科学論文からの原子分子データの収集、データベース化の自動化支援に関する研究

[1]佐々木明、[2]村田真樹、[3]柏木裕恵、[3]城和貴、[4]ピフル　ルカーシュ、[5]加藤太治、[5]村上泉
原子力機構[1]、情報通信研究機構[2]、奈良女子大[3]、国際基督教大学[4]、核融合科学研究所[5]
(sasaki.akira@jaea.go.jp)

１．はじめに

原子分子の物理・化学的な特性のデータ（原子分子データ）は、さまざまな基礎研究、産業応用における重要な基礎データとして活用され、データベース化が進められている[1]。本研究は、情報処理、言語処理技術の活用によって、従来手作業によって行われてきたデータの収集・評価作業の自動化支援を行うことを目的とする。そして、原子分子データベースをより大規模で有用なものへと発展させる可能性を明らかにすることを目指している。

自然界に存在する化学物質の種類の多さからも分かるように、原子分子の状態や、それらのかかわる化学反応の速度係数などのデータは無限とも言えるほど多い。そして、基礎研究の進展や、応用上の新しい興味の対象が生まれるに伴い、大量のデータが新たに生産され、また必要とされるので、データベース構築を効率化することは重要な課題である。

原子分子データのデータベース化の作業は、通常原子分子物理学の専門家による原子分子データの理論計算や測定の結果が書かれている論文を収集すること、論文中の図、表に記載されているデータを取り出すこと、データの精度を評価することなどの要素から構成される。

本報告では、このうち論文の収集の自動化に注目して行っている研究の現状について述べる。

原子分子データが発表される論文雑誌は Phys. Rev. 誌をはじめ 20 種類程度で、その中の論文総数は 10^4 件/年のオーダーであるのに対して、収集の対象となる論文の数は 100 件/年程度である。対象論文の発見は大きな労力を必要とし、自動化できればデータベース構築の効率化に大きく役立つと考えられる。一方、収集対象が論文全体に占める割合が小さいことから、対象を漏れなく収集することが重要と考えられる。

２．論文収集の自動化支援の方法

本研究では、まずこれまで原子分子データの収集に従事してきた研究者からヒアリングを行った、その結果、対象論文の採否の判定のために、タイトルやアブストラクトの情報が重要な役割を果たしていると考えられたことから、収集の対象となる原子分子データが含まれているかどうかの判定に、アブストラクトのコンピュータによる解析を応用する可能性に着目した。

原子分子データについて記述している論文は、共通な性質を備えていると考えられるので、テキスト分類の手法によって、論文全体の集合からそれを選び出すことができると考えられる。われわれは、これまでの原子分子データベース開発を通じて、原子分子データに関する論文アブストラクトがデータベース化されていることに注目し、これを教師データとする機械学習によって、未知の論文に原子分子データが含まれているかどうかを判定することにした。まず、アブストラクトに出現する語の頻度 (tf-idf 値) を属性として用い、LVQ(学習ベクトル量子化)法で機械学習を行ったところ、適合率、再現率は 70%程度であった[2]。

判別の精度を高めるには、属性の選択が重要であり、そのためには、原子分子データ分野の専門知識を属性として活用することが重要と考えられる。専門知識は、分野に固有な用語や表現に現れると考えられる。そこで本研究では原子分子の状態や化学反応を表す記号や表記方法が原子分子データの分野に特有であることに注目し、それらを検出、抽出する方法とその応用について検討した[3]。

