電子文書による情報アクセスと情報の共有

東京大学大学院情報学環教授／英国マンチェスター大学教授　辻井　潤一

PROFILE
国際機械翻訳協会（IAMT）およびアジア太平洋機械翻訳協会（AAMT）前会長、
AAMT／Japio 特許翻訳研究会委員長、国際計算言語学会（ACL）元会長、国際計算言語学委員会 (ICCL) 永久メンバー

tsujii@is.s.u-tokyo.ac.jp　TEL 03-5841-4120

1 はじめに

日本語ワープロの出現は 70 年代の終わりで、企業などで実際に使われだしたのは 80 年代の半ばからなので、25 年の年月が経過した。

この 25 年間で手書きの文書作成から、電子的な文書作成の時代への移行が完了した。現在では、手書きは苦手でまともな文書は電子的にしか作れないという人が過半数を超えるであろう。当初は手書きでないと思考がまとまらず、手書き原稿を作ってから電子的なものを作る人も多かった。現在では、筆者を含めて、私的な手紙などでも電子的に原稿を作ってから手書きする人、最初からは手書きできないという人が、むしろ多くなっている。

文書を作るだけではない。小説、新聞、図や表を含んだ論文なども紙媒体ではなく、電子的なものをスクリーンで読むという人が増えつつある。

紙媒体でないと読めない、あるいは、読んだ気がしないという人が、筆者の世代にはまだ多いが、電子媒体でないと読めない人が過半数を超える時代がすぐに到来する。

読みながら、メモの書き込みやハイライトができる、特定の単語やフレーズがでているページに即座に移動できるなど、紙媒体では不可能な機能が実現されている。学生の論文に赤鉛筆で修正をいれる時代は終わった。

国際会議の大部で持ち運びに苦労する論文集も、もう昔ばなしになっている。電子媒体の書籍、書類、論文が一箇所に集積され、必要な部分だけをダウンロードするクラウド化が急速に進んでいる。

書籍や書類の電子化により、情報へのアクセスや情報共有の方法が大きく変化している。この変化は、言語処理やセマンティックウェッブの技術が付け加わることで、さらに加速されていくであろう。

特許情報の管理やアクセスも、それにともなって大きく変化しよう。本稿では、近未来の情報アクセス、情報の構造化と共有がどのようなものになっていくかを考える。

2 情報共有とセマンティックウェッブ

ネットワークに蓄積された個々の情報単位に識別可能なアドレス (URL) を与えて、情報のアクセスと共有化を行うこと、このウェッブ技術の基礎を提唱したのが、ティム・バーナード・リーである。その彼が、2001 年から、現在のウェッブをさらに一歩進化させたものとしてセマンティックウェッブを提唱している [(1)]。

ウェッブには、画像や音声、音楽など言語以外の情報資源も多い。ただ、ウェッブ検索では、文書、すなわち、言語による情報を対象とするものが圧倒的に多い。画像検索も、画像があらわれる周辺の文書中の単語を使ったり、ユーザが付けるタグ（単語や句）を手がかりにしたりすることが多い。

しかしながら、日本語や英語などの自然言語には、同じ単語や句が違ったもの表現する曖昧性、あるいは、違った表現（単語や句）で同じ対象を記述する多様性といった、厄介な問題がある。自然言語で書かれた情報にアクセスしたり、それを共有したりするのは、存外難しい。

セマンティックウェッブの構想は、自然言語の代わりに、曖昧性や多様性のない人工的な言語で情報内容を書き、この記述を手がかりに情報アクセスを行なおうとい

うものである。

URL という、情報内容とは無関係な物理的なアドレスを頼りにした情報アクセスと共有ではなく、セマンティックウェッブでは、情報の内容によるアクセスと共有を目指す。しかし、情報内容が自然言語で書かれている限り、自然言語の曖昧さと多様性のために、現在の情報検索と同じような問題を抱えることになる。

情報アクセスを情報のセマンティックス（意味）に従って行うためには、ウェッブサイト（文書）の情報内容がある種の人工言語で表現されていると都合がよい。自然言語で書かれた情報をオブジェクト・レベルでの情報表現と呼び、この人工言語による意味内容の記述を、メタデータと呼ぶ。あらゆるサイト（文書）にこのメタデータが付与されていれば、アドレス（URL）によるアクセスではなく、このメタデータを使った意味内容によるアクセスが可能となろう。

