



Review Paper

論文アブストラクトから原子分子の状態の情報を
検出，抽出する方法の研究

佐々木 明，村田真樹¹⁾，金丸敏幸¹⁾，白土 保¹⁾，
井佐原 均¹⁾，上島 豊，山極 満

(日本原子力研究所関西研究所光量子科学研究センター，情報通信研究機構情報通信部門¹⁾)

Methods to Extract Information on the Atomic and Molecular States from Scientific Abstracts

SASAKI Akira, MURATA Masaki¹⁾, KANAMARU Toshiyuki¹⁾, SHIRADO Tamotsu¹⁾,
ISAHARA Hitoshi¹⁾, UESHIMA Yutaka and YAMAGIWA Mitsuru

*Advanced Photon Research Center, Kansai Research Establishment, Japan Atomic Energy Research Institute, 8-1 Umemidai,
Kizu-cho, Soraku-gun, Kyoto, 619-0215 Japan*

¹⁾*National Institute of Information and Communications Technology, Information and Network Systems Department, 3-5 Hikaridai,
Seika-cho, Soraku-gun, Kyoto 619-0289, Japan*

(Received 25 May 2005 / Accepted 15 August 2005)

We propose a new application of information technology to recognize and extract expressions of atomic and molecular states from electrical forms of scientific abstracts. Present results will help scientists to understand atomic states as well as the physics discussed in the articles. Combining with the internet search engines, it will make one possible to collect not only atomic and molecular data but broader scientific information over a wide range of research fields.

Keywords:

atomic and molecular data, database, text mining, language processing

1. はじめに

核融合，天文学などの基礎科学，プラズマプロセス，医療，環境などのさまざまな応用研究において原子分子過程のシミュレーションは広く用いられているが，正確なシミュレーションを行うためには，原子分子のエネルギーレベルや，電子や光との相互作用のレート係数などの素過程データが必要である[1]．

原子分子データは，物理学の基本データのひとつとして，NIST（アメリカ標準局）や核融合科学研究所，日本原子力研究所などに置かれたデータセンターにおいて，収集・評価，データベース化が行われている．そして，収集されたデータは，データ集として出版されたり，インターネットを通じて公開されている[2]．しかし，原子分子データはその種類や量がほとんど無限と言えるほど多いため，原子分子データの理論計算，実験による測定を限られたデータセンターだけで行うことは困難であり，原子分子物理学コミュニティに属する世界中の多くの専門家の参加，協力を得ることが不可欠である．実際には，原子分子物理学の専門家が論文として投稿しコミュニティの審査を経て発表されたデータが，応用研究に利用されるようになって

したがって，原子分子データを利用するためには，多くの学術雑誌から情報を収集する必要がある．ところが，従来は対象のデータが記載されている論文を見つけ，データを抽出する過程はすべて手作業で行われていた．最近になって，学術雑誌はオンライン化が進み，キーワードや雑誌情報によって論文を検索するシステムが提供されるようになり[3]，Webサイトを通じて論文を検索し，閲覧，入手することは一般的になったが[4]，論文に必要な原子分子データが記載されているかどうか，そのデータが正確かどうかの判断は，以前同様人間に委ねられている．

情報科学の分野では，より積極的に，Web ページの内容を分析し，利用者に役立つ情報を取り出す研究が進んでいる[5]．例えばニュース記事であれば，日時，場所，人などが固有情報，重要情報であり，これを知ることが記事の内容を理解する助けとなる．これらの情報は，記事の中でそれぞれ特有なパターンに従って記述されると考えられる．そして，パターンが明白な書式で表される場合だけでなく，文章の文法的，意味的な関係によって初めて明らかになる場合も，機械学習によって認識可能であると考えられている[6]．

本研究ではこのような情報科学，言語処理の分野の成果

author's e-mail: sakaki@apr.jaeri.go.jp

を、原子分子データの収集に応用する可能性に着目した。原子分子データが記述されている論文には、必ずどのような原子、分子およびどのような素過程が対象であるかが書かれているはずである。すなわち、原子・分子の状態に関する情報が、原子分子データが記載されている論文を特徴づける固有情報であると考え、電子化されたアブストラクト中から検出、抽出する研究を行った。そして、抽出した情報を手がかりにして、原子分子データが記載されている論文の発見に応用する可能性について考察した。

