特定分野の学術論文をピンポイントで抽出し，いかに検索するか？

Defect dat@baseの実践例

A pinpoint search system for a specialized research area in physics and engineering

"Defect dat@base" for defects in semiconductors and semiconductor devices

梅田 享英[1] ｜ 萩原 茂[2] ｜ 水落 憲和[1] ｜ 磯谷 順一[1]

UMEDA Takahide[1]; HAGIWARA Shigeru[2]; MIZUOCHI Norikazu[1]; ISOYA Junichi[1]

1 筑波大学大学院図書館情報メディア研究科・知的コミュニティ基盤研究センター（〒305-8550　茨城県つくば市春日1-2）
Tel : 029-859-1307/ 029-859-1321/ 029-859-1594
E-mail : umeda@slis.tsukuba.ac.jp/mizuochi@slis.tsukuba.ac.jp/isoya@slis.tsukuba.ac.jp

2 筑波大学知的コミュニティ基盤研究センター・非常勤職員（〒305-8550　茨城県つくば市春日1-2）
E-mail : hgwrsgr@gmail.com

1 Research Center for Knowledge Communities, Graduate School of Library, Information and Media Studies, University of Tsukuba (1-2 Kasuga Tsukuba-shi, Ibaraki 305-8550)

2 Research Center for Knowledge Communities, University of Tsukuba (1-2 Kasuga Tsukuba-shi, Ibaraki 305-8550)

原稿受理　(2008-10-08)　　(情報管理　51(9):653-666)

1. はじめに

1.1 学術論文の大量生産時代

インターネットの世界では「情報爆発」という言葉をしばしば耳にする。Webやブログによって大勢の人が情報を発信できるようになり，そのために大量の情報がネットワーク上に無秩序に置かれるようになった。このような状況は，インターネットが取り扱う（専門性や厳密性といった意味で）ルーズな情報に限ったことではなく，極めて専門的な情報についても当てはまるようになってきている。

図1をご覧いただきたい。これは執筆者のグループ注)が主に投稿するアメリカ物理学会American Physical Society（APS，http://www.aps.org/），アメリカ物理学協会American Institute of Physics（AIP，http://www.aip.org/）の主要学術誌の年間ページ数の変遷をグラフにしたものである。今後もこれらの雑誌を分析に使用するので少し詳しく説明しておくと，APSが発行する学会誌がPhysical Review（PR，後にシリーズA～Eに細分化），その速報誌がPhysical Review Letters（PRL），AIPの学会誌がJournal of Applied Physics（JAP）で，その速報誌がApplied

注) 執筆者たちが所属する知的コミュニティ基盤研究センターは2002年に筑波大学に設置された。情報発信/蓄積の基盤となるさまざまな種類の知的コミュニティ（knowledge communities）に関する研究を行うのが目的である。執筆者のグループは物理学や工学を専門とし，情報通信基盤を支えるデバイスや材料の研究を主に行っているが，執筆者たちが所属する学術専門分野における知的コミュニティについても研究を行っている。

Physics Letters（APL）である。図1のグラフからは，1970年前後よりページ数が増加傾向を示し，2000年以降はその増加傾向に拍車がかかっていることが読み取れる。合計ページ数は4誌だけでも年間10万ページにも及び，人間の処理範囲を超えてしまっていることがうかがい知れる。

このページ数の増加は論文数の増加によるものである。もともと，速報誌であるPRLとAPLにはページ制限（1論文当たり3〜4ページ）があるため，図1の結果は論文数の増加にほかならない。ページ制限の無いレギュラー誌で見ても，論文1件当たりのページ数はPhysical Review B(PRB)で7.7ページ（70年）→7.5ページ（06年），JAPで4.8ページ（80年）→5.6ページ（06年）と各年代で大きくは変わっていない。このような論文数の急激な増加は，(1)論文執筆者数（つまり研究者数）の増加，(2)1人当たりの執筆論文数の増加，(3)高いインパクトファクターをもつ雑誌への投稿の集中，の3要因によってもたらされていると推測される。現役の研究者として肌身に感じていることだが，(1)については従来の欧米＋日本の研究者に加えて新興国の研究者が明らかに目立つようになってきており，(2)や(3)については，どこの国でも研究予算獲得のために(論文の数)×(掲載誌のインパクトファクター)が重要視されるようになっていることが効いている。このような増加傾向は，その原因から考えても今後ますます拍車がかかることが確実である。学術論文を発行する側の学会や出版社も，論文を読む側の研究者も今後ますます情報の処理に労力を払わなければならない。