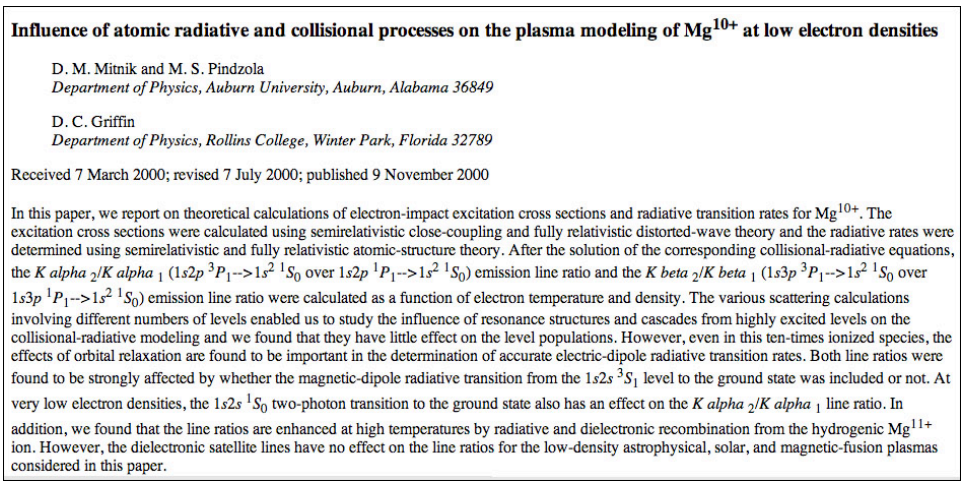
３．原子分子の状態の情報の検出、抽出

原子分子データ分野では、物質の最小単位は原子であり、物質は原子とその集まりである分子から構成されると考えられる。したがって、論文中で現象を議論する際も原子の記述が基本になる。

図 1 に原子データに関する記述を含む論文アブストラクトの例を示す。原子分子の状態に関する情報が、Mg^{10+}や $1s3p\ ^3P_1$、$1s^2\ ^2S_0$ のように、記号と上付き、下付き文字、斜体の文字などによって表されていることを示す。

科学論文では、原子の状態は、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。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、スペクトル項である。原子種は、良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。



Influence of atomic radiative and collisional processes on the plasma modeling of Mg^{10+} at low electron densities

D. M. Mitnik and M. S. Pindzola
Department of Physics, Auburn University, Auburn, Alabama 36849

D. C. Griffin
Department of Physics, Rollins College, Winter Park, Florida 32789

Received 7 March 2000; revised 7 July 2000; published 9 November 2000

In this paper, we report on theoretical calculations of electron-impact excitation cross sections and radiative transition rates for Mg^{10+}. The excitation cross sections were calculated using semirelativistic close-coupling and fully relativistic distorted-wave theory and the radiative rates were determined using semirelativistic and fully relativistic atomic-structure theory. After the solution of the corresponding collisional-radiative equations, the *K alpha* $_2$/*K alpha* $_1$ ($1s2p\ ^3P_1$-->$1s^2\ ^1S_0$ over $1s2p\ ^1P_1$-->$1s^2\ ^1S_0$) emission line ratio and the *K beta* $_2$/*K beta* $_1$ ($1s3p\ ^3P_1$-->$1s^2\ ^1S_0$ over $1s3p\ ^1P_1$-->$1s^2\ ^1S_0$) emission line ratio were calculated as a function of electron temperature and density. The various scattering calculations involving different numbers of levels enabled us to study the influence of resonance structures and cascades from highly excited levels on the collisional-radiative modeling and we found that they have little effect on the level populations. However, even in this ten-times ionized species, the effects of orbital relaxation are found to be important in the determination of accurate electric-dipole radiative transition rates. Both line ratios were found to be strongly affected by whether the magnetic-dipole radiative transition from the $1s2s\ ^3S_1$ level to the ground state was included or not. At very low electron densities, the $1s2s\ ^1S_0$ two-photon transition to the ground state also has an effect on the *K alpha* $_2$/*K alpha* $_1$ line ratio. In addition, we found that the line ratios are enhanced at high temperatures by radiative and dielectronic recombination from the hydrogenic Mg^{11+} ion. However, the dielectronic satellite lines have no effect on the line ratios for the low-density astrophysical, solar, and magnetic-fusion plasmas considered in this paper.

図１ 原子分子データに関する情報を含む論文アブストラクトの例[4]

図２に半古典論のボーアのモデルによる原子の構造の例を示す[5]。原子は原子核とそれを取り巻く電子からなり、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、原子分子物理学ではそれから１個ずつ電子を取り去ったイオン種も個別に興味の対象となる。イオン種は、Li^+, Li^{2+}, Li^{3+}...のように書かれる。

一方、原子に属する電子は 1s, 2s, 2p...などと、表される、とびとびの決まった軌道上にあると考えられている。電子軌道は整数で表される主量子数と、s、p、d、f、g... のような英字一文字で表される方位量子数の組で表される。方位量子数は、実際の原子では楕円形をなしているとされる電子軌道の長軸と短軸の長さの比に対応する。そして各軌道に何個の電子があるかを示す電子配置によって原子の状態が決まる。電子配置は、軌道とその占有電子数の並びで、

$$(\text{軌道 1})^{(\text{占有電子数})}\ (\text{軌道 2})^{(\text{占有電子数})}\ \ldots \quad (1)$$