一方で、言語処理の結果の精度や信頼性には人間の書く文章の持つ曖昧さなどに起因する本質的な限界があり[7]、人間による作業の代替というよりも、人間による文章の読解を支援する立場をとることの重要性が指摘されている。本研究も原子分子データ収集において人間の判断を助ける情報処理技術の可能性を示すことをめざしている。

2. 原子分子物理の重要情報の抽出の方法

本研究では、原子分子の状態を表す情報が、電子文書中で所定の規則に従って記述されることに注目し、原子と簡単な分子の情報を抽出する方法の検討を行った。

原子分子物理の論文では、原子分子の状態は、歴史的に決められた規則に従った表現で記述されている[8]。すなわち、原子の状態を指定する主な要素は、原子種、イオン種、電子配置、微細構造である。原子種は文字で書き下される場合もあるが、普通はよく知られているように1文字か2文字の英文字で表す元素記号によって、H, He, Li...のように表される。常温では、原子は同数の陽子と電子を持ち中性の状態にあるが、原子物理ではそれから1個ずつ電子を取り去ったイオン種が個別に興味の対象となる。イオン種は、

Li⁺ (1a)

Li II (1b)

heliumlike Li, He-like Li (1c)

のように書かれる。

(1a)-(1c)は、イオン種の表現が多様性を持っていることを示す。(1a)は(元素記号)^{イオン価数+}の形で単純にイオンの価数を示すのに対し、(1b)は分光学的表記と言い、中性状態を1として、イオン価数+1の値をローマ数字で表す。(1c)は対象となるイオンが2個の電子を持ち、中性Heと等しいという、He等電子系列に属するイオンであることを示す。

原子の状態は、次に電子配置によって決められる。電子配置は、各軌道とその占有電子数の並びで、

(軌道1)(占有電子数)(軌道2)(占有電子数)... (2)

のように表される。電子軌道の表現には、全角運動量 j で区別する相対論的表記と区別しない非相対論的表記の二通りがあり、一般に原子番号が大きく、価数が多いほど相対論的効果が大きくなり、原子の性質を表現するために相対論的表記が必要になる。非相対論的表記での軌道は、主量子数と方位量子数を表す1文字で、1s, 2s, 2p...のように

表される。

さらに詳細な原子の微細構造は、スペクトル項によって表される、LS結合を仮定する場合は、

(多重度)(軌道角運動量を表す一文字)(全角運動量) (3)

と表される。軌道角運動量はS, P, D, F...等で表される。相対論的効果の大きい原子・イオン種では、スペクトル項もLS結合よりも jj 結合で表す方が適切である。

これらの記号や表現は、原子物理学や分光学の進展に伴って決められたもので、原子・イオン種、電子配置、スペクトル項のそれぞれについて、原則として与えられた状態に対して一意に決まるように定義されている。しかし、その表現はあくまでも人間が理解することを前提とするので、表現を簡単にするために一部が省略されて表されることや、原子核物理学のような異分野で似た表現が別の意味を持つ場合があり、表現の定義が複雑になる。

例えば中性Naの基底状態の電子配置を、

1s² 2s² 2p⁶ 3s (4)

とは書かずに、1sから2pまでの軌道を省略し、

3s

とだけ書くことがある。これは、原子の性質が主に主量子数の最も大きい最外殻の軌道の電子数によって決まり、電子で満たされている内側の閉殻の影響が少ないという性質に基づいている。このことは、論文中の表現から真の電子配置、原子の状態を特定するために文脈からの推論が必要な場合があることを示唆する。

原子状態の集合を表すためにも表記の省略が行われる。一般に原子の状態は階層構造を持ち、ひとつの原子・イオン種には多くの電子配置で表される状態があり、ひとつの電子配置には多くの微細構造が対応する。論文の中でひとつの原子・イオン種あるいは電子配置に属する複数の状態のデータに言及する際には、個々の状態をすべて書き下すよりも、それを階層構造に従ってまとめて記述する方が普通である。