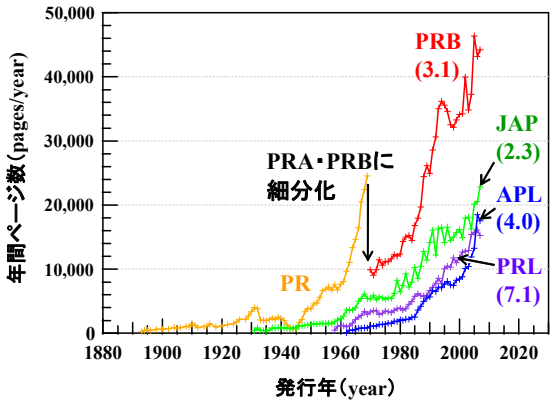

図1　アメリカ物理学会の学術雑誌の年間ページ数の増加

Physical Review（PR）及びPhysical Review B（PRB）：http://prb.aps.org/。Physical Review Letters（PRL）：http://prl.aps.org/。Journal of Applied Physics（JAP）：http://jap.aip.org/。Applied Physics Letters（APL）：http://apl.aip.org/。かっこ内の数字はThomson Scientific社による各雑誌のインパクトファクター（2006年の値）。ちなみに日本の同分野では，日本物理学会欧文誌Journal of The Physical Society of Japanが1.9,日本応用物理学会欧文誌Japanese Journal of Applied Physicsが1.2となっている。PRは現在，A〜Eの5分野に分かれており，その中で最大のPRBはCondensed matter and materials physicsをカバーしている。

1.2 論文大量生産時代への対応

このような論文の大量生産時代に対応するためには，まずは論文の電子化・データベース化が非常に大切である。上述の4誌はすでに発行初年にまでさかのぼって論文の電子化を完了しており，他の雑誌もそれに追随すべく電子化を進めている。そうして論文すべてをデータベース化し，検索エンジンで必要な論文を探し出す仕組みを作る。これはインターネットの世界と同じ解決方法である。APSやAIPはそれぞれのデータベースのクロス検索機能（SPIN＋Scitation®，http://scitation.aip.org/）を無料で公開しており，対象となる220万件の論文の検索が可能な体制を整えている。また，検索エンジンの改良も随時進んでいて，最近ではGoogleベースの検索エンジンCrossRef Search（http://prola.aps.org/xrs.html）が提供されるなど，さまざまな取り組みが行われている。

しかし現役の研究者から見て，それだけでは問題が解決しないような気がしている。インターネットの検索エンジンと同様に，データベースが巨大化するほど，検索をすると大量の論文が引っかかり，的の絞れた検索を行うのは非常に大変なのである。学術論文も，Webサーチと同様にトップに表示される幾つかの論文を見るだけでよいの

だろうか。また，新しい分野を始める時など，自分の専門から外れた分野の検索では検索結果の妥当性が判断できないことがよくある。本当に重要な論文が網羅されているのだろうか。そうした場合，現役の研究者はどのような行動をとるだろうか。それは（そう簡単に実行できる訳ではないが）その分野の専門家に直接聞いてみることである。そうすると例えば「こういう論文があります」「以前はこの論文の説が正しいと考えられていましたが，今はこちらの論文の方が定説です」といった耳寄りな情報を教わることができる。学術論文にも間違いや修正はあるし，また異論や対立もある。それが学会での議論を通して，ある説，ある理論に集約され，確定していく。専門家の頭の中には，そのような知識が整理されて収められている。この専門家の知識をデータベース化できれば，論文の検索にとって有効ではないだろうか。

