Vol.2012-NL-209 No.12
2012/11/23

【招待講演】電子図書館と自然言語処理

長尾　真[†]

[†]京都大学

電子図書館と自然言語処理

長尾　真
京都大学名誉教授
2012年11月23日

Ⅰ 書籍の形態の歴史的変遷

Through knowledge we prosper

2

書籍形態の変遷

時代	内容	媒体	道具・手段	量	形	特徴
古代	文章	石板、粘土板	のみ、へら	1枚	板	1～2次元表現
<グーテンベルグ>	文章 図	竹簡、木簡、パピルス、羊皮紙	筆と墨筆写	1組	巻物	
		紙	版木	多数冊	冊子(頁という概念)	
	文章、図、写真		活字印刷	ぼう大な冊数	本(目次、索引)	
ディジタル時代(フェーズ1)	文章、図、写真 音、動画像、イメージ	電子読書端末	キーボードスキャナー電子表示	任意冊数	電子読書端末にぼう大な数の本が入れられる任意の本の欲しい部分を取り出せる検索機能	3～4次元表現
ディジタル時代(フェーズ 2)		著者と読者の間のやりとりの出来る機能をもった電子読書端末	電子ペンや音声による入力機能			

3

- 日本においては1994年頃に電子読書端末が発売されたが、コンテンツがほとんどないなどのことから普及しなかった。
- 4、5年前から何万という短い単純な小説が携帯電話端末のために作られ、若い人達に読まれて来ている。
- その中にはベストセラーとなったものもあり、それらは本として印刷出版された。

電子読書端末

4

- 米国でKindleやReader、iPadなどが発売され、これらは日本にも入って来ている。
- 最近はスマートフォンに移行しつつある。
- それに伴って国内の出版社・印刷会社などが10万冊を越える電子コンテンツをそれぞれの流通プラットフォームから発売しはじめた。

5

電子読書端末と電子書籍
(マルチメディア書籍)

- 電子読書端末は文字のほかに音や映像が扱えるマルチメディア端末である。
- 電子書籍は文字だけでなく図、表、音、動画像などマルチメディアの著作物となってゆくだろう。
- このように紙の本では表現できないことが電子読書端末を通じて表現できる。

6

（インターラクティブ書籍）

・読者が端末表示画面に指やペン、音などによって意図を伝達することができるようになる。

- したがって読者と電子書籍との間で対話の関係が生じる。
- 端末装置はネットにつながっているから、読者は他の読者と共同で読書をしたり、著者などと対話できることになり、著作活動、読書活動に新しい世界が開かれる。

7

Ⅱ　理想の電子図書館を目指したアリアドネ（Ariadne）

10

電子書籍革命

- グーテンベルグ革命は印刷術における革命であった。
- 電子書籍の革命はコンテンツ表示の革命であるだけでなく、人間の表現できる内容が文字・図形・写真から、音や映像などの多次元世界に拡大されたこと、また読者が反応することが出来ることになり、全く新しい世界が展開される大きな影響力のある革命である。

8

日本で最初の電子図書館

- 1990年に計画を立ち上げ1994年にプロトタイプを完成（京都大学長尾研究室）。
- 書物の構造化と種々の強力な検索システム。
- 電子読書において、しおり挿入機能、メモ記入機能、自動読みあげ機能、日英・英日機械翻訳機能等を実現。
- これらの成果を「電子図書館」（岩波書店、1994）という本にまとめて出版した。2010年3月にその新装版が出た。