一方、原子状態を表すそれぞれの表現の中に現れる英字や数字には、物理的な制約がある。元素はたかだか100種類であり、存在し得るイオンの範囲は、まれに現れる負イオンを除くと、中性原子から軌道電子をすべて取り去った完全電離イオンまで、0から原子番号の範囲に限られる。電子配置に現れる主量子数 n 、方位量子数 l 、占有電子数 n_e の間には、

$$0 \leq l \leq n-1$$

$$0 \leq n_e \leq 2(2l+1)$$

で表される制限がある。文字の並びから得られる情報と原子物理学の知識を組み合わせることによって、認識の精度を高めることができると考えられる。

このような原子の状態の記述は複雑であり、原子分子物理学の専門家以外の人にとってはわかりにくいものである。しかし、HTMLで書かれた電子文書では、原子・イオ

ン種、電子配置、スペクトル項それぞれの記述を英数字と上付、下付文字を表す<sup>, <sub>などのタグの並びとして定義することができて、コンピュータプログラムで認識できると考えられる。

3. 原子状態の論文アブストラクトからの抽出の実際

本研究では、以上のような原子状態を記述する表現の特徴を利用し、HTMLで書かれたテスト用の論文アブストラクト中から、原子・イオン種、電子配置、スペクトル項などの固有情報を検出し、抽出するとともに、結果をwebブラウザ画面上で色分けしてマーキングして表示されるようにし、見た人がアブストラクトの中で扱われている原子の状態を理解、判別することを支援するソフトウェアを作成した。

実際にはインターネットから入手したアブストラクトには、タイトル、著者などの書誌情報の他、webページとしてのタイトルなどの情報、ナビゲーションのための情報や、表示の体裁を整えるための要素が含まれているので、物理的に意味のある記述が含まれていると考えられる、タイトルとアブストラクトの部分を取り出した上で固有情報の抽出を行った。

具体的なソフトウェアの作成ではperl言語の正規表現を用いた。正規表現とは、テキストの中の文字の並びの規則を一般化したもので、電子文書の中の情報の検索、抽出、処理のために広く使用されている。そこで、2章にまとめた原子物理における原子状態の記述のルールと、実際の論文アブストラクトに現れる記述をそれぞれ分析し、原子・イオン種、電子配置、スペクトル項に対応する正規表現を決定した。

このうち、原子・イオン種、電子配置の検出に用いたルールは以下の通りである。

原子・イオン種には、HeやInなどのように英語単語と重なるものがあるため、いったん大まかな規則で抽出を行ったあと、原子・イオン種でなさそうなものを除外するようにした。すなわち、はじめに、

「e+/- あってもよい 上下付文字あってもよい
原子名 上下付文字あってもよい
IVXivxの表現があってもよい “like”/“ic”が
あってもよい」の一回以上の連続 (5)

という規則で抽出を行った。ただし、式中の/は並びの中のいずれかひとつを示す。規則は、必須の原子名の前後にいろいろな修飾を加えたものとして定義される。原子名とは予め記憶させておく元素記号と名称とその別名を指す。最初のeの表現は、今回対象としたテスト論文が主に電子衝突過程を扱っていて、e+Xの形式の表現がよく現れるのを認識するために加えた。IVXivxの表現とは、

「I/V/X/i/v/x」の0回以上の連続

を指し、価数をローマ数字で表す分光学的表現に対応する。like/icの表現は等電子系列に対応する。

次に、抽出した表現のうち、以下を原子・イオン種でなさそうなものとして除外した。

- ・英語アルファベット大文字一文字
- ・英語アルファベット大文字一文字の連続
- ・Rev (Phys. Rev.などの雑誌名にしばしば出現)

については、周辺に原子・イオン種であることを示す手がかり表現 (like/ic) がない場合は除外し、

- ・文頭のAs, In, At, Heなど

は、前置詞、代名詞の可能性が高いとして除外した。あとで示すように、これらの追加的な規則により、論文中に現れる原子・イオン種とそうでないものはほぼ確実に区別できることがわかった。

また、(5)式において原子名に添字があっても良く、表現が複数回繰り返しても良いとして、同じルールによってO₂, NH₄などのような簡単な構造の分子の表現も認識できるようにした。

一方、(2)および(4)式に示した電子配置は、

「数字あってもよい s/p/d/f/g 上下付文字あってもよい」の1回以上の連続 (6)