そこで微力ながら，私たちが本業とする研究分野において，そのような環境の実現を試みることにした。それが本稿に述べるDefect dat@baseシステム（http://www.kc.tsukuba.ac.jp/div-media/defect/）である[1)]。このほかに，私たちは「電子スピン共鳴分光（ESR/EPR）」という特殊な専門分野における強力なデータベース兼シミュレーター「EPR in Semiconductors」（http://www.kc.tsukuba.ac.jp/div-media/epr/）も公開している[2)]。どちらにしても，私たちの視点はデータベースの開発者側というよりは，ユーザーの側に近い。私たち自身がデータベースのユーザーであり，またその中身（コンテンツ）の生産者でもあるからである。そのような少し毛色の違った立場から，ある特定の論文を学術雑誌から人間の専門家と同じように抽出する技術についての研究結果を述べる。また，そのようにして抽出された論文を極めて専門的な内容でピンポイントで検索するための私たちのシステムについて紹介したい。

2. Defect dat@baseシステムの実践例

2.1 題材について

Defect dat@baseは，私たちが研究している「半導体の結晶欠陥」を題材にした公開データベースで，ソーシャルブックマーク技術を応用した検索システムに特徴がある。システムの話をする前に，題材についてまずご紹介したい。

「半導体」とは，物理学や電子工学の教科書を見てみると「電気をよく流す金属と電気を全く流さない絶縁体のちょうど中間の電気抵抗をもつ物質」と書かれている。もっと正確に言えば，半導体は絶縁体の１種であり，基本的には電気を流さない。しかし，ある特殊な条件下（電圧がかかる，光が当たる等）では電気を流すことができる。この性質をうまく使うと電気と光の精密なコントロールが可能となり，これが現代の高度なエレクトロニクスの基盤技術となっている。しかし半導体の性質は，結晶の純度や結晶の完全性に極めて敏感に反応する。「結晶欠陥」とはそのような純度を左右する不純物や，結晶の規則正しい並びが乱れた部位を指し，半導体の研究では最も大切な要素の1つである。

半導体の結晶欠陥については専門の国際会議ICDS（International Conference on the Defects in Semiconductors）があり，隔年開催で2007年度で24回（48年間）を数えている。参加人数は300～500名。半導体研究分野をもっと広範にカバーする会議としてICPS（International Conference on Physics of Semiconductors）もあり，ICDSと交互に隔年開催されている。さらに半導体の種類ごとに専門の国際会議・シンポジウムがあるので，毎年，相当数の会議が世界各地で開かれていることになる。

半導体が注目を集めるようになったのは1947年のゲルマニウム（半導体の1つ）を用いたトランジ

スタの発明からで，以来60年を超す歴史の中でさまざまな半導体が誕生し，半導体産業はいまや世界で最も巨大な産業となった[3]。現在の主流であるシリコンはマイクロプロセッサやメモリといった大規模集積回路（LSI）を作ることのできる唯一の半導体であり，現在でも技術革新が続いている。しかし90年代以降は，シリコンのもつ物理的な性能限界が次々に明らかにされて，新しい半導体の開発が現実味をもって進められるようになった。そこでも，シリコンが過去（1970～80年代）にたどったように，結晶欠陥をいかにコントロールするかが技術の鍵となっており，半導体の結晶欠陥の研究には尽きることなくテーマが提供されている。この分野の学術論文の傾向としては，実験系の論文は1970～80年代の研究が今もって引用されている。他方，理論系の研究は，コンピュータの発達によって計算規模や精度が大きく変わってくることもあって，より最新の研究の方が引用される傾向にある。論文は物理学（solid state physics, condensed matter physics, materials physics）や応用物理学の雑誌に投稿されることが多い。以上が，ざっと見たDefect dat@baseがカバーする研究分野の傾向である。

2.2 「タグ」を使ったピンポイント検索システム

Defect dat@baseがやろうとしているのは，この分野の専門家の知識を集めて検索に利用できるような仕組みをネットワーク上に作ることである。現役の研究者は毎週のように自分と関連のありそうな学術論文を探して読んでいる。そうして研究者の頭の中には「この論文にはこのような内容が書いてある」といった専門知識が蓄えられていく。この情報をネットワーク上のデータベースに蓄積できないかと考えた。