$1s^2\ 2s^2\ 2p^2$ のように表される。

それぞれの電子は、ひとつの軌道に入ることができる電子数が方位量子数を l として $2(2l+1)$個に制限される量子力学の Pauli の排他律を満たす限り、任意の軌道に入ることができる。主量子数が大きいほど軌道半径は大きくなり、その値には理論的な上限はない。従って、その組み合わせで決まる電子配置は無限に存在し、原子分子物理ではそのような状態を区別して記述できることが求められる。

さらに、スペクトル項とは、原子が発する光のスペクトルを詳細に観測した時に見られる微細構造で、原子の詳細な内部状態に対応し、

$$^{(\text{多重度})}(\text{軌道角運動量})_{(\text{全角運動量})} \quad (2)$$

と表され、具体的には $^1S_{1/2}$, $^3P_1$ などと表記される。軌道角運動量は S, P, D, F... 等の大文字一文字で、多重度は整数、全角運動量は整数または半整数で表される。

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、例えば 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る。すなわち、原子・イオン種、電子配置、スペクトル項は記号の並びと考えることができ、正規表現によって表わして、文書中の表現を検出、抽出することができる。例えば、(1)で表した電子配置を一般化した記号の並びは、

「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　(3)

と表すことができる。

本研究では、オンラインジャーナルにある HTML 形式の論文アブストラクトの表現の分析を行って、原子、イオン種、電子配置、スペクトル項に対応する規則を求め、それによる情報の検出、抽出の性能の評価を行った。

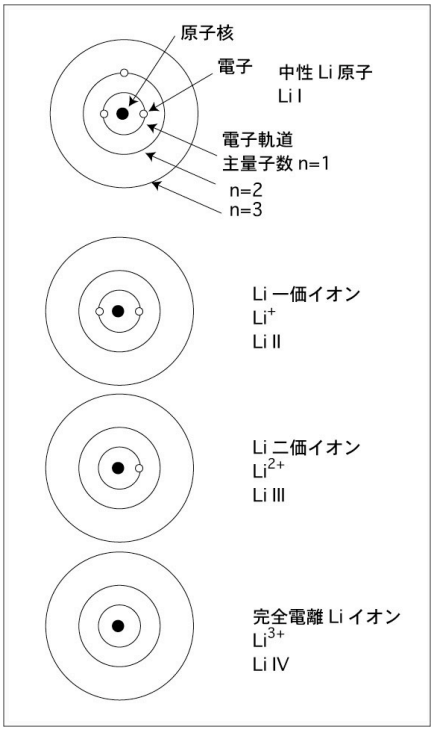


図 2　ボーアのモデルによる原子の構造とその表記方法

4．結果と考察

本研究では、市川が収集した、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アブストラクト 379 件を用い[6]、設定した規則を網羅するだけの色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、結果の正しさの判定を行った。

表 1 は、抽出した原子・イオン種、電子配置、スペクトル項の情報の適合率と再現率である。今回の例では、原子・イオン種の表現の抽出の適合率とは、抽出され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エアによって正しく抽出できたものの割合である。結果は表 1 のように、原子・イオン種や電子配置の情報を開発したソフトウエアがほぼ漏れなく検出できることを示す。スペクトル項の情報について再現率が低いのは、スペクトル項の *S*, *P*, *F*...を原子の硫黄、燐、フッ素と誤るためである。

次に、このような原子分子の表記が、論文の分類に役立つかどうかについて検討を行った。テスト文書として用いた Phys. Rev. A-E 誌一巻分(vol.69、vol.1)には、それぞれ原子・分子・光(A)、凝縮系・固体物理(B)、原子核物理(C)、素粒子物理・重力・宇宙(D)、非線形現象・統計物理(E)という、物理学の中の異なった専門分野の論文が採録されている。

表 1 原子分子データを含む論文から原子種、電子配置、スペクトル項を検出した結果

種類	適合率	再現率
原子種	81.5% (274/336)	98.6% (274/278)
電子配置	98.6% (274/278)	98.6% (141/143)
スペクトル項	100% (27/27)	40.3% (27/67)

結果は表 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、原子分子物理と関係が薄いと考えられる Phys. Rev. E 誌には予想通り稀にしか現れないことを示す。

一方、今回の結果は、Phys. Rev. B、C、D の各誌にも原子分子の状態の表現が多く含まれるという結果を示す。

表 2　雑誌毎の原子分子の記述が出現する論文の数と割合

雑誌名	論文数と割合	
市川[6]	368/379	97%
Phys. Rev. A	23/86	27%
Phys. Rev. B	55/88	63%
Phys. Rev. C	26/50	52%
Phys. Rev. D	28/67	42%
Phys. Rev. E	11/151	7%