11

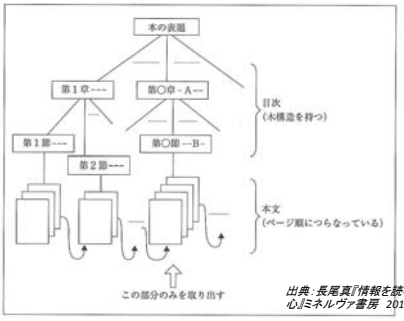
電子読書端末の特徴

<table>
<tr><th>時代</th><th>内容</th><th>媒体</th><th>道具・手段</th><th>量</th><th>形</th><th>特徴</th></tr>
<tr><td>古代</td><td>文章</td><td>石板、粘土板</td><td>のみ、へら</td><td>1枚</td><td>板</td><td rowspan="4">1～2次元表現</td></tr>
<tr><td rowspan="3"><グーテンベルグ></td><td rowspan="2">文章
図</td><td>竹簡、木簡、パピルス、羊皮紙</td><td>筆と墨筆写</td><td>1組</td><td rowspan="2">巻物</td></tr>
<tr><td rowspan="2">紙</td><td rowspan="2">版木</td><td rowspan="2">多数冊</td></tr>
<tr><td>文章、図、写真</td><td>冊子
(頁という概念)</td></tr>
<tr><td>ディジタル時代
(フェーズ1)</td><td rowspan="2">文章、図、写真
音、動画像、イメージ</td><td>電子読書端末</td><td>キーボード
スキャナー
電子表示</td><td rowspan="2">任意冊数</td><td rowspan="2">電子読書端末にぼう大な数の本が入れられる
任意の本の欲しい部分を取り出せる検索機能</td><td rowspan="2">3～4次元表現</td></tr>
<tr><td>ディジタル時代
(フェーズ 2)</td><td>著者と読者の間のやりとりの出来る機能をもった電子読書端末</td><td>電子ペンや音声による入力機能</td></tr>
<tr><td>活字印刷</td><td>ぼう大な冊数</td><td>本
(目次、索引)</td></tr>
</table>

9

書籍の構造化と検索

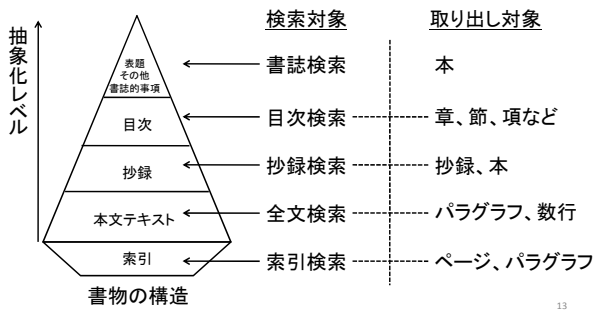
- 電子書籍は目次や索引によって構造化することができる。



出典：長尾真『情報を読む力、学問する心』ミネルヴァ書房　2010年　p.137

12

- 電子書籍に対しては種々の検索をすることができる。検索出力の単位は書物、書物の章や節あるいはパラグラフなど任意の単位となる。



13

- 連想機能をもつ種々の検索システムによって必要とする情報・知識を取り出せるようになるだろう。
- この知識システムは一種の百科辞典とみることもでき、これに検索質問を出すことによって書誌情報でなく質問に対する答が取り出せることになるだろう。これは事実検索である。

16

書籍の解体と再編成

- 構造化された書籍の検索によって必要な部分だけを取り出すことができる。
- すなわち書籍は部品に解体され、必要なところだけを取り出せる。
- いろんな書籍の必要なところを取り出し自分の筋書きにしたがって並べなおし、新しい著作物を作ることができる。

14

Ⅲ　国立国会図書館（NDL）における電子図書館

知識ネットワークの構築

- 全ての書籍を部品に解体し、種々の因果関係によって部品同士をリンクすることができる（ハイパーテキスト構造）。
- 因果関係としては同義/類似関係、反義関係、上位下位関係、原因結果関係、全体・部分関係などいろいろのものが考えられる。
- こうして電子図書館を人間の頭脳内の記憶のように、知識のシステムの形に構成することができるだろう。

15

電子図書館の大切さ

- NDLは来館者だけでなく、全ての国民に同等のサービスを提供することが理想である。
- そのためには図書館資料を全て電子化して利用者に送信できることが必要である。
- 図書・資料を電子化すれば、冊子体での図書館サービスよりもはるかに優れた知的で柔軟なサービスを実現できる。