そこで，Webページの分類に使われているソーシャルブックマーク（social bookmark）技術[4]を応用することを考えた。図2(a)にはDefect dat@baseのインターフェースが示されている。メイン画面には半導体結晶欠陥に関する重要な学術論文がリストアップされ，電子ジャーナル上の各論文のWebページ（抄録ページ，図3）へとリンクしている。このページから論文本体へのアクセスも可能である（ただし通常は有料）。各論文がどのような内容であるかを示すために，基本的な書誌情報と抄録の断片が表示され，さらに検索用の「タグ」がその論文を読んだ専門家の手によって与えられている。これらはWebページのソーシャルブックマークの本家サイトであるDelicious（http://delicious.com/）に倣って設計した（図2(b)）。ソーシャルブックマークがインターネットで使われ始めたのが2004年で，Deliciousもこの時期に生まれている[4]。私たちが開発を始めたのが2005年で，Defect dat@baseの公開は2006年7月，アメリカで開催された結晶欠陥に関するシンポジウムにおいて行われた[1]。そこから現在に至るまでの論文やタグの数の変遷を図4に示した。

タグは論文の内容を表す語句であり，その論文を読んだ専門家あるいは論文の著者自身によって与えられるが，キーワードのように必ずしも中心話題を表す語句ではなくてもよいところがポイントである。通常，論文には著者が選んだキーワードが5個前後与えられているが，このようなキーワードは大分類にしか使えない。それに対して，タグは専門的でピンポイントな検索に対して効果を発揮する。例えばDefect dat@baseの場合は，特定の結晶欠陥について検索するという場合が多いが，結晶欠陥の名前（Label）は機械的に名付けられたものが多く，例えば「A1～An」（適当なアルファベット＋整数n），「NL1～NLn」（NLはオランダの意），「G1～Gn」（Gは研究者のイニシャル）などがある。このような「記号」を普通の検索エンジ

