Japio

概念辞書によるシステマティックなイノベーション支援に向けて

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター　MASTAR プロジェクト
言語基盤グループ グループリーダー
鳥澤　健太郎

PROFILE
1995 年東京大学助手、北陸先端科学技術大学院大学准教授を経て、2008 年より現職、自然言語処理の研究に従事。

torisawa@nict.go.jp　TEL
http://mastarpj.nict.go.jp/~torisawa/

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター　MASTAR プロジェクト
言語基盤グループ 研究員
風間　淳一

PROFILE
北陸先端科学技術大学院大学情報科学研究科助教を経て 2008 年 8 月より現職、自然言語処理の研究に従事。

kazama@nict.go.jp　TEL
http://mastarpj.nict.go.jp/~kazama/

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター　MASTAR プロジェクト
言語基盤グループ 主任研究員
村田　真樹

PROFILE
1997 年京都大学にて日本学術振興会リサーチ・アソシエイト、1998 年郵政省通信総合研究所入所を経て 2008 年より現職。自然言語処理の研究に従事。

murata@nict.go.jp　TEL
http://mastarpj.nict.go.jp/~murata/

1 はじめに

イノベーションという言葉が流行語となって久しい。ちまたには、イノベーションを加速するためのノウハウ本が流通している。本稿では IT、さらには Web 上などに存在するテキスト情報を用いて、イノベーションを加速する構想について述べてみたい。

2 生成とテスト

コンピュータのアルゴリズムには、「生成とテスト」(generate-and-test) と呼ばれる一連のアルゴリズムがある。これは正しい出力を一意に求めるための方法が明確になっておらず、ただ、正しい出力が満たすべき条件だけが分かっている場合に有効である。「生成とテスト」自体の原理は非常に簡単で、なんらかの、時にはある意味いい加減な方法で出力の候補を一つ求める。ついで、その候補が正しい出力になり得るかどうか、つまり、出力がみたすべき一定の条件をクリアしているかどうかを確認し、もし、条件がクリアされるのであるならば、それを出力とする。こういったことをすべての出力の候補において検討して、出力の集合がすべて出揃うまで繰り返す。つまり、出力の候補を全て見つけ出して、その候補をテストし、最終的な出力を決定するというわけである。特に出力の候補を見つけ出すことを、コンピュータサイエンスでは、出力の候補を「生成する」(generate) という。

具体例として、詰め将棋を考えることができる。詰め将棋の答えは、駒の動きということになるが、その動きの背後にあるはずの指し手の思考は非常に複雑なもの

であり、これをアルゴリズムとして再現するのは容易ではない。こうした状況では、一般に「生成とテスト」が有効である。つまり、すべての可能な駒の動きを仮の指し手として生成して、それが相手の王将の詰めになっているかどうかをテストしていけば、いずれ答えにたどり着く。ここで重要なことは、「生成とテスト」のアルゴリズムには、すべての駒の動きを数え上げることと、具体的な駒の動きが詰めになっているかどうかの判定の二点だけが必要であるということである。詰めになるかどうかのチェックは比較的簡単である。特に、すべての駒の動きを生成するということは、将棋盤上の味方の駒の各々に対して、ルールが許す全ての動きを逐一数え上げるだけで良く、いわゆる定跡であるとか、冴えた直感だとかは一切必要ない。従って、「生成とテスト」による計算は、前述したような指し手の思考、特にいろんな意味で良さげな駒の動きをピックアップするプロセスをアルゴリズムで真っ正直にシミュレートすることに比べれば遥かに単純であり、実際に実現可能である。つまり、非常に複雑なプロセスであっても、「生成とテスト」によって、計算機のプログラムとして実現可能な場合があるのである。

もちろん、これは非常に単純化された議論であって、「生成とテスト」をアルゴリズムとして採用することによるデメリットは多々ある。例えば、生成される候補の数が膨大であれば、計算には膨大な時間がかかる。また、生成される可能な候補が無限にある可能性があれば、計算が終了し、正しい答えが出力される保証もない。しかしながら、例えば、イノベーションのように非常に複雑でその定式化が難しいような問題を計算機で解こうとすれば、何らかの形で「生成とテスト」のアルゴリズムを導入することは、ある意味、コンピュータサイエンスにおける「定跡」なのである。本稿では、こうした発想に基づいて、我々が現在開発している概念辞書という巨大な辞書を使って、イノベーションに「生成とテスト」的な発想を持ち込む構想について述べる。