と書くことができる。(6)式における数字は電子の主量子数に、上下付文字がそれぞれ占有電子数と相対論的表記における全角運動量jの記述に対応する。非相対論的、相対論的表記をあえて区別しないのであれば、ひとつの正規表現でまとめて指定できることを示す。対応するperl言語の正規表現は、

```
my $rule = "(?:${x}¥¥d*${x}(?:[spdfg]))$
            {x}${y}${x})+"; (7a)
```

```
my $x = "(?:<¥¥/?(?:I|RM|i|rm|B|b|SMALL|
            Small|small)>|
            <(?:FONT|Font|font)[^<>]*><¥¥/
            (?:FONT|Font|font)>|
            <(?:SUP|SUB|sup|sub)>¥¥s+<¥¥/(?:
            SUP|SUB|sup|sub)>¥¥&nbsp;sp¥¥;)*"; (7b)
```

```
my $y = "(?:${x}<(?:SUP|sup)>${x}(?:${x}
            (?:[ ^<]+)${x})) *${x}
            <¥¥/(?:SUP|sup)>${x}|${x}<(?:SUB
            |sub)>${x}(?:${x}
            (?:[ ^<]+)${x})) *${x}<¥¥/(?:SUB|
            sub)>${x}) *"; (7c)
```

などと表される。ここで「あってもよい」とは、その表現が0回以上連続して現れても良いという意味を指し、正規表現では*記号で表される。また、1回以上の連続は+記号で表される。(7a)式において、[spdfg]は括弧の中の1文字のいずれか(方位量子数)に、¥¥d*は主量子数に、\${y}は占有電子数に対応する。この際、HTML文書では上付、下付文字はそれぞれ<sup>および<sub>タグで指定され、しかも雑誌によってその書式が大文字の場合と小文字の場合があったり、フォント、サイズなどの指定が加わる場合、余計なスペースが加わる場合がある。多くの雑誌を

対象として正しい認識を行うには、すべての書式を網羅する必要があるため、規則が複雑になる傾向がある。

このように、本研究では原子状態を表す表現をすべて規則(ヒューリスティックルール)として定義して、テスト用アブストラクトの中からこれらの固有情報の抽出を行った。

Fig. 1 は、In, Xe イオンの電子衝突電離断面積の測定に

関する論文のアブストラクトについて、原子状態を表す表現の検出を行い、結果を web 画面に表示した文章において、文字の背景色を、原子・イオン種および簡単な分子種をピンク色で、電子配置を黄色で表して強調して表示した例である。灰色はいったん原子・イオン種と認識されたが適当でないと判断して棄却したものである。図はイオン種

(a)



☐ View Page Images or PDF (992 kB)

Electron-impact ionization of In^+ and Xe^+

E. W. Bell, N. Djuric, and G. H. Dunn

Joint Institute for Laboratory Astrophysics, University of Colorado National Institute of Standards and Technology, Boulder, Colorado 80309-0440

Received 14 June 1993

Absolute ionization cross sections for In^+ and Xe^+ by electron impact have been measured from below threshold to 200 eV using the crossed-beams technique. The cross sections for In^+ were possibly enhanced by indirect ionization processes. The excitation of the ion from the $4d^{10}5s^2$ ground state to the $4d^95s^25p$ state followed by autoionization has been postulated. The In^+ cross sections show a peak value of $15.9 \times 10^{-17} \text{ cm}^2$ at about 80 eV. The cross sections for Xe^+ peak at a value of $25.6 \times 10^{-17} \text{ cm}^2$ at about 35 eV. Experimental measurements are compared to configuration-averaged distorted-wave calculations [M. S. Pindzola *et al.*, J. Phys. B **16**, L355 (1983)], the semiempirical formula of Lotz [Z. Phys. **216**, 241 (1968)], and, in the case of Xe^+ , previous experimental results [C. Achenbach *et al.*, J. Phys. B **17**, 1405 (1984)]. Also presented are ionization-rate coefficients and fitting parameters for both ions for temperatures in the range $10^4 \text{ K} \leq T \leq 10^7 \text{ K}$.