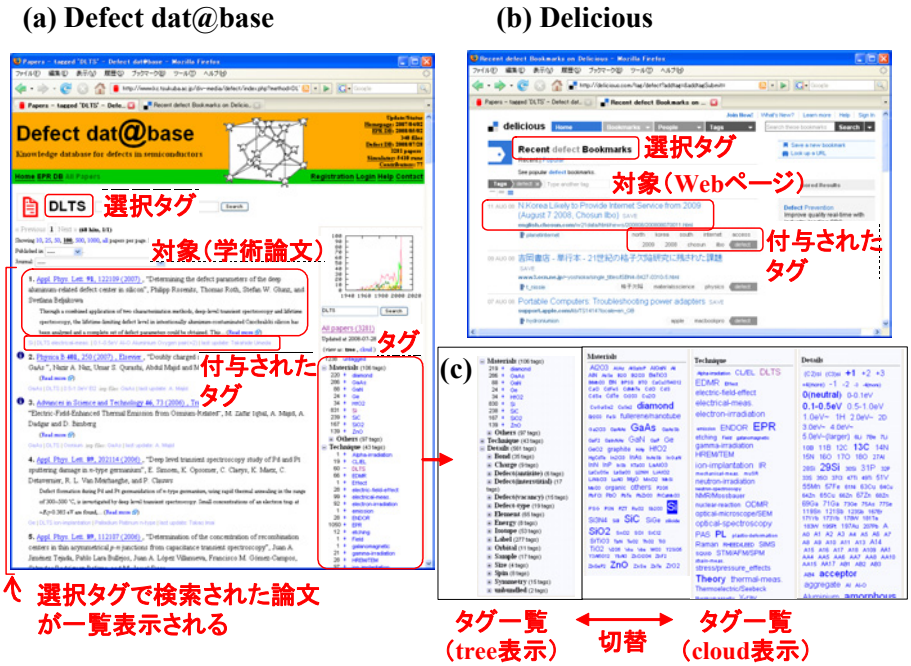

図2　Defect dat@baseとDeliciousのインターフェース

(a) Defect dat@baseのメイン画面（http://www.kc.tsukuba.ac.jp/div-media/defect/）。「半導体の結晶欠陥」に関する重要な学術論文や文献を，専門家の入力した「タグ」で分類する。(b) Deliciousのメイン画面（http://delicious.com/）。DeliciousはWebページのソーシャルブックマーク分類の先駆け的サイトである。2008年7月に大幅なリニューアルがあり，インターフェースがそれまでとは一変した。この画面は最新版のもの。(c) Defect dat@baseのタグ表示。Deliciousと違って，学術用途に合わせるために階層的なタグ管理を行っている。左側は階層を明示したtree表示，右側はDeliciousで採用されているcloud表示で，好みに応じて切り替える。ユーザーは検索に使いたいタグを見つけてクリックする。すると，そのタグが付与されたメイン画面に論文が表示される。tree表示の時は，各タグに対して論文数が表示される。

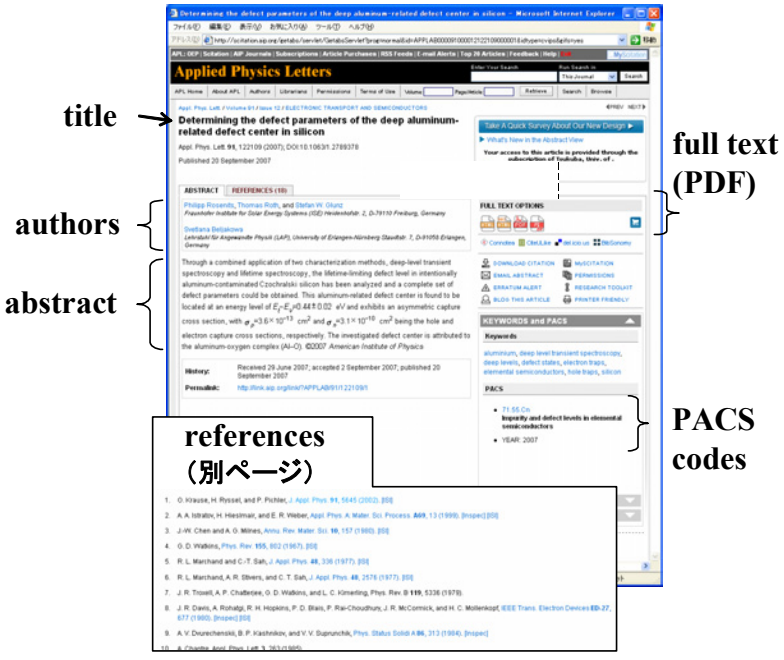

図3　電子ジャーナルの抄録ページ

AIPのApplied Physics Letters（APL）の例。無料で公開されており，DOI（Digital Object Identifier, http://dx.doi.org/）を使ってURLが与えられているためにリンク切れの心配もない。そこから先の論文本体（full text）へのアクセスは有料となる。したがって，Defect dat@baseがリンクするのは抄録ページまでとしている。抄録ページに載っている情報はどの雑誌もほぼ共通になってきたが，機能やレイアウトなどは雑誌によって大幅に異なる。古い論文についてはWebページがまだ整備されていない雑誌が多い。

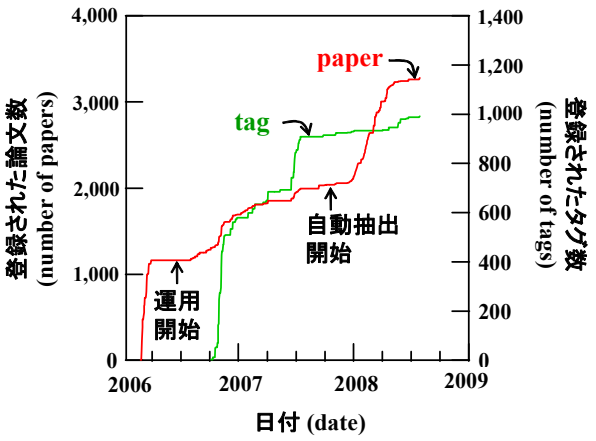

図4　Defect dat@baseにおける論文とタグの数の変遷

2006年7月より公開。2007年9月より自動論文抽出システムを試験的に稼動。システムの評価が終わるまで論文の登録を延期していたため，論文数の増加にはタイムラグが発生している。2008年5月に電子ジャーナルのWebページの大幅なリニューアルがあり，自動論文抽出システムを一時停止させた。右軸のタグ数は「登録されたタグの数」で1,000個近くに上っているが，登録後に統廃合されたタグがあるので現存しているのは約700個である。