3 イノベーションと「生成とテスト」

さて、ここで本題のイノベーションに話を移すが、イノベーション自体は非常に複雑なプロセスである。しかしながら、多くの場合、イノベーション自体は、

1. 仮説，アイディア候補の作成。
2. 実験 / 試作による検証。ただし、いわゆる「思考実験」も含む。

の二つのステップに分けることができるであろう。この二つのステップ自体、先の「生成とテスト」における、出力候補の生成と、出力候補のテストに対応づけることが可能であろう。ただし、「生成とテスト」の場合、テストは計算機の上での計算でしかないのに対して、実験、試作ともなれば、普通は計算だけではすまない。一方、仮説、アイディアの生成に関しても、人間によるイノベーションの場合には、直感、イノベーションのキーとなる自然現象等の発見、気づき等、非常に複雑で現状、計算機上で実現できないようなプロセスが含まれているが、仮説、アイディア候補を生成すればよいだけに計算だけで可能な部分というのはテストに比べれば多いであろう。また、詰め将棋の例とのアナロジーで行けば、アイディア候補の生成では、本来イノベーションを行なう人間が行なう非常に複雑な思考のプロセスをな

ぞる必要は本質的には必要でないかもしれない。以下では、こうした仮説の生成の部分を、現在我々が Web から自動構築している概念辞書を用いて自動化する構想について検討したい。

詰め将棋の話に戻ると、人間が詰め将棋をする場合と、「生成とテスト」のアルゴリズムに基づいて詰め将棋の答えを見つける場合の大きな差は、指し手を生成するという操作が、単純にシステマティックに行なわれているかどうかである。「生成とテスト」のアルゴリズムでは見落としがないようにすべての可能な駒の動きをシステマティックに数え上げるのに対して、人間が駒の動きを選ぶ場合にはすべての可能な手を少なくとも意識的に検討しているという保証はない。実際に将棋の下手な人間が詰め将棋をする場合には、見落としが生じて、実際に答えにたどり着くことができないことが良くある。

本稿で述べる構想で仮説の生成を自動化することによって得ようとしているメリットは、まさに見落としがないようにすべての可能性をシステマティックに数え上げる、もしくは可能な限り多くの可能性を数え上げ、人間による検討、思考実験、さらには実際の試作、テストに供することにある。

エジソンが白熱電球を発明するにあたって、数千種類のフィラメントの材料を試し、とりあえず、京都産の竹を材料として見いだしたというのは有名な話である。実際問題としてそうした作業は研究開発においては日々必要になるであろう。また、仮にエジソンがトライした数千種の中に、竹を含めて彼の手近にある有用なものがぬけていたとすれば、それだけ白熱電球が世の中に出るのが遅れた、あるいはそもそも世の中に出なかったということになろう。我々の狙いはまさにそうした事態をさけ、さらにはイノベーションのプロセス自体を加速させるものである。

4 概念辞書と具体例

前置きが長くなったが、ここで、現在我々が開発している概念辞書を用いて、具体例を示したい。この概念辞書は Web にある 1 億ページの情報をもとに自動で構築されているものであり、多様な語の間にある多様な意味的関係を含んだ辞書である。現在約 200 万語の語またはフレーズの間の意味的関係を含んでいる。また、これは検索支援システム「鳥式改」（図 1）で利用されており、意外でありながら、有用な情報を検索することができるようになっている。（参考文献 1、2）この概念辞書の「トラブル」「原因」「対策」といった意味的関係の一部を示したのが、図 2 である。この図では、「洗濯機」に関係するトラブルとして、当然予想されるものとしての「黒カビ」、「カビ」から、多くの人にとって意外であろう「アトピー」、「アレルギー性皮膚炎」、「溺死」と言ったようなものが提示されている。さらに、このトラブルの間の因果関係も概念辞書には含まれており、例えば、「黒カビ」、あるいは「カビ」が「アトピー」の「原因」になることがわかる。また、「カビ」に対する「対策」として、「光触媒」、「銀イオン」、「安息香酸」、「ヒノキチオール」がある。