18

資料のディジタル化における問題

- 著作権のある図書のディジタル化は著作権者の許諾を必要とする。
- 著作権者を探し出すのに時間と費用がかかる。
- 著作権者不明の図書が多い（孤児出版物）。
- 孤児出版物は著作権者が出て来た時に著作権料を払うことにして、ディジタル化し、配信できるようにするべきである。

19

著作権法改正（1）

- 国立国会図書館における特例

国立国会図書館においては、図書資料保存の目的で許諾なく図書資料のディジタル化をすることができる（この場合のディジタル化は利害関係者との話し合いでディジタルイメージであって、文字化ができないことになっている）。

著作権法改正（2）

- 研究開発のための特例

（情報解析のための複製等）
第四十七条の七　著作物は、電子計算機による情報解析（多数の著作物その他の大量の情報から、当該情報を構成する言語、音、影像その他の要素に係る情報を抽出し、比較、分類その他の統計的な解析を行うことをいう。）を行うことを目的とする場合には、必要と認められる限度において、記録媒体への記録又は翻案を行うことができる。ただし、情報解析を行う者の用に供するために作成されたデータベースの著作物については、この限りでない。

著作権法改正（3）

- 障害者のための著作権法の改正

・これまでの点字図書館から、政令で定める図書館（国立国会図書館をはじめ、公共図書館等）にまで広げて図書のディジタル化による提供が可能となる。

・図書館間でのデータ送信、図書館から視覚障害者の方々への送信が可能となる。

・将来、視覚障害者の範囲が発達障害者等に拡大されることが期待される。

NDLにおけるディジタル化の現状（1）

- 国会会議録(戦後)は全て文字テキスト化し、種々の検索が可能。
- 帝国議会会議録は全ての資料がディジタル画像データとして読むことができる。
- 日本法令索引で1867年以降の全ての法律、条約等が検索できる。

23

NDLにおけるディジタル化の現状（2）

- 資料のディジタル化のために補正予算127億円を獲得。
- 1968年までの図書、雑誌、博士論文、官報、古典籍等を2009年度から2年間でディジタル化した。
- ディジタル化はイメージのレベルであり、文字化は出版界の強い反対により出来ない。

24

国立国会図書館 所蔵資料のデジタル化の状況（平成23年8月末）

資料種別	所蔵数（H22年度末）(A)	デジタル化実施済*1 (B)	デジタル化未実施 (A-B)	実施割合 (B/A)
古典籍	29万冊	7万冊	22万冊	1/4
和図書	427万冊	88万冊	339万冊	1/5
和雑誌	455万冊	101万冊	354万冊	1/5
博士論文	39万冊*2	14万冊	25万冊	1/3
合計	950万冊	210万冊	740万冊	1/5

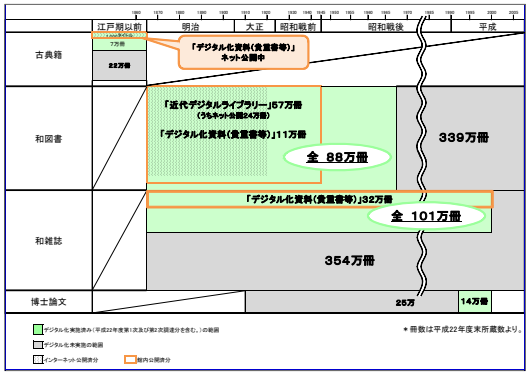
*1：デジタル化実施済刊行年代は次のとおり。
【古典籍】江戸期以前
【和図書】明治期～1968年刊行
【和雑誌】明治期～2000年刊行（商業出版との調整タイトル等を除く。）
【博士論文】平成3（1991）年度～平成12（2000）年度受入れ
*2：平成12年度までの所蔵数。平成13年度以降は各大学においてデジタル化することになっている。

