言語処理技術が拓く学術情報流通基盤

Human Language Technologies and Digital Library

名古屋大学情報連携基盤センター
Nagoya University Information Technology Center

松 原　茂 樹
MATSUBARA, Shigeki

Abstract

Development of an advanced digital library by utilizing the information technology is expected strongly. This paper describes the availability of the human language technologies in a scholarly information exchange infrastructure. It focuses on the information exchange model consisting of three processes, collection, conversion and publication, and shows that language processing technology can be used for the classification, translation, summarization, mining, retrieving of scholarly information documents.

１．はじめに

World Wide Web の開発が始まってから 15 年余り経過した。その間、ネットワークの高速化が進んだこともあり、Web 技術は急速に発展し、現在では情報流通を支える基盤として揺るぎない地位を築いている。Web は、特に、デジタル文書の流通において、その迅速性、広域性、利便性を武器に大いに威力を発揮する。その意味で、Web をベースとした学術情報の流通技術には大きな可能性があり、すでにそのいくつかは実現され、利用されるに至っている。今後、情報技術の一層の進展により、Web 上のデジタル文書群によって形成する新たなデジタルライブラリが実現されることも期待できる。

情報流通とは、情報が滞らず流れ通ることをいい、そこには、情報を作り出す生産者と情報を利用する消費者が存在する。学術情報流通の場合、学術上の論文やデータなどの生産から消費への流れを意味する。優れた学術情報流通環境を構築するためには、

- 価値ある学術情報を網羅的に収集すること
- 学術情報を機能的に管理できるように適切に整備すること
- 消費者が利用しやすい形式で学術情報を提供すること

が重要となる。流通対象となる学術情報が大規模になれば、これらを手作業で遂行することは困難であり、機械的に処理できることが望まれる。学術情報の多くを占めるのがテキストであり、テキストに記されている言語情報を計算機で扱うことができれば、上述の作業のいくつかを計算機が代行することも不可能ではない。

計算機が人間の言葉を理解したり生成したりする技術は、言語処理技術とよばれる。これまで翻訳、対話、検索等の知的処理の実現を目標に研究が進められてきたが、デジタルライブラリの開発は、今後、言語処理研究の重要なアプリケーションになると予想される。というのも、翻訳、対話、検索をはじめとする言語処理技術のほとんどがデジタルライブラリの高度化に利用できるためである。

本稿では、言語処理技術を利用した学術情報の流

通基盤について述べる。本研究では、学術情報流通を収集、加工、発行の３つのプロセスから構成されるとして議論を展開する。これらの３つのプロセスのいずれにおいても言語処理技術が効果的に機能する可能性がある。以下、第 2 章では、学術情報流通モデルについて概説し、本稿における議論の前提を示す。第 3 章では、言語処理技術について概観し、第 4 章では、言語処理に基づく情報流通基盤技術を、流通プロセスごとに提示し、デジタルライブラリの新たな方向性を示す。

２．学術情報流通

本稿の議論で前提となる学術情報流通モデルを図１に示す。本モデルでは、学術情報の生産、流通、消費の３つのプロセスから構成される。各プロセスは以下の機能を備えている。

- **生産**　学術的な調査、分析、実験、あるいは、論文執筆などにより、学術情報を作成すること。その主体の多くは研究者である。
- **流通**　学術情報が生産されてから消費されるまでの流れのこと。学会や学術機関、出版社などがその中心である。
- **消費**　作成された学術情報を利用すること。新たな学術情報の生産に使用されることもある。

このうち「流通」プロセスは、その機能を整理することにより、さらに収集、加工、発送の３つのプロセスに区分することができる。

- **収集**　学術情報を特定の場所に集めること。
- **加工**　収集した学術情報を整理・分類し、適切な形式に変換すること。
- **発行**：利用者の要求に応じて学術情報を提供すること。

いうまでもなく、学術情報の流通は、Web 技術が生まれる前から実現されていた。例えば、学術論文が生産されてから消費されるまでには、研究者が論文を作成し（生産）、論文の投稿を学会が受理し（収集）、査読や編集により論文に手を入れ（加工）、論文誌として出版し会員に発送し（発行）、会員が論文を閲覧する（消費）といったプロセスを経ることになる。これは、論文の生産から消費までの一般的な流れである。

