとして，問題の階層や単位が明示されることにより，分業が容易となる点がある.
- 構造と外観との分割
  - HTMLとCSS（Cascading Style Sheets）
- アプリケーションと外観・インタラクションとの分割
  - XUL（User Interface Language）
  - XAML（Extensible Application Markup Language）

課題2-2：CSS

- Operaを用いて文学部のページの外観を変えてみてください.
  - http://www.bun.kyoto-u.ac.jp/index-j.html
- 前回レポートの外観をCSSを用いて定めてください.
- CSSの記法についてはWebを調査し，定められた仕様のいずれかに沿ってください.

第2回課題：文書の構造化・形式化

- 前回のレポートをHTMLを用いて構造化・形式化し，CSSを用いて外観を与える.
- 各自の調査内容も可能な限り追加すること．前回の分と合わせ，A4用紙に印刷しておよそ2ページ以上となるようにする.
- 期限は10月14日（土）17:00
- レポートはCSSとHTMLを用いて記述された文書として作成する.
  - ファイルがひとつの場合はそのまま電子メールに添付して提出
  - ファイルが複数の場合はlzh/zip/tgzのいずれかの形式でアーカイブし，電子メールに添付して提出
  - レポートをWebページとして作成し，URLを電子メールで連絡する方法でも構いません
- あて先は　久保田　kubota@ii.ist.i.kyoto-u.ac.jp

おわりに

- 参考図書
  - 西尾，太田，横田，西田，佐藤「情報の共有と統合」，岩波講座マルチメディア情報学7（岩波書店）
  - Sinan Si Albir「入門UML」（O’REILLY）
  - 杉山，長田，下嶋編「ナレッジサイエンス」（紀伊國屋書店）
- 予定していた，形式化と意味，メタ情報（RSS等）の話は第9回へ回します