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、Phys. Rev. C 誌の場合は核種(原子核)の表現が多く抽出されていることが分かった。

原子、イオン種とスペクトル項の記述において、左肩字は、核種の違いを表す原子量と多重度の 2 つの意味に用いられるので、*S*, *P*, *F* の原子種については、形式的には両者を区別できない。しかし、原子物理では、原子量の値は原子番号の 2 倍を中心とした限られた値であることが知られている。原子番号 16 の S では 32-36 の範囲にあるものが代表的な同位体と呼ばれる核種である、一方、多重度の値は 0 から 5 くらいまでがほとんどである。従って、肩字の数値により、^{35}S は核種、^{1}S はスペクトル項と判定できる。このような原子種と核種を判別する処理を行ったところ、50 件のアブストラクト中に現れた原子種の記述 121 件のうちの 83 件(69%)は核種の表現であることが分かった。

それに対して、Phys. Rev. D 誌に掲載されている素粒子物理学の論文では、K 中間子、B 中間子が K^+、B^0 のように原子のカリウム、ホウ素のイオンなどと

まったく同じように表現され、記号だけでは原子種と判別ができない場合があることが分かった。

さらに、原子種、スペクトル項と核種の判別を行い、同様の処理によってと誤って処理されていたスペクトル項の表現を正しく抽出し直したあとで、原子分子の各情報がそれぞれの雑誌に現れる割合を求めた結果を図3に示す。核種の表現は Phys. Rev. C 誌に偏って現れ、原子分子データが記載されている論文にはほとんど現れないことを示す。

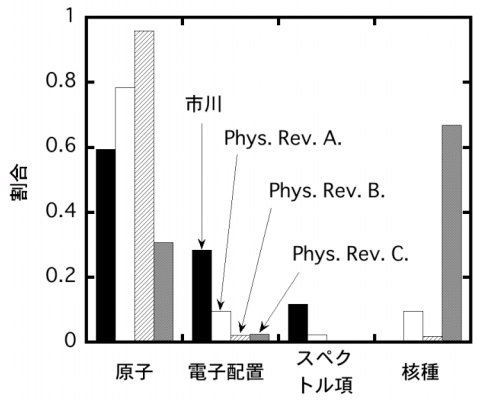


図 3 雑誌毎に出現する原子分子の記述のうちの原子、電子配置、スペクトル項、核種の情報の割合

５．まとめ

本研究では、原子分子データが含まれている論文の自動的な収集の手段の一つとして、論文中の表記の規則を利用して原子分子に関する情報を検出し、抽出する方法を開発し、評価を行った。

開発したプログラムは論文中の情報をほぼ漏れなく検出でき、さらに表現のより詳細な特徴を用いて情報を分類すると、論文雑誌毎の特徴がより明らかになることも分かった。本研究の成果を活用することでテキスト分類の精度が向上することが期待される。

抽出された情報は、論文の内容を特徴付ける重要情報である。論文中で図4のように抽出した情報にマーキングして表示することにより、研究者の理解を支援する用途などへ応用する可能性も考えられる。

謝辞　本研究は一部文科省科研費基盤研究 C の補助を受けて行われた。

参考文献

[1] “プラズマ原子・分子過程の展望”、加藤隆子他：プラズマ・核融合学会誌, **75**, 1124 (1999).
[2] “Design and Implementation of an Evolutional Data Collecting System for the Atomic and Molecular databases” 、A. Sasaki, J. Kazuki, H. Kashiwagi, C. Watanabe, M. Suzuki, L. Pichl, M. Ohishi, D. Kato, M. Kato, and T. Kato, J. Plasma Fusion Res. Ser. 7, 2005, to appear.
[3] “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、金丸敏幸、白土保、井佐原均、上島豊、山極満：プラズマ・核融合研究, **81**, 717 (2005).
[4] “Influence of atomic radiative and collisional processes on the plasma modeling of Mg^{10+} at low electron densities.” 、D. M. Mitnik, et al., Phys. Rev. **A62**, 062711 (2000).
[5] 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 年.
[6] Y. Itikawa, ADNDT **80**, 117 (2002)

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 x 10^{-17} cm^2 at about 80 eV. The cross sections for Xe^+ peak at a value of 25.6 x 10^{-17} 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 K <= T <= 10^7 K.

図 4　論文アブストラクトから抽出した原子分子の情報を web 画面上でマーキングした例