情報検索から情報構造化、情報発見へ向けて

東京大学名誉教授／マイクロソフトリサーチアジア研究所首席研究員 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

1 はじめに

計算機システムの中には、大量のテキストが蓄積されている。ウェッブ中で一般にアクセスできるテキストだけではない。会社や組織の中で流通するテキスト、学会などでアーカイブされているテキスト、特許データベースに蓄積されたテキストなども、急速にその量が拡大している。

こういった共有テキストベースは、普通のウェッブ検索では検索できないが、テキストを共有するコミュニティが存在し、そのコミュニティのメンバーが不断にテキストを生産し、それらを共有している。

テキストは、情報流通の基本的な形態であり、そこには構造化されていない大量の情報が埋め込まれている。テキストは、発信者と受信者の間の情報流通を支えてきたが、それらが計算機システム内に蓄積されることで、流通時には意図していなかった受信者、潜在的に大量の受信者に対する情報発信が可能になった。言い換えると、それまでは第一義的な受信者が情報を受け取るとそれで使命を終えていたテキストが、現在では、蓄積され長期に渡って参照できる。このことで、直接の情報流通に関与していなかった、非常に多数の第３者が時間的、空間的な隔壁を越えて情報を共有できるようになった。

このような性質が、ウィキーリークスなど、情報流通への直接の関与者が持つ意図とは無関係に情報が流通するという、現代的な問題を生むことにもなっている。

ここでは、大量のテキストが長期に渡って蓄積され、それが常に利用可能になった状況でのテキスト処理と知識処理の技術的可能性について議論する。

2 検索を超えて

大量のテキストが常に利用可能になった状況で、最初に発展した技術はウェッブ検索の技術であった。ウェッブ検索によって、テキスト情報は空間という物理的な制限がなくなり、それまでの図書館、新聞、各種の情報サービスがウェッブに置き換わることになった。

取得したい情報をキーワードでウェッブ検索することは、日常化している。ただ、このウェッブ検索も、蓄積される情報の巨大化、時間的経過がもたらす蓄積情報の陳腐化など、徐々にその限界が明らかになりつつある。

必要な情報を取り出すためにキーワードをいくつも入れ、組み合わせを変えて何度もウェッブ検索にかけた、という経験を持つ人は多いであろう。使っているソフトウェアや電子機器が故障して、その対処方法を検索しようとして、長時間をウェッブ検索に費やした、といった経験である。

このような経験がすこしづつ軽減されだしているのは、Yahoo 知恵袋のような Q ／ A サイトの登場であろう。テキストに埋め込まれた情報をうまく検索で取り出すことは、実は、それほど容易ではない。Q ／ A サイトは、ある人が知りたい情報は他の人にも有効であろうという前提で、質問と答えをあらかじめ構造化しておく。同様に、大部な百科事典の現代版として、大勢の人が使っている Wikipedia も、ユーザが興味を持つであろう特定の人物や組織、概念ごとにそれらに関する情報

をあらかじめすべてまとめて構造化しておくことで、情報アクセスを容易にしている。

ただ、Q／A サイト、Wikipedia は、いずれもあらかじめ情報を構造化して組織化した二次的な情報源である。この二次情報の、構造化の作業をコミュニティーとして行う、今はやりの言い方では Crowd Sourcing で大規模に構築することで成功事例である。ただ、このようなあらかじめの構造化は、大勢の人が必要とすることを前提にし、Crowd Sourcing とはいえ、かなりの人手と時間が必要となる。したがって、本来のウェッブ検索が持っていた、新鮮な情報の取得、個人が必要とするニッチな情報の入手という、利点を損なうことにもなっている。特許情報の検索のように、個々の事例ごとの検索が必要な場合には、あらかじめの構造化には限界があろう。

本稿では、Crowd Sourcing による知識の構造化に続く技術の可能性について議論する。

3 構造化データベースとテキスト処理

Wikipedia が有用なのは、ある個別の人物、組織、概念といったもの（一般に Entity、と呼ばれる）に関する情報が一箇所に集められていること、ということである。実際、英語版での Wikipedia（2012 年 9 月 23 日時点）では、Michael Jordan という名前を持つ有名な人物には 10 名の異なった人物があり、そのうち 9 名の人物についての Wikipedia の項目が用意されている。もっともよく知られたバスケットボールの選手以外にも、機械学習の著名な研究者（この人物は、われわれの研究分野の研究者にとっては、バスケットボールの選手以上になじみが深い）らの項目がある（図 1）。

このように言語の待つあいまい性（たとえば、同じ名前で異なった人物が指される）や多様性（同じ人物の名前が違って表記されたり、複数の名前を持つ場合）を言語処理手法を使って解消し、Entity を中心にして情報をあらかじめ構造化しておくシステム（Entity-based System）は、この Wikipedia 以外にも数多く見受けられるようになってきた。