©1993 The American Physical Society

URL: <http://link.aps.org/abstract/PRA/v48/p4286>

DOI: 10.1103/PhysRevA.48.4286

PACS: 34.80.Kw

(b)

Title:

Electron-impact ionization of In^+ and Xe^+

Abstract:

Absolute ionization cross sections for In^+ and Xe^+ by electron impact have been measured from below threshold to 200 eV using the crossed-beams technique. The cross sections for In^+ were possibly enhanced by indirect ionization processes. The excitation of the ion from the $4d^{10}5s^2$ ground state to the $4d^95s^25p$ state followed by autoionization has been postulated. The In^+ cross sections show a peak value of $15.9 \times 10^{-17} \text{ cm}^2$ at about 80 eV. The cross sections for Xe^+ peak at a value of $25.6 \times 10^{-17} \text{ cm}^2$ at about 35 eV. Experimental measurements are compared to configuration-averaged distorted-wave calculations [M. S. Pindzola *et al.*, J. Phys. B **16**, L355 (1983)], the semiempirical formula of Lotz [Z. Phys. **216**, 241 (1968)], and, in the case of Xe^+ , previous experimental results [C. Achenbach *et al.*, J. Phys. B **17**, 1405 (1984)]. Also presented are ionization-rate coefficients and fitting parameters for both ions for temperatures in the range $10^4 \text{ K} \leq T \leq 10^7 \text{ K}$.

Fig. 1 An example of an abstract (from E.W. Bell, N. Djuric, and G.H. Dunn, Phys. Rev. A48, 4286 (1993)) on atomic data, (a) on the web page of American Physical Society (<http://www.aps.org>), and (b) with atomic states and configuration marked and color coded on the text. Words colored pink correspond to atomic and ionic state, and yellow correspond to electron configuration.

と電子配置の情報が正しく検出されていることを示す。

このような文書処理のソフトウェアが意図した通りに動作しているかどうかを確かめるには、設定した規則を網羅するだけの様々な種類の原子状態の記述を含むテスト用文書を実際に処理して、結果の正しさを判定する必要がある。本研究では、テスト用文書として市川[9]が収集した、原子分子と電子の衝突による電離や励起過程のデータを含む論文の文献データベースに収められている論文のアブストラクトを用いた。

次に、抽出した原子・イオン種、電子配置、微細構造の情報について適合率 (Precision) と再現率 (Recall) を評価した。適合率、再現率はこのような情報抽出ソフトウェアの性能を評価するために用いられる一般的な尺度である[10]。今回の例では、原子・イオン種の表現の抽出の適合率とは、ソフトウェアが抽出した表現のうちで正しく原子・イオン種の表現であったものの割合、再現率とは対象とする全文書の中の原子・イオン種の表現のうちでソフトウェアによって抽出できたものの割合である。結果は Table 1 のように原子・イオン種や電子配置の情報をほぼ漏れなく検出できることを示す。微細構造の情報について再現率が低いのは、分光記号 S , P , F ... を原子の硫黄、燐、フッ素と誤るためであることがわかった。

4. 原子状態の抽出ソフトウェアの応用可能性

このように、コンピュータプログラムを用いて、論文アブストラクト中の原子状態の情報を高い精度で抽出できることがわかったので、この情報を、原子分子データが含まれている論文の発見に応用する可能性について検討した。

はじめに、原子状態の情報の存在が、その論文に原子分子データが記述されているかどうかを判断するために役立つかどうかについて検討を行った。そのためには、前述の電子衝突過程に関する論文のアブストラクト336点に加えて、Phys. Rev. A~E 誌各1号分 (vol. 69, No.1) をテスト文書として選び、原子状態の情報が文書中に現れるかどうかを調べた。結果は Table 2 のように、電子衝突過程に関する論文のアブストラクトの中にはほとんど必ず原子状態の情報が現れるのに対して、Phys. Rev. A~E 誌では雑誌によりそれぞれ出現の確率が異なり、統計、非線形物理、ソフトマターなどを扱い、原子分子物理と関係が薄いと考えられる Phys. Rev. E 誌には予想どおり稀にしか原子状態の表現が現れないことがわかった。

一方、今回開発したソフトウェアは、凝縮系物理を扱う

Table 1 Accuracy and reproduction of the extracted expressions of atomic and molecular state, configuration and fine structure.

	precision	recall
atomic species	81.5% (274/336)	98.6% (274/278)
electron configuration	98.6% (141/143)	99.3% (141/142)
fine structure	100% (27/27)	40.3% (27/67)

Table 2 Probability of appearance of the expressions of atomic and molecular state in different journals.