3 メタデータと統制キーワード

情報の内容を記述したメタデータを情報単位にあらかじめ付与しておいて情報検索を行うことは、別段新しいアイデアではない。医学や生命科学分野では、論文内容を表すキーワードを MeSH(Medical Subject Headings) という統制キーワードの集合から選択し付与することで、これを使った論文検索サービスが、米国医学図書館 (NLM: National Library of Medicine) から提供されている。特許文書の検索でも、IPC コードの付与とそれによる検索は、すでに一般的に行われている。

セマンティックウェッブの構想は、これを個別の検索システムの内だけでなくて、ウェッブというグローバルな情報空間で行うことで、多数の独立したシステム間でも情報共有しようとするところに新しさがある。また、これまでの統制キーワードがそれ自身の内部には構造を持たない、単純なワード（語）であったの対して、文やテキストの意味内容など、より構造的なものまでも構造化して表現し、それをウェッブ空間全体で共有しようとすることにある。

文書やウェッブサイトの意味内容を標準化されたメタデータで表現すること、また、この標準化を単純な統制キーワードではなく、ある種の構造をもった人工言語で書いておくことによって、情報のアクセスと共有を可能にしようとするセマンティックウェッブの考え方は、今後の情報システムに大きな影響を持つと考えられる。

4 情報共有とURI

検索システムで使われてきた統制キーワードとセマンティックウェッブ構想の差の一つは、後者がメタデータの解釈をグローバルに分散した不特定多数の情報システム間で標準化することで、検索自体よりも情報共有に重点を置いていることである。

ティム・バーナード・リーは、現在のウェッブでの情報アクセスを支える URL に変わるものとして、意味的なメタデータを捉えている。このために、ウェッブ中の場所に与えられる URL(Universal Resource Locator, Uniform Resource Locator) に代わり、個々の意味的概念に与えられる URI(Universal Resource Identifier) を提唱する。URI は、現在のような URL ウエッブ中の物理的なアドレスではなく、特定の個人や会社、製品などのように個別の個体や概念などのように、それに対して様々な情報（たとえば、個人の場合には年齢、性別、職業など）が付与される基本的な単位に対する識別子である。

言語処理分野では、人名・地名・組織名・製品名など、情報が集積する単位を固有名 (Named Entity) と称して、この生起をテキスト中で認識する研究が行われている (NER: Named Entity Recognition)。また、固有名であることの認識だけでなく、同姓同名の人を区別したり、逆に、違った表現であっても同一要素を指す場合には、それらを認識して同一の識別子を与える標準化処理 (Normalization) も研究されている。

NER での識別子は一種の URI であるが、特定のシステム中での識別子であるだけでなく、ウェッブ中に分散して存在する文書 (ウエッブサイト) すべてに共通して使われる識別子であることが重要である。たとえば、JAPIO, Japio, 日本特許情報機構、特許情報機構など、現実の世界では同一の組織であるものが、複数の文書や

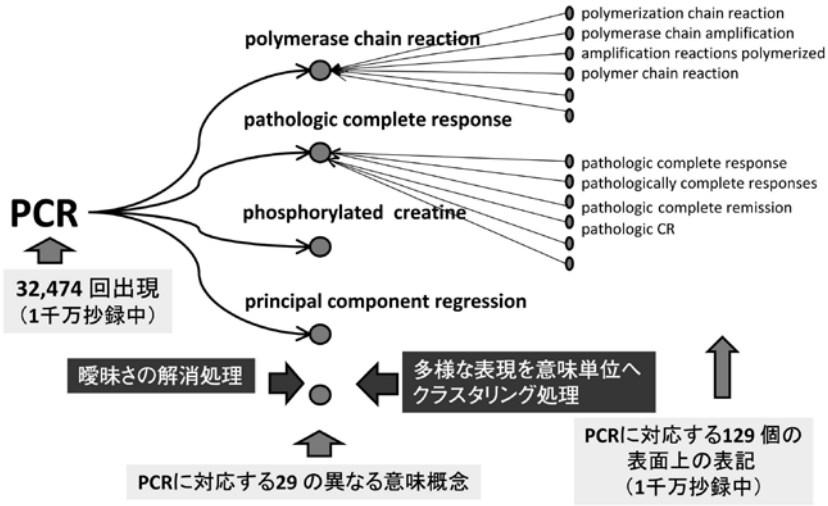

図１　多様性と曖昧性

ウェッブサイトでは、表面上違った単語や表現で指し示されることがよくある。これらすべての一見違った表現が同一の組織を指し示すことが認識され、その結果、同じ識別子で索引つけされることで、「Japio」に関する情報が、複数の文書やウェッブサイトに分散して存在していても、また、たとえそれらの文書やサイトで違った呼称が使われていたとしても、すべて一括して検索することができる。