こうした意味的関係は、例えば、対策の場合「A が B に有効である」というパターンに現れる語 A、B を Web 文書から抽出することによって得られている。ヒノキチオールがカビと「対策」という意味的関係を持つ

図１　検索支援システム「鳥式改」(洗濯機に関するトラブルを表示したところ)

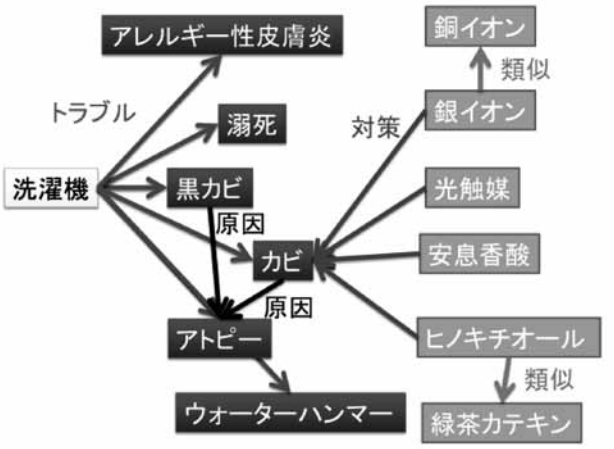

図２　概念辞書の一部抜粋

ことは、そうしたパターンを「ヒノキチオールがカビに有効である」のような文にマッチさせることで得られている。また、重要なポイントとして、「A が B に有効である」以外にも同じ意味的関係を表すパターンは存在する。こうした例として「A が B に効く」「A が B に効果がある」といった表現があるが、我々が開発した技術を使えば、「A が B に有効である」というパターンと「A が B に効く」というパターンが同じ意味を持つことを認識でき、そうした同義のパターンを自動発見して、さらに大量の対策の関係を持つ語のペアを抽出することができる。

このようなアルゴリズムを使って Web から意味的関係を抽出するということは、多くの Web ユーザが提供したいわば集合知を抽出するということになる。この集合知は我々一人一人が持つ知識を簡単に凌駕する。例えば、執筆者にとっては、銀イオンがカビの対策になるな

どということはまったく思いもつかないことであった。これから述べる構想はそうした集合知を高度に活用し、イノベーションを加速させようという試みでもある。

さて、こうした意味的関係の抽出において、現時点では、あくまで関係を持つ二つの語だけが考慮されている。例えば、上に挙げた意味的関係の例、「洗濯機」とそのトラブルとしての「アトピー」、さらに「アトピー」の原因としての「カビ」、さらには「カビ」の対策としての「光触媒」の３つの関係が一つのシナリオの中で位置づけられ、「アトピーの原因となる洗濯機のカビを光触媒でやっつける」というシナリオ、アイディアが Web 上に実際にあるかどうかは、概念辞書を見るだけではわからない。そうしたアイディアが既にあるかどうかを確認するためには、その意味的関係にある「洗濯機」「アトピー」「カビ」「光触媒」の４つの単語を検索エンジンに入れ、これらを全部含むような文書が存在するかどうか確認する必要がある。こうした検索を実際に行なってみると、光触媒によって、カビをクリーニングする洗濯機の特許が見つかり、これがアトピーなどに効果があることがうたわれている。もちろん、これはすでに特許として出願済みであって、新たなイノベーションを生み出す情報とは考えられないが、まだメジャーになっていない新規なアイディアのようである。

同様に、カビの対策である銀イオンと洗濯機、カビ、アトピーの関連を、検索エンジンを使って調べると、やはり、銀イオンを利用したカビに強い洗濯機がすでに商品化されていることがわかる。さらには、ヒノキから抽出されるという薬品、ヒノキチオールに関しても洗濯機のカビを抑えるための洗浄剤が実際に商品化されていることがわかる。一方で、やはり、カビの対策としての安息香酸については、洗濯機との関連は本稿の執筆者が調べた限りでは見つからず、実現にあたっては様々な困難があるかもしれないが、次なるイノベーションのための出発点としては検討しても良い新規なアイディアといっても良いであろう。