journals	number of samples and probability	
abstracts on atomic data[9]	142/147	97%
Phys. Rev. A (atoms, molecules and photons)	23/86	27%
Phys. Rev. B (condensed matter)	55/88	63%
Phys. Rev. C (nuclear physics)	26/50	52%
Phys. Rev. D (particle physics, gravity, space)	28/67	42%
Phys. Rev. E (non-linear, statistical physics, soft matter)	11/151	7%

Phys. Rev. B 誌や、原子核物理を扱う Phys. Rev. C 誌にも原子状態の表現が多く含まれるという結果を示した。抽出の内容を検討したところ、Phys. Rev. B 誌の場合は分子の表現が多く抽出されていること、Phys. Rev. C 誌の場合は、核種の表現が抽出されていることがわかった。原子核物理分野では、 ^{32}S のように左の肩字で同位体の原子量を表し、これと分光記号の S 項とは形式上同じである。 S , P , F などは元素記号としても分光記号としてもしばしば現れるので区別する必要がある。これらの元素については、存在し得る原子量の範囲が、 S では32~36, P では30~33, F では18~19であり、分光記号の左の肩字に現れる多重度とは取りえる値の範囲がかけ離れていることから識別ができると考えられるが、それには原子(核)物理学の知識が必要である。

さらに、抽出された原子状態の情報をさらに論文検索、抽出に利用するために結果の意味の解釈についての検討を行った。今回開発したプログラムは、原子状態の表現の多様性を考慮し、原子・イオン種の複数の表現を認識する。しかし、その結果を利用して新しい検索を行う場合を考えると、複数の表現が同一であることを理解すること、あるいは複数の表現のうちで基本的な共通な表現(基底表現)に変換して処理を行う必要があると考えられる。Table 3 は原子・イオン種の表現を基底表現に変換することを試みた結果である。変換前には様々な表現で書かれている原子種を元素記号で、イオン種を(元素記号)イオン価数 $^{+}$ の形に変

Table 3 Example of conversion of various expressions of atomic state into general forms.

abstract	atoms	ions	extracted expression
Phys. Rev. A62, 062720 (2000)	O	O^{5+}	Li-like O^{5+}
J. Phys. B33, L391 (2000)	Fe	Fe^{14+}	Fe XV At
J. Phys. B33, 1013 (2000)	O Be C B	B^{2+} C^{3+} O^{5+} Be^{+}	B^{2+} Li-like C^{3+} O^{5+} Be^{+}
Astron. Astrophys. Ser. 142, 317 (2000)	Fe	Fe^{7+}	Fe^{7+}
J. Phys. B33, 3449 (2000)	S Ca Fe	S^{11+} Fe^{24+} Ca^{18+}	S^{11+} He-like Fe^{24+} Ca^{18+} At
J. Phys. B33, 597 (2000)	Ne	Ne^{2+}	neon $^1\text{S}_0$ Ne III $^1\text{S}_0$ $^3\text{P}_{2,1,0}$
Phys. Rev. A62, 062711 (2000)	Mg	Mg^{10+} Mg^{11+}	hydrogenic $^1\text{S}_0$ Mg^{10+} Mg^{11+}

換した例を示す。

5. 議論とまとめ

本研究では、電子化された科学論文のアブストラクトを解析して、原子状態を表す重要情報を抽出し、web 画面上に表示するソフトウェアの開発とその評価を行った。作成したソフトウェアはアブストラクトの原子・イオン種の情報をほぼ漏れなく検出することができることを示した。

本研究で開発したソフトウェアは、分子に関しては、原子の記述の拡張による単純な記述によるものに限り認識することができる。さらに分子の名前を辞書化することと、軌道角運動量を $\Sigma, \Pi, \Delta, \dots$ などで表す分子についての分光記号を取り入れる拡張を行えば、原子分子物理学でしばしば研究の対象になるような、分子と電子、イオン、光などとの相互作用の記述を認識できるようになると考えられる。一方、化学や生物学で興味の対象となる高分子については、表記方法が複雑であるばかりでなく、種類が極めて大きいのでそれを処理するにはデータベースと組み合わせた新たな手法を取り入れる必要があると考えられる。もし対象が TeX 文書であれば、ChemTex[11]のような化学式の記述方法を認識に利用できると考えられる。一方で、新しい化学物質に対する命名規則[12]を計算機で認識できるようルール化するか、もしくは機械学習することにより、論文中に未知の語、名前が現れた時に、それがどのような分子構造に対応するかを推測する方法も考えられる。