25

NDL資料の全国配信

- 著作権法を改正し、市場で容易に入手できない書籍等はNDLから公共図書館、大学図書館等に配信できるようにした。
- これは平成25年の早い時期から実施される。
- 著作権の存在する資料のこのような広域配信は世界で初めてである。

国立国会図書館所蔵資料のデジタル化の状況（平成23年8月末）



26

Ⅳ　Web情報の収集

ネット上で読めるように するための努力

- ディジタル化された資料は館内の端末でのみ見ることが可能、同時に見られるのは蔵書冊数以下。
- 著作権者から許諾を得る作業も進めており、許諾の得られたものは、順次ネット上で公開してゆく。その場合、利用は自由である。

27

インターネット上の情報

- born-digital 情報の存在。
- 国の文化財として収集・保存・活用すべきもの。
- Webサイトはどんどん開設されるとともに、消えていくものも多い。
- 安定的に開設されているWebサイトも内容は常に変化している。
- これらを常に追跡して保存することが大切。

30

国立国会図書館における Webアーカイビング

- WARP（Web Archiving Project）と略称。
- 2002年から約2700サイトのWeb情報と、約2000タイトルの電子雑誌の収集を行っている。
- イベント、町村合併等、消去の可能性が高いWebサイトを優先。

31

WEBサイトの収集

・国立国会図書館法を改正し、国、地方公共団体、国公立大学、独立行政法人等のWebサイトを許諾なく収集できるようにした。
・深層Webで収集のできない部分については送ってもらう。
・収集したWebサイトの情報は、許諾を得てインターネット上に公開する。
・許諾の得られないものはNDL内でのみ利用可能。

32

Ⅴ　電子書籍時代における図書館の諸問題

Through knowledge we prosper

33

電子納本の大切さ

- 出版物を中心とする日本の文化創造活動は収集、保存され、新しい創造のために利用されるべきである。
- 日本の出版物は納本制度により国立国会図書館で収集され、利用に供されている。
- 電子出版物についても電子納本、保存され、利用に供されるべきである。
- 電子出版物とは？

34

MARC作成の自動化

- MARC（Machine Readable Catalogue）
- 図書に詳細なMARCを付けるには知識・経験が必要で時間がかかる。
- MARC付与作業はコストが高い。
- 電子図書の場合、どこまで自動化できるか。
- Web情報の場合にどのような書誌情報を付与するか（ダブリンコアに基づくメタデータ付与）。

35

書誌情報はどのように必要か

- 図書を分類するという考え方
- 分類の詳しさ
- 全文検索などが出来る時代における分類、書誌情報の意味をどう考えるか

36

図書館は何を集めるべきか

- 出版物とは何か
- 図書、小冊子、逐次刊行物、楽譜、地図、映画フィルム、複製文書・図面、レコード
- 電磁的手段による文字、映像、音またはプログラムを記録した物
- その他NDL法に規定していないもので集めるべきものは？

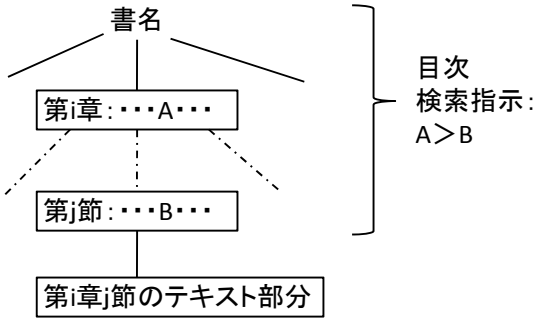
  脚本、アニメ、ゲーム、パンフレット、広告類？

  Net上のBlog、Twitter、Facebook、．．．．．．？

・ラジオ、TV放送データの収集

37

目次の階層構造検索



40

多様な検索の可能性

- 過去の全ての出版物を容易に調べられる。
- コンテンツの全文検索、スニペット表示などで書物の立ち読みにあたることが出来て、欲しい書物を入手できる。
- 種々の検索が可能となり、冊子単位だけでなく、冊子の中の章、節、項や頁、パラグラフ、文や図など、種々の単位で取り出せるようにすることが必要である。