NER の対象となる固有名は、人名・地名といったものに限らない。特許や科学技術文書では、同様な役割を専門用語が果たしている。専門用語は特定分野での重要な概念を表現し、その概念に付随した、様々な種類の情報が集積される結節点としての役割を持っている。また、同じ専門的概念が違った表現（そのままの英語表現やカタカナ語、略語、複数の日本語訳など）で指示されたり（多様性）、表面上同じ表現が文脈によっては違った専門概念を指したり（曖昧性）する[(2)]。

これらの自然言語の問題に解消して標準化し、表面的な用語ではなく、概念として一つのものとみなされるものにすべて同じ一意の識別子を与えることにより、特許や科学技術文書への意味アクセスが可能となる。筆者の研究グループでは、生命・医療分野での意味にもとづく文献アクセスシステムを構築しているが、この分野では、文献中に略語が頻繁に出現する。略語には、自然言語の欠陥である多様性･曖昧性（図１）が顕著に現われ、表面上おなじ略語でも文脈によって異なった専門概念を

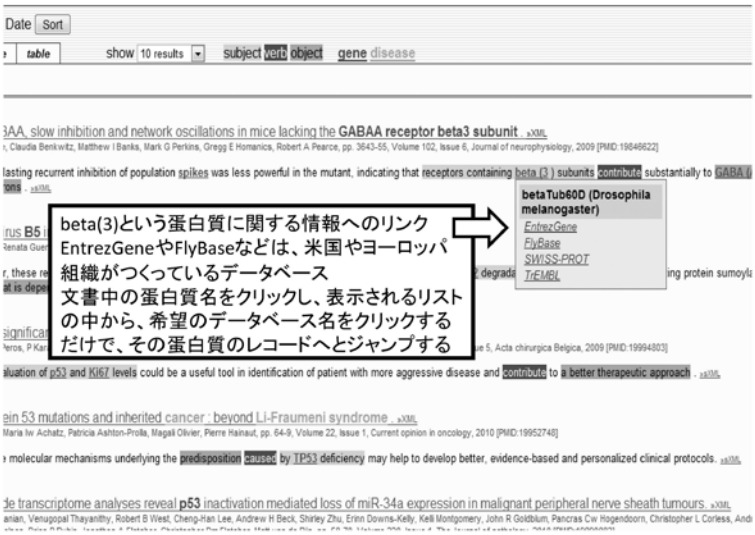

図２　文献検索システム　MEDIE の画面

表現していること（曖昧性の解消）、また、一見違う表現が実は同じ専門概念を表現していること（多様性の解消）を認識することで、各段に情報検索の精度を向上させることができる[(3)]。

5 文書のハイパーテキスト化

文書の自動的なハイパーテキスト化は、固有名へのハイパーリンクづけなど、一部のサイトではすでに行われている。個別のシステム内部の識別子ではなく、ウェッブ中の多くのサービスシステムが、識別子 (URI) の体系を共有することで、システムの境界を超えたハイパーリンクの結びつきが可能となる。図2は、我々の研究グループがサービスしている文献サービスシステム (MEDIE) である[(4)]。このシステムでは、蛋白質·遺伝子、病名といった固有名認識を行い、個々の固有名にハイパーリンクを付与している。ユーザは、固有名のクリックで、固有名が蛋白質や遺伝子の場合には、UniProt、EntrezGene、GenAtlas、HUGO、TrEMBL という米国・ヨーロッパの組織が構築しているデータベースのレコードへと直接ジャンプできる。さらにはウェッブ中に散在する様々な App を起動し、たとえば、対応する蛋白質の3次元構造を表示することも可能となろう。

このようなさまざまな情報源に散在する情報の即時的な相互リンクは、紙媒体では不可能であり、電子媒体のみがもつ利便性であろう。

6 細かい粒度での情報アクセスと相互リンク

論文などの文書は冒頭から末尾まで通して読まれることを前提にするが、熟達した研究者や技術者は、必要な箇所を素早く見つけ、そこだけを詳しく読むなど、メリハリの利いた読み方をしている。

このようなメリハリの利いた論文や技術文書の読み方を積極的にサポートすることが、今後の技術開発のひとつの焦点になろう。伝統的な文献検索が、文献を単位として検索していたのに対して、現在研究されているパッ