このように、情報の検出、抽出の精度をさらに高めるためには、原子分子物理学の知識を利用して、記述されている内容を理解することと、同時に文章の文法的な性質、現れる用語の確率的な性質を利用することが必要と考えられ、原子分子物理学と言語・情報処理のいっそうの協力が重要と考えられる。

本研究の結果は、現在一般に行われているインターネットの検索エンジンを利用した研究情報の収集を進めて、専門知識を利用した能動的な情報の分析が行える可能性を示す。今回開発したソフトウェアを web ブラウザ/検索システムと組み合わせれば[13]、本研究の成果は、広く原子分子過程に関する論文の読解を助ける目的に役立つと考えられる。

さらに研究を論文中の図表から画像処理によって具体的に数値データを読み取る処理や、データの信頼性の評価へと発展させることにより、原子分子データの収集や、データベース化の支援へと発展させることも可能と考えられる。後者のためには対象のデータがどれだけ他の研究で利用されているかを調べる計量書誌学的手法も役立つと考えられる。

現在、原子分子物理学の研究とその応用研究は世界的規模で進展し、とりわけ中国や韓国をはじめとするアジア諸国での科学技術の進展も著しい。今後は最新の研究結果が中国語や韓国語で発表される機会も増加すると考えられる。このような多言語の論文を読みこなす語学力を得るこ

とは人間の研究者には困難である。それに対して、本研究で示した研究情報の計算機による処理は、論文の記述そのものは言語に依存するが、抽出される原子分子データはより抽象的であることに注目するものであり、多言語で記述される研究情報の収集にも発展させられる可能性がある。逆に海外の研究者にそのようなシステムが普及すれば、日本語による研究成果の普及、日本における研究活動の優先権を主張するために役立つ可能性も考えられる。

謝辞

本研究を行うに当たり、奈良女子大城和貴教授、柏木裕恵さん、会津大学 Lukas Pichl 講師、鈴木学さん、核融合科学研究所加藤隆子教授、加藤太治助手との討論に感謝します。また、日本原子力研究所関西研究所田島俊樹所長、光量子科学研究センター木村豊秋センター長、James Koga 主任研究員、森林健悟副主任研究員、同那珂研究所炉心プラズマ研究部の久保博孝主任研究員のご理解、ご支援を受けたことに感謝します。本研究の一部は、けいはんな情報通信オープンラボの技術的援助を受け、また文部科学省科学研究費基盤研究 C「原子分子論文抄録読解データベースの研究」の補助を受けて行なわれた。

参考文献

- [1] 加藤隆子他：本誌小特集「プラズマ原子過程の展望」プラズマ・核融合学会誌 75, 1124 (1999).
- [2] 例えば核融合科学研究所 web サイト, <http://dpc.nifs.ac.jp/admrc/index-j.html> を参照.
- [3] 例えば <http://scitation.aip.org/> を参照.
- [4] 馬場 肇：「パーソナルコンピュータ (PC) の有効利用法」プラズマ・核融合学会誌 78, 237 (2002).
- [5] 難波英嗣：「WWW 上のテキスト情報の知的統合」人工知能学会誌 19, 311 (2004).
- [6] 村田真樹：「質問応答システムの現状と展望」電子情報通信学会誌 86, 959 (2003).
- [7] 長尾 真編：「自然言語処理」岩波講座ソフトウェア科学 15 (岩波書店, 1996).
- [8] ヘルツベルグ著、堀健夫訳；原子スペクトルと原子構造 (丸善, 1964).
- [9] Y. Itikawa, ADNDT 80, 117 (2002).
- [10] 北 研二、津田和彦、獅々堀正幹：情報検索アルゴリズム (共立出版, 2002).
- [11] <http://www.physik.uni-leipzig.de/Computer/latexdoc.html>.
- [12] 日本化学会、岩澤康裕編：化学便覧 (丸善, 2004).
- [13] L. Pichl, M. Suzuki, K. Joe, and A. Sasaki, "Networked mining of atomic and molecular data from electronic journal databases on the internet", Proceedings of the 4th International Workshop, DNIS 2005, LNCS 3433, pp. 159, Springer, 2004.