生産者と消費者に望ましい流通を実現するために、流通の「質」と「量」を高めることが重要である。学術情報流通における「質」とは、真に必要とする学術情報のみを消費できることであり、「量」とは、大量の学術情報を消費できることである。上述のモデルにおける「流通」プロセスを高度化することにより、そのような流通環境を実現することが可能となる。

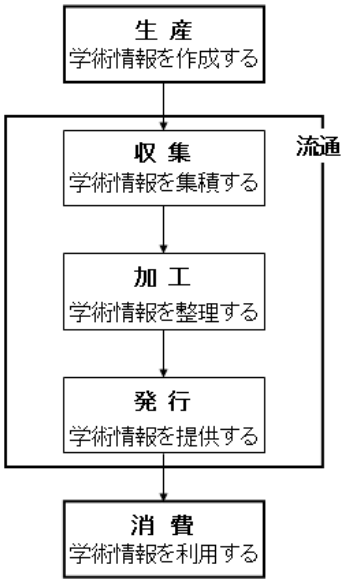

図１．学術情報流通モデル

一方、Web など、近年の情報技術の発展にともない、新しい学術情報流通機構の構築が進められつつある。情報技術を活用した学術情報の流通機構は、一般的な「モノ」の流通と同様、研究者が生産した学術情報を利用者が消費するための仕組みを提供する。この場合、生産とは、学術情報を作り出しそれをオンライン化すること、また、消費とは、オンライン上の学術情報にアクセスし活用することを意味し、流通のインフラとしてインターネットを使用することを前提とする。情報技術を用いた学術情報流通を図２に示す。この場合、流通を構成する各プロセスの機能は以下の通りである。

- **学術情報の収集**　インターネット上に分散する学術情報を集積する。収集ロボットによる自動的な収集と人間が一部介在する半自動的な収集がある 5) 6)。
- **学術情報の加工**　利用者ニーズに合致するような形態に学術情報を加工する。情報群の組織化や情報の変換などがある 7) 8)。
- **学術情報の発行**　利用者が学術情報に容易にアクセスできるような環境を提供する。情報検索インタフェースがその典型例である。

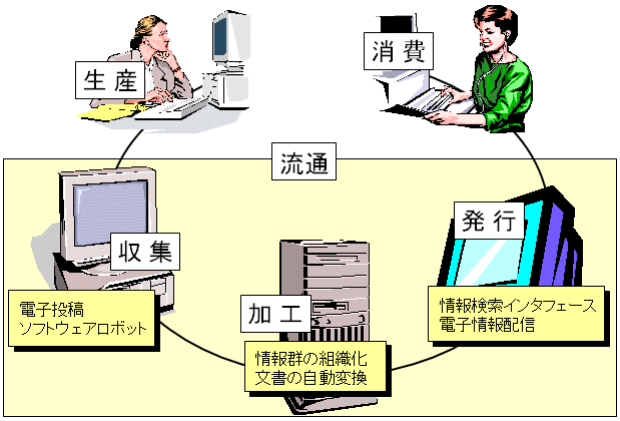

図２ 情報技術を利用した学術情報流通

３．言語処理技術

自然言語とは、日本語や英語、中国語など、人間が日常的に使用する言語であり、計算機によって自然言語を扱うことを自然言語処理という。自然言語処理に関する研究が開始されてからすでに半世紀余り経過しており、これまでに、かな-漢字変換、自動翻訳等、いくつかの技術が実用化されるに至っている。現在でも、全文検索や翻訳電話、会話ロボット等の応用システムの実現を目指して研究が進められており，今後の進展が期待されている．

言語は表層的にみれば、書き言葉の場合は単なる文字列、話し言葉の場合は単なる音素列に過ぎない．しかし、そこにはいくつかの構成素とその間の関係が潜んでおり、言語に潜むそのような性質を正しく分析することなしに言語を正確に理解することはできない。例えば、日本語文「ブラックホールが成長する仕組みが実証された」という文は、単なる２１文字からなる文字列であるが、その中には「ブラックホール」や「仕組み」などといった文字列を、文の中の構成素として認められるし、「ブラックホールが」と「成長する」の間には、修飾・非修飾という構成素間の関係を認めることができる。実際、このような分析がこの文の理解に欠かせない。言語を計算機によって分析することを言語解析という。言語解析は、自然言語処理において基盤となる要素技術であり，分析のレベルによって、形態素解析、構文解析等が存在する。