研究者にとって馴染みが深いシステムでは、マイクロソフトのアカデミックサーチ（Microsoft Academic Search: MAS）、国立情報学研究所の CiNii システムがある。MAS での Entity は、個々の研究者だけでなく、個々の論文、論文誌、国際会議なども Entity として取り扱われ、これらの Entity ごとの情報の集積と Entity 間の関係が構造化されている。したがって、各論文が別のどの論文によって参照されているか、複数の研究者がどれだけ共著論文を書いているかを即座にみることができる（図 2）。また、その論文を参照する別論文のテキスト中での箇所も即座に見ることができる（図 3、図 4）。

もう少し科学技術情報によった分野での Entity 中心のシステムに、私が東大に在職中に開発した、生命科学分野での知識管理システム（MEDIE）がある。このシステムでは、論文中に現れた Entity（生命科学の分野では、基本的な Entity にたんぱく質、遺伝子、DNA、化学物質といったものがらう）を認識して、これをユーザがクリックするだけで、その Entity に関する外部のデータベース中での情報が参照できる機能を付けてい


Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

Michael Jordan is an American basketball player.

Michael Jordan may also refer to:

- Michael Jordan (mycologist), English mycologist
- Michael Jordan (footballer) (born 1986), English goalkeeper (Arsenal, Chesterfield, Lewes)
- Michael Jordan (born 1931), English businessman, former senior partner of Cork Gully
- Mike Jordan (born 1958), English racing driver
- Michael Jordan (Irish politician), Irish Farmers' Party TD from Wexford, 1927–1932
- Michael B. Jordan (born 1987), American actor
- Michael I. Jordan (born 1957), American researcher in machine learning and artificial intelligence
- Michael H. Jordan (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- Michael-Hakim Jordan (born 1977), American professional basketball player

See also

- Michael Jordan statue, Chicago statue of the basketball player Michael Jordan
- Jordan (name)

図 1　Wikipedia の項目

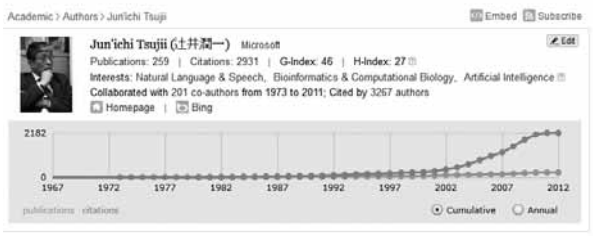

図 2　MAS の著者に関するのページ

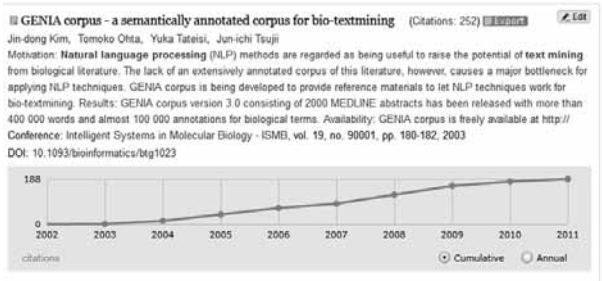

図３　特定の論文を参照する論文数の推移

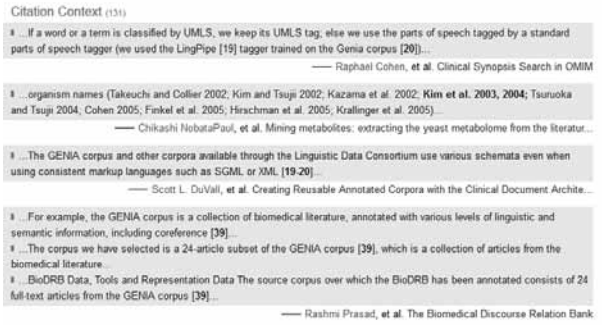

図４　論文の被参照箇所の提示

た。ここでは、同じように見える名前が科学的にみると違った Entity を指しているとか、全然別の名前で同じ Entity を指しているものを認識する必要があった。

重要なことは、テキスト中での Entity を認識する技術は、Named Entity Recognition（NER）技術として整備されて、Wikipedia のような Entity 中心の情報構造化が、あまり人手を掛けずに実現できるようになってきたことである。

キーワード検索で、関係のありそうな Web ページや文献のリストを見せるだけの単なる検索技術から、Web ページや文献中に散在しているある特定の Entity に関する情報を集約して、かき集めて、それらを網羅的に見せる技術の基盤技術が出来つつあることである。