ここで行なわれていることをすこし別の角度から見れば、「なんらかの製品の「トラブル」の「原因」に対する「対策」を施すことは新たな商品開発に結びつくかもしれない」という発想に基づき、具体的に、それらの関係を持つ語を組み合わせることで、アイディア、シナリオを生成するということと言っても良い。その内のいくつかのアイディアについては、それらに基づく特許、商品がすでに存在することがわかったが、そうした特許、商品等がまだ存在しない新規なアイディアもいくつか見つかった。これらの内、実際に有効なアイディアは、数は少ないかもしれない。しかしながら、これは、概念辞書というデータベースに含まれるデータの単純な組み合わせによって、アイディアを自動的に生成するという、これまでにない方法で見つかったアイディアであり、今後こうしたアイディアは人の頭を一切煩わせることなく、Web 中の集合知を利用して無数に作ることができる。こうした自動的な組み合わせは、詰め将棋の指し手の候補を生成するように、いわば、システマティックに行なうことができるのであって、人がアイディアを出す時に陥りがちな見落としを防ぐことが可能になる他、集合知を利用することで、そもそもイノベーションを行なおうとしている一個人が知らない情報に基づいて、その個人の頭を使うこと無く、アイディアを出力することすら可能になる訳である。

また、概念辞書には、同じ文脈で使われることが多

い単語のリストが含まれているが、このリストには意味的に類似した単語が含まれていることが多い。概念辞書には、完全に自動的な処理によって、１００万語に対して、そうした単語のリストが付与されている。これによれば、銀イオンには銅イオンが、また、ヒノキチオールには緑茶カテキンが意味的に類似していそうだということがわかる。この類似性を利用すると、例えば、洗濯機のカビの対策に銅イオンを利用する、あるいは、緑茶カテキンが利用できるかもしれないという仮説を生み出すことができる。実際に、銅イオンのケースを Web 検索で探してみると、商品化はされておらず、その有効性は不明であるが、10 円玉を洗濯機に入れることでカビを防止するというアイディアが Web 上で披露されている。一方で、緑茶カテキンに関しては目立つようなアイディアは出ておらず、素人考えではあるが、今後商品化の出発点となるかもしれない。こうした意味的な類似を利用する方法は、いわゆるアナロジーに相当するものである。これまで、アナロジーとは人間の頭脳の中でしか行なわれていなかった。ところが、以上のような例は、このアナロジーを計算機上で行なうことが可能であり、やはり、自動的にアナロジーに基づくアイディアを無数に生成することが可能となり、見落としを防ぎつつ効率よくイノベーションを推進することにつながると考えている。

5 イノベーション支援システムに向けて

さて、ここで、もう少し大きな視点から上に述べたような具体例が何を意味しているのか、さらに今後こうした概念辞書を用いてどのようなシステムを構築し、イノベーションの支援を行なうべきかを検討してみたい。以上の具体例では、概念辞書中の意味的関係の組み合わせによって、一連の語で表される新しいアイディアの候補をシステマティックかつ自動的に生成できる可能性を示した。これはイノベーションを「生成とテスト」アルゴリズムによって実現するとすれば、まさに「生成」のステップに対応する。この生成されたアイディアは、上の具体例で見たように、同時に Web 検索にかけることで、既に特許申請されているもの、あるいは既に商品化されているもの等をふるい落とし、新規性という点で検討に値するものだけをユーザに提示することも可能である。(実際には、Web 文書の構造をある程度分析しないと、生成されたアイディアに現れる単語が、Web 文書中では一つのアイディア中の要素として記述されているかどうかは判別できず、今後若干の研究を要する。複数の単語が一文書に現われているものの、実際には全く別の複数のアイディアの記述に現れているだけであるというのも往々にしてある。こうしたケースを自動判別するのは我々の今後の研究課題である。)