計算機で自然言語を扱うには、これらの解析技術の利用が不可欠であるが、これまでの研究成果により、かなりの部分が自動化可能な水準まで達しており，学術情報処理等においてもこれらの技術を比較的容易に活用できるような環境が整いつつある。例えば，形態素解析とは、意味をもつ最小の言語単位に分割し、それらの言語的役割を明らかにすることであり、あらゆる言語処理の基礎となる技術である．計算機による形態素解析の自動化のために、さまざまなツールが提供されており、JUMAN 3)や ChaSen 1)など、多くのツールがフリーウェアとして利用できる。例として，先の例文を JUMAN で形態素解析した結果を図 3 に示す. 各行が一つの形態素を示しており、品詞等が記されている．形態素解析は、自然言語処理において最も成熟した技術の一つであり、新聞等の書き言葉に対して 99%を超える精度に達している．構文解析とは、文を構成する要素間の構造的関係を明らかにすることである。計算機による構文解析についても、形態素解析ほどではないものの、いくつかのフリーウェアが整備されつつある 2) 4).

Result of JUMAN

Input: ブラックホールが成長する仕組みが実証された

ブラックホー(ぶらっくほー)ブラックホ	普通名詞	-	-	-	-
が	(が)	が	格助詞	-	-
成長	(せいちょう)	成長	サ変名詞	-	-
する	(する)	する	動詞	サ変動詞	基本形
仕組み	(しくみ)	仕組み	普通名詞	-	-
が	(が)	が	格助詞	-	-
実証	(じっしょう)	実証	サ変名詞	-	-
さ	(さ)	する	動詞	サ変動詞	未然形
れた	(れた)	れる	動詞性接尾辞	母音動詞	タ形
EOS	-	-	-	-	-

juman@kc.t.u-tokyo.ac.jp

図３ 日本語文の形態素解析結果

Result of KNP

- Input: ブラックホールが成長する仕組みが実証された
- Result:

```
# S-ID:1 KNP:2004/07/26
ブラックホールが―┐
          成長する―┐
                仕組みが―┐
                      実証された
EOS
```

knp@kc.t.u-tokyo.ac.jp

図４ 日本語文の構文解析結果

４．学術情報流通における言語処理技術

第 2 章で示した学術情報流通モデルに基づいて、収集プロセス、加工プロセス、発行プロセスにおける言語処理技術の利用について論じる。

４．１ 学術情報の収集

学術情報を自動的に収集するときには、大量に生

産される情報の中から学術情報の範疇に入るもののみを取り出す必要がある。このために、学術情報を定めるプロファイルを作成し、それを用いて情報フィルタリングを実行することは一つの方法である。高い精度を備えたフィルタリングには、優れたプロファイルの存在が不可欠であり、取り出したい学術情報をあらかじめ詳細に観察し、精緻な特徴づけを与えることが重要となる。

４．２ 学術情報の加工

学術情報の加工には２つのタイプがある。１つは、学術情報群の組織化であり、もう１つは、学術情報の編集である。

学術情報群の組織化のための主要な技術は、文書分類である。文書分類の概要を図５に示す。学術情報が適切に分類されていれば情報アクセスが容易となる。文書分類では、文書を単語の集まりとして扱うことが一般的であったが、文書を構文解析し、それに基づいて文書を特徴付ける手法が提案され、その有効性が示されている 9)。

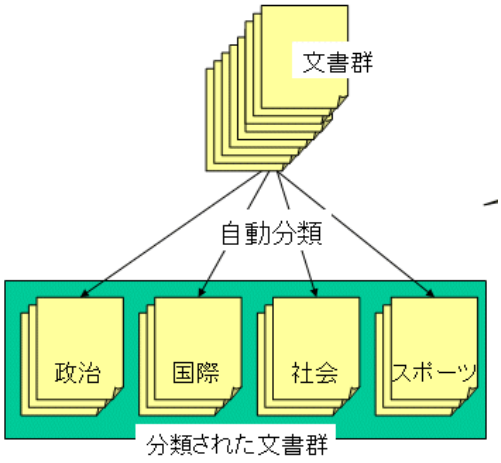

図５．文書の自動分類

文書分類をさらに進めると学術情報群の体系化が可能となる。これは分類以上に高度な組織化であり、類似や対立、評価など、学術情報間の意味的関連付けを与えることを目的とする。そのような試みの一つとして論文の参照情報に基づく関連付け方法が提案されている。その一部は、引用文献情報サイトとして Web 上でのサービスが展開されている 10)。