図３　細かい粒度での情報アクセス

セージ検索では、個々の文書の内部にまで立ち入って、検索要求への答えを含んでいそうなパラグラフ、質問応答システムの研究では、質問への直接の答えを返すこと、を目標としている。

このような細かな粒度での情報の検索や関係付けも、すでに一部のサービスでは開始されている。例えば、図3は、米国医学図書館（NLM）のサービス（PubMed）の画面を示したものである。このサービスでは、論文本体の横にその箇所に関係する他の論文が表示され、これをクリックすることで、別論文へと直接にジャンプできる。現時点では、本文中で引用されている文献の中で重要なものと判断される論文の表示や論文の種類（Review や Research paper）の区別、その概要の

図４（b）　Path Text

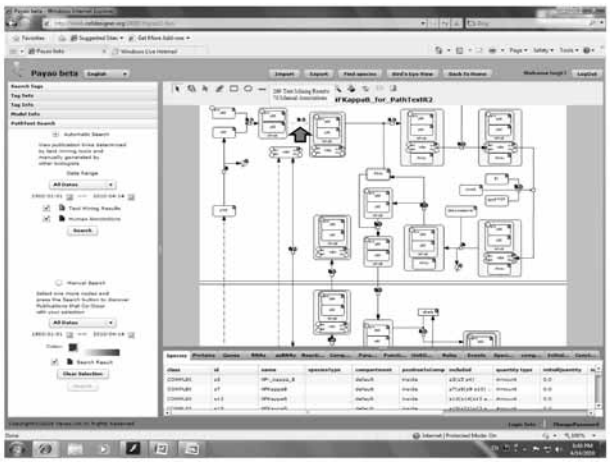

図４（a）　Path Text
青い矢印部分をクリックすると、このモデル中のこの箇所に特に関係する論文の部分が表示される（図４（b）参照）

一部を追加することで、読者がその論文へとジャンプするかどうかの判断をサポートする程度であるが、論文の内部に立ちって、その部分に対して関係する他の論文へのハイパーリンクを貼っていく試みとして、興味あるものとなっている。ここで、重要なことは、個々の論文にも概念などの固有名と同じように一意に識別可能な識別子が割り振られていることである。この識別子を用いることで、ある論文から別の論文へのジャンプが可能になっている。

現在のところ、NLM の試みは、本体論文と引用論文の意味内容に関する処理はなんら行っていないために、ジャンプ先は、引用論文の冒頭になっている。しかし、次節でのべるような論文の意味内容が処理できると、ジャンプ元の論文の特定箇所と特に関係が深い引用論文の特定の章やパラグラフにジャンプするなど、粒度のより細かな情報リンクが可能となろう。

図４(b) は、このような細粒度の情報のリンク付けを行うシステムとして、我々が開発しているシステム (PathText) の画面を示したものである [(5)]。

PathText では、ジャンプ元は論文ではなくて研究者が興味を持っている生体系のネットワークモデルであり、ネットワークの一部をクリックすること (図４(a)) で、ネットワーク中の特定のリンクが表現している特定の生命現象が記述されている論文の部分へとリンクが貼られている。

このように、電子媒体の文書は、外部から文書中の内部の任意の箇所へと直接リンクすることを許すことにより、冒頭から末尾へと読み進むという従来の文書とは異なる読み方、すなわち、自分の興味に関係する箇所へと外部から直接ジャンプして、そこから文書を読むという読み方をも可能にすることになろう。科学技術論文や特許文書など、大量の文書情報を読みこなすためには必須の道具となろう。

7 複合構造を持った意味索引

固有名からの情報アクセスは、固有名認識 (Named Entity Recognition) 技術の発展により、かなり実現されてきている。これに対して、セマンティックウェッブ構想に見られる構造的なメタデータは、どのような情報アクセスを許すことになるのだろうか？

生命医学系の統制キーワード MeSH や、特許の IPC にあらわれる専門概念には、専門用語として一つの単語のようにみえるが、実際には、句や文で表現されるような複合的な内容、内部構造を持つものも多い。たとえば、図５は、生命医学系で使われる統制キーワード（オントロジーと言われる）である GO(Gene Ontology) の統制キーワードの一つが、文書中でどのように現れるかを示したものである。

この統制キーワード (STAT protein nuclear translocation) は、「STAT ファミリーに属する蛋白質が核外から細胞核へと移動する現象」を指すもので、