現在の技術では、多くの文献やウェッブページに散在する冗長な情報を除去して、本質的な情報のみに整理することはできない。ただ、多くの研究グループが、ウェッブ中の情報をかき集めることで、Wikipedia の項目に対応するものを自動構築する研究に取りくみ始めている。

4 今後の方向

Wikipedia のような Entity 中心の構造化は、非常に有効であった。ただ、構造化された情報を常に最新のものに保持しておくためには、新たな項目を書き下すのと同じくらいに大変な努力が必要となる。とくに項目が爆発的に増えている現在、それら項目の内容を常に Up-Date なものに保持していくコストは、Crowd Sourcing という、掛け声だけでは解決しない。

実際、Wikipedia が現在のように有名な人間や組織だけを対象としているときには、Crowd Sourcing だけで情報を最新のものに保っておくことは可能であるようにみえる。

しかし、この技術をより大量の、あまり有名でない Entity に適用したり、さらには、Entity 間の関係までも構造化していった場合には、それらの情報を、最新の、かつ、信頼できる状態に保持するには、膨大な人手が必要となる。

同様な問題は、生命科学の知識管理システムにおいても顕著である。この分野では、たんぱく質の機能に関して、日々新たな論文が出版され、これに対応して、２次的に整理された構造化データベースをの内容を更新、維持していく必要がある。このためには、各データベースごとにキュレータと呼ばれる専門家、数十人の専門の研究者が論文をよみ、日常的に２次データベースを更新している。発表される論文の加速度的な増加、構造化しておくべき情報項目や Entity の増大に伴い、人手による情報の管理は、非常に高コストなものとなる。

構造化されていないテキスト情報がもつ即時性と構造化データベースが持つ情報の質、信頼性、集約性とをどのように連続させ、相互の利便性を向上させていくことは、Wikipedia 的な情報の構造化を成功させるキーテクノロジーとなりつつある。

5 検索から発見へ：テキストマイニング

情報検索の基本は、必要な情報がある Web ページ、やテキストを見つけてユーザに提示することであった。巨大なテキスト群に埋もれた情報をユーザに提示するという基本は、前述の Entity 中心のシステムでもどうようである。

これに対して、巨大なテキスト群を分析することで新たな情報を見つけ出すことを目指す技術がテキストマイニングである。特許文書の集合を全体として分析することで、個別会社や会社群の技術開発動向を発見するとか、技術間の相互依存関係を見つけ出すという、個々のテキストには明示的に記述されていない新たな情報を生み出す技術である。

私が、2011 年まで勤務してきた英国国立テキストマイニングセンター（NaCTeM）では、生命科学の分野で薬と症状、病気とたんぱく質の相互関係を大量のテキストから自動抽出し、それを視覚化して提示するシステムを構築した。この Facta+ と呼ばれるシステムは、当時 NaCTeM に在職していた鶴岡慶雅さん（現在、東京大学工学系大学院准教授）が中心となって開発したものであるが、多くの生命科学者の注目を浴びることになった（図５）。実際、ある特定の病疾患を研究している専門家からは、それまで気がついていなかったその疾患とたんぱく質との関係に気がついた、といった感想が聞かれた。

NaCTeM が開発した Facta ＋は、論文中の概念間の統計的な共起関係を分析の基礎としているが、この分析が単語レベルで行われるのではなく、前述の Entity 中心の技術や Entity 間の構造的な関係分析の結果をもとに統計的な共起関係を計算することで、精度を向上させたものであった。

今後は、Entity 中心のような知識の構造化手法と Facta ＋に見られるマイニング手法が統合されていくことで、あらたな情報技術を生み出していくものと思われる。

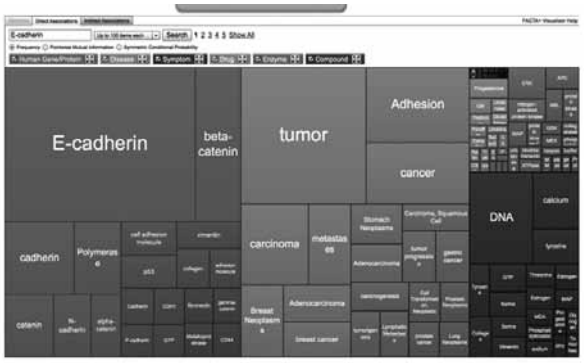

図５　Facta+_ の視覚化インターフェース
（E-Cadherin というたんぱく質に関連した疾患、遺伝子、薬などが関連の強さに応じたタイルで示されている。このインターフェースから、関連する論文に直接移動することができる）

6 終わりに

ウェッブ検索から情報の構造化、新情報の発見への動きは、いまその端に着いたばかりである。今後、これらの方向での技術開発が加速化することが期待される。