学術情報の編集としては、文書の自動要約、自動翻訳、自動言い換えなどがある。このうち自動要約は、一般には、重要文抽出によって実現することができ、その機能はいくつかの文書作成ソフトウェアにも搭載されている。また、自動翻訳については実に多くの製品が市販されるに至っており、実用化した技術とみなせる。しかしながら、学術情報を編集するためのツールとしては、性能はまだまだ不十分であり、今後の進展が大いに望まれる。

４．３ 学術情報の発行

大量の学術情報の発行においては、利用者が探している学術情報を容易に見つけることが可能な環境が重要であり、その代表的な技術が情報検索である。情報検索とは、データ群の中から情報を取り出すことをいい、そのようなソフトウェアを一般に検索システムという。検索システムといえば、最近では、全文検索機能を備えることが常識的になっており、事実、文書処理ソフトウェアのテキスト検索機能やインターネット上の検索エンジンが隆盛を極めている。

そのような現存する検索システムのほとんどは、検索質問として単語もしくは単語列がキーボード入力によって与えられ、それを含むデータや文書を返すという形式に従っている。全文検索を初めとする情報検索技術は、ここ数年の間に飛躍的に向上したものの、単語による検索という検索スタイルの制限に不満を抱いている利用者は少なくない。このような現状に対して、情報検索に高度な言語処理技術を導入することにより、新たな検索環境を提供できる可能性がある。

現在、様々な検索システムが使用されているが、その検索対象はいずれもテキストである(画像や音声などのマルチメディア検索システムも存在するものの、そのシステム構成はテキスト検索がベースになっている)。いうまでもなく、テキストは自然言語で記述されている。これは、テキストが単なる単語の集まりではなく、単語の並びによって意味を形成していることを示している。

それにも関わらず、情報検索のための検索キーを単語入力に限定することは不自然である。自然言語による情報検索とは、自然言語文を検索キーとして扱う情報検索を意味する。理想的には、検索キーとして入力された自然言語文の意味を理解し、それと同等の意味を有するテキストを検索結果として返す必要がある。ただし、計算機が自然言語文の意味を正確に理解し、意味が同等であるか否かを正しく判定することは現状では必ずしも容易ではないため、それに近い機能をどのように実現するかが問題とな

る。

全文検索の他にも、言語横断検索、対話的検索、質問応答などがあり、いずれも知的情報アクセス環境の構築に不可欠な技術である。

５．まとめ

本稿では、言語処理技術を利用した学術情報の流通基盤について述べた。まず、学術情報流通のモデルを提示し、モデルを構成するプロセスごとに、言語処理技術の利用可能性を論じた。

デジタルライブラリは、言語処理研究によって極めて重要な目標を提供する。しかしながら、デジタルライブラリの開発における言語処理技術の貢献は必ずしも大きくなかった。今後は、これまでに築かれた言語処理技術を結集することにより、真に使いやすい学術情報流通システムを実現することが課題である。

参考文献

1) Chasen.
http://chasen.naist.jp/hiki/ChaSen/
2) Cabocha.
http://chasen.org/~taku/software/cabocha/
3) JUMAN
. http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html
4) KNP
. http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html
5) Crow R. “Case for Institutional Repositories: A SPARC Position Paper.” ARL Bimonthly Report. 2002.
6) Johnson, R. K. “Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication.” D-Lib magazine 8(11), 2002.
7) Cliff, P. “Building ResourceFinder.” Ariadne 30. 2001.
8) Heery, R. et al. “Renardus Project Developments and the Wider Digital Library Context.” D-Lib Magazing. 7(4). 2001.
9) 鈴木ほか. 構文情報を用いたテキストの自動分類. 東海支部連合大会. 2004.
10) NII 引用文献情報ナビゲータ. http://ci.nii.ac.jp/