38

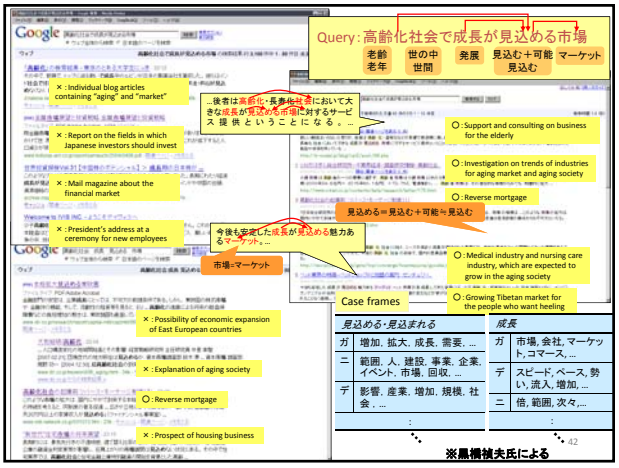
IBMのワトソンがクイズ番組で勝利

- 2011年2月ワトソンと称するIBMの質問応答システムがクイズ番組 Jeopardy! の王者と対戦、勝利した。
- このシステムは本や辞書、百科事典、ウィキペディアなど、2億ページのテキストデータ（約100万冊相当）を記憶し参照した。
- 構文解析、情報検索、機械学習、知識表現と推論、問題解決など、種々の手法を用いている。対話的要素も含んでいる。
- ハードウエアは2880個のCPUを並列に動かして人に負けない速さで回答を出した。

テキスト検索の種々

- 書誌的事項、キーワード等による検索。
- シソーラス等を用いたあいまい検索、連想検索。
- 目次の階層構造検索。
- 全文検索、コンテキストを用いることによって、パラグラフ単位の検索が可能。
- 自然言語を用いた質問。
- 対話型検索。

39



42

情報の信頼性、信憑性の検証

- グーグル検索の問題点
- 取り出された情報が学問的知識と比較して矛盾していないか（信憑性）
- 取り出された情報の内容、文体、発信者の属性などが信頼できるものであるか（信頼性）
- グーグル検索でトップあたりに来る情報と正反対の情報がロングテールのどこかに存在しないかどうか

43

Ⅵ　知識システムの構築

知識社会の時代

- 物の時代から情報の時代へ
- 量から質の時代へ
- 知識が富を生み出す
- 製造から、設計、世界標準、知的所有権へ

45

知的活動に集中する

- 地球環境、エネルギー、資源、廃棄物、人口減少などの問題を抱える中で、日本が生産性を上げ、世界のトップグループに残るためには知的労働に集中することが必要。

46

- 情報価値、知識付加価値の高いものに集中すること（企画、設計、先端技術、知識産業、情報産業、コンテンツ産業、メディア産業、芸術、など）。
- そのために情報の網羅的収集が必要。

47

メディア文化財についての課題

- 本以外の情報メディアの重要性。
  写真、地図、パンフレット、・・・
  演説、語り、歌謡、音楽、ラジオ放送、・・・
  CD、DVD、TV放映、映画フィルム、・・・
  舞台芸術など
- これらのうち、パンフレット、CD、DVD、TV等は公共的立場からの本格的な長期保存の目的でのアーカイブはほとんどなされていない。
- TVプログラムなどは台本テキストと対にして保存することが大切。

48

文化財のディジタル保存

- 絵画のディジタル化と保存・再生（美術館）
- 3次元物体のディジタル計測と再現（博物館）
- 遺跡のディジタル記録と再現（東大 池内克史教授）
- 無形文化財（たとえば踊り）の3次元記録と再現（京大 松山隆司教授）
- インターネット上の情報の保存、Blog、Twitter等の情報