ンで検索しても全然ピントの合わない結果しか得られないが，Defect dat@baseでは「Label」を表すタグが277種類登録されており（図2(c)），タグ「G1」を選べば，G1欠陥について書かれた論文が直ちに見つかる。

第2の例としては，結晶欠陥を構成する元素（例えば水素，H）で調べようとした時も，Defect dat@baseの「Element」を表すタグ（現在65種類）で「Hydrogen」を選択すれば，水素で構成された結晶欠陥を調べた論文が直ちに表示される。一方，検索エンジンで「Hydrogen」を調べた場合は，結晶欠陥とは関係のない「Hydrogen」の論文が大量に引っかかり，その中から選び出さなければいけなくなる。

第3の例として，タグを使えば数値検索もある程度可能になる。Defect dat@baseでは１つの試行として，エネルギー値を表すタグ「0～0.1eV」「0.1～0.5eV」… を作って，論文にこれらのタグを付与している。検索エンジンでは単位をもった数値を検索することは不可能であるが，タグを使用するとさまざまな対象について検索の可能性が開けてくる。もちろん，複数のタグを使っての絞り込み検索も可能であり，例えば「シリコン(Si)中の水素（hydrogen）と関連した欠陥を赤外吸収分光法（Infrared spectroscopy: IR）で調べた論文」を探したければ，「Si」「Hydrogen」「IR」というタグを選べばよい。さらに，通常の検索エンジンと同様の語句検索もサポートしている。

学術論文が対象のDefect dat@baseでは，本家のDeliciousとは違って，タグの階層的な管理を行っている。どの論文でも「研究する半導体（研究対象）」「研究に使用した手法」「研究結果の詳細」が記述されているはずなので，まず「Materials」「Technique」「Details」という3分野にタグを区分けした（図2(c)）。これを「バンドル」と呼んでいる。その下層にさらにバンドルを作ることも可能である。特に561種類ものタグのある「Details」では，前出の「Label」「Element」「Energy」などのバンドルによって大量のタグをうまく整理している（図2(c)）。

Defect dat@baseのもう1つの特徴は，各タグに使用頻度が表示される点である。Deliciousでは使用頻度の高いタグを目立つようにするcloud表示が使われている。Defect dat@baseでも同様の表示オプションが使えるが，標準はディレクトリ形式の表示を採用していて，その横に使用頻度が表示されようになっている（図2(c)）。これは，どの半導体がどのくらい研究されているのか，どのような手法がよく使われているのか，どの結晶欠陥がどの程度調べられているのかなどを知る目安となり，Defect dat@baseを使うユーザーにとって有益な情報を提供してくれる。

3. 特定分野の学術論文をピンポイントで抽出する

3.1 抽出アルゴリズムと，その評価

Defect dat@baseは，上述の編集機能を使って，まずは専門家の人に半導体結晶欠陥に関する欠かせない学術論文またはその他の文献（書籍）を登録してもらい，次にタグ付けを行ってもらうシステムである。後者の作業は極めて専門的な内容に対してピンポイントで行われるので人間（専門家）が行うことを前提としているが，前者については対象が広がる分，コンピュータでもある程度できるのではないかと考えた。そこで，電子ジャーナルから対象論文を専門家並みの精度で抽出してDefect dat@baseに登録する自動論文抽出システムを開発することにした。

このシステムの要（かなめ）は必要な論文をいかに見分けるかという判断作業にある。専門家は論文の内容・意味を理解して判断することができる。しかしコンピュータの場合，あるいは自然言語処理の立場では，論文にどのような語彙（ごい）が使われているか，それが文中のどこに登場したのかを解析することになる。いかにしてコンピュータに専門家と同じような判断をさせるか。そのために，専門家の判定とコンピュータアルゴリズムの解析結果とを大規模に比較検討する評価実験を行った。本節ではその研究内容について述べる。

語彙の解析には各論文のダイジェスト版とも言える抄録ページ（**図3**）を使用する。抄録ページはインターネット上