一方で、このように自動生成されたアイディアを実際のイノベーションに結びつける際には、最低、思考実験、さらには試作／検証というプロセスが必要である。前述したように、このプロセスには人間が関与せざるを得ず、生成されるアイディアが非常に大量にある場合には、明らかに実施不可能となる。こうした状況においても有効なイノベーション支援を行なうためには、生成されるアイディアをより精密な方法でフィルタリングすること、あるいは、有望そうなアイディアが上位に来るようなランキング手法を開発することが必要となる。例え

ば、アイディアとしてある薬物を洗濯機のカビ防止に役立てるというシナリオが生成された場合に、その薬物に関して副作用がないのか、どうか、概念辞書に今後含める予定の他の意味的関係を見ることでチェックし、もし副作用があるならばアイディア候補から除くなどの処理が考えられよう。また、同様に洗濯機のカビ防止のシナリオのランキングでは、使用する薬物、物質の値段、入手のしやすさなどをやはり、概念辞書や Web 上の情報を用いて計算する方法などが考えられる。今後は、こうした利用に耐えるように、概念辞書や、Web 上の情報の解析手法を開発していく予定である。

ここで、重要なポイントとして、以上に述べてきたようなシステムの構想では、イノベーションのシナリオで使われる薬品などの「実物」が一切現れないということである。処理はすべてテキスト上で行なわれる。こうした点は、イノベーションの現場におられる方々からすればおそらく奇異に感じられることと思われる。イノベーションにおいては、アイディアで使われるアイテムの手触り、外見といった物理的要素も大きな役割を果たすはずで、テキスト情報だけで何ができるのか、というご意見もあろう。これに関しては、我々はなんら反論はない。ただ、今後イノベーションを加速するにあたっては、本稿で述べたようなシステマティックなアイディアの生成、検討は望ましいと考えている。そうしたアイディアをアイテムの物理的要素を考慮にいれて行なおうとすれば、そうしたアイテムの物理的観点からのデータベース化は必須である。ところが、現状、物理的観点からの物質等のデータベース化はかなり限られたものでしかなく、また、一旦データベース化されてしまえば、画像等の情報を除き、そうした物理的データであってもテキスト情報でしかなくなる。したがって、仮に広い範囲のアイテムを視野に入れたシステマティックなアイディアの生成を行なおうとすれば、少なくともここしばらくは、いずれにせよ、テキスト上の情報を使うより他ないと考えている。また、Web 上のテキスト情報も往々にして非常に詳細なものを含んでいる。例えば、例で用いた、ヒノキチオールであれば、その沸点、融点、水溶性などは Wikipedia に記述されている。こうした豊かな Web 上の情報をイノベーション等の重要なプロセスで有効活用するのは重要な研究テーマであろう。

6 さいごに

本稿では、Web から自動的に構築された概念辞書と呼ばれるデータベースを用いて、イノベーションの元となるアイディアをシステマティックに生成し、イノベーションの加速に役立てるという構想について述べた。イノベーションの重要性は論を待たず、ビジネス、工学、その他のありとあらゆる分野でイノベーションを加速していく必要があろう。Web という史上最大の情報源をそれに役立てるというのは非常に重要な研究テーマであると考えている。今後は特許等の分析、解析なども行ないつつ、そうした研究を推進していきたいと考えている。なお、本稿で述べた概念辞書は、その一部あるいはそれを自動構築するツール群を ALAGIN フォーラムにて、配信していく予定である。ご興味のある読者は ALAGIN フォーラムの Web サイト (http://www.alagin.jp) をご覧ください。

参考文献

1. *キーワードサーチを越える情報爆発サーチ - 自然言語処理で価値ある未知をマイニング-*, 鳥澤健太郎, 中川裕志, 黒橋禎夫, 乾健太郎, 吉岡真治, 藤井敦, 喜連川優 , 情報処理学会学会誌「情報爆発」特集号 , Vol.49, No.8, pp. 12?18, 2008.
2. *Torishiki-kai, an Autogenerated Web Search Directory*, Kentaro Torisawa, Stijn De Saeger, Yasunori Kakizawa Jun'ichi Kazama, Masaki Murata, Daisuke Noguchi, and Asuka Sumida, To appear in Proc. of ISUC 2008, pp. 179-186, Osaka, Japan, December, 2008.