49

ディジタル情報の保存における課題

- 各種メディア情報におけるメタデータ、データの国際的標準フォーマットの設定
- オフライン媒体での保存では何年かすると媒体が変質したり、また再生機器がなくなる可能性がある。
- オンライン媒体での保存では、数年ごとに機器の更新とデータの移行を行わねばならず、その経費は高い。
- 何百年も変化しないオフライン記憶媒体の開発が必要である。

50

知識インフラの必要性

- 知識の拡大再生産のためには、知識の創造と集積・流通・活用のサイクルの構築が必要。
- 課題解決型の研究には様々な学問分野がかかわるシステム的アプローチが必要。

51

- 課題を設定するためには、その課題についてこれまでどのような研究がなされて来たか、何が未解決か、イノベーションをおこせる可能性があるか、社会に対するインパクトはどうなりそうか等を調べねばならない。

52

知の共有化

- 多くの分野がかかわるシステム的課題の場合、理工系の研究者だけでなく、政策立案者、人文社会系の研究者や市民もが調査してアセスメントができる環境を作る必要がある。
- あらゆる学問の成果は当然のこと、企業社会、人間社会、自然社会等の知識・情報を収集整理し、自由に利用できるようにしなければならない。

53

知識インフラの構造

- 研究情報基盤の整備が謳われてきたが、通信ネットワークが中心であった。
- 必要なものは学術情報コンテンツ、知識コンテンツの組織的な整備である。
- 分野を超えた知識の関連付けが必要である。
- 日本中に散在するコンテンツをクラウドに移し、そこに検索をかければ関連する全ての必要なコンテンツが得られるようにする。

54

- 知識は関連するものが有機的に結合され、ネットワーク的に統合化されたもの（単に情報を集めたものではない）である。
- 日本中にある人文社会科学を含んだあらゆる学問・研究のコンテンツ、数値データ、研究データ、研究ツール、社会状況データ等が知識の形に組織化される必要がある。
- 諸外国の同様なシステムとリンクがとれる必要がある。

55

- 全ての書籍を部品に解体し、種々の因果関係によって部品同士をリンクすることができる。
- 因果関係としては同義/類似関係、反義関係、上位下位関係、原因結果関係、全体・部分関係などいろいろのものが考えられる。
- 世界の電子図書館の有機的統合による言語の壁を越えた利用を推進すべきであろう。

58

Ⅶ　理想の電子図書館（要約）

- こうして電子図書館を人間の頭脳内の記憶のように知識のシステムの形に構成することができるだろう。
- 連想機能をもつ種々の検索システムによって必要とする情報・知識を取り出せるようになるだろう。
- この知識システムは一種の百科辞典とみることもでき、これに検索質問を出すことによって書誌情報でなく質問に対する答が取り出せることになるだろう。これは事実検索である。

59

知識ネットワークの構築

- 書物が解体され、必要な部分だけが取り出されて使われるような検索方式。
- 関連する知識情報がリンクされて取り出せる知識構造。
- 知識インフラの各拠点がこのような知識構造になっていて、横断的に取り出せること。

Ⅷ　理想の電子図書館を作るための技術要素

Through knowledge we prosper

60

テキストコーパスの構築

- 著作権法の改正によって言語の性質を調べたりする研究目的には許諾なく他人の電子テキストを利用できる。
- Internet 上のテキストを何億文と集めて解析し言語の性質を調べたり、辞書を作ったりできる。

61

ソフトウェア技術

- データベース
- プログラミング
- アルゴリズム
- ネットワーク

62

情報処理技術

- パターンマッチング、パターン認識
- マルチメディア
- 画像、映像処理技術
- クラスタリング
- キーワード抽出、シソーラス
- テキストマイニング
- 関連性検出

63

- 文解析、文章解析
- 対話システム
- 情報圧縮、要約
- 情報分析
- 機械翻訳
- ・・・・・・

64

文化学的知識

- 概念とは、知識とは
- 記号論理学
- 知識の組織化、辞書学
- 古文書学、書誌学
- 図書館と文書館
- 著作権法、プライバシー保護法
- 経営管理、公共サービス論

65

知識は我らを豊かにする

Through knowledge we prosper