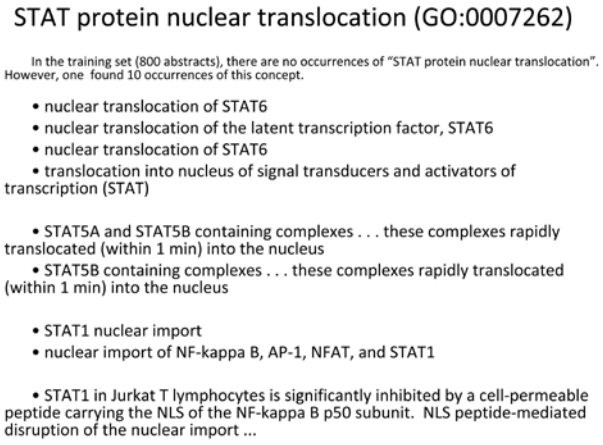

図５　統制キーワードと言語表現

実際の文書中では句や節、場合によっては、パラグラフにより記述されている（図５参照）。

また、この統制キーワードは、「蛋白質の細胞内での移動」、「蛋白質の核外から細胞核への移動」、「STAT ファミリーの細胞内での移動」、「STAT ファイミリーの核外から細胞核への移動」…. といった他の統制キーワード、同じように複合的で内部構造をもつ統制キーワードと意味的な関係を持つ。これら統制キーワードの関係を系統的に捉え、かつ、文書中での表現と対応付けるためには、このようなキーワードをこれ以上分析できない単語と見るよりも、たとえば、

[Event: Translocation
Theme: STAT family
To: Nuclear
From：Cytoplasm]

という内部構造を持った意味単位と考えるのが自然であろう。これらの意味構造をメタデータ記述として使い、図５の多様な表現をこの意味構造へと標準化することで、固有名と同様なハイパーリンクづけが句や節、パラグラフを単位として可能となる。

言語処理技術の研究では、多様な文表現によって記述されている事象 (Event) を実は同一の出来事を表現しているものとして認識し、標準化する研究が、活発化している。これらの研究は、前述の固有名認識 (Named Entity Recognition: NER) の研究に対して、事象認識 (Event Recognition: ER) の研究と呼ばれている。６節の PathText は、この ER の成果を細粒度の情報アクセスへと適用したものである [(6)]。

8 おわりに

本稿では、文書が紙媒体から電子媒体へと変わり言語処理技術の進展と結びつくことで、近未来にどのような情報アクセスが可能になるかを議論した。

文書作成の過程は、紙媒体から電子媒体への移行に25 年かかった。現在では、文書作成は電子媒体によるものが主流となっている。

これに対して、文書を読む側は、まだ紙媒体での文書が大きな役割を果たしている。しかし、電子書籍や iPAD の発売など、読む側にも電子媒体の役割が大きくなってきている。特に、研究者や技術者が論文をウェッブ中から検索し、ダウンロードして読む過程が一般化するに伴い、紙媒体に印刷する過程を経ずに読むことが普通になりつつある。さらに、本稿で議論したような意味に基づく細粒度の情報アクセスを可能にするシステムが一般化するに伴い、紙媒体の果たす役割はさらに低下していくものと思われる。

IPC コードをさらに拡張した意味的なメタデータの付与など、このような大きな技術の流れの中で、今後の特許情報システムのあり方を考えてゆく必要があろう。

[参考文献]

(1) Tim Berners-Lee, James Hendler, Ora Lassila: The Semantic Web, Scientific American, May 2001
(2) Ananiadou, Sophia, Douglous Kell and Junichi Tsujii. Text Mining and its potential applications in systems biology. Trends in Biotechmology. 24(12). Elsevier, 2006.
(3) Okazaki, Naoaki, Sophia Ananiadou and Jun'ichi Tsujii. Building a High Quality Sense Inventory for Improved Abbreviation Disambiguation. Bioinformatics. OUP, 2010.
(4) http://www-tsujii.is.s.u-tokyo.ac.jp/medie/index.cgi
(5) B.Kemper,T.Matsuzaki,Y.Matsuoka,Y.Tsuruoka,H.Kitano,S. Ananiadou, J.Tsujii :PathText: a text mining integrator for biological pathway visualizations, Bioinformatics, Vol.26 (12), OUP, 2010
(6) Ananiadou, Sophia, Pyysalo, Sampo, Tsujii, Jun'ichi and Kell, Douglas B.. Event extraction for systems biology by text mining the literature. Trends in Biotechnology. 28(7). pp. 381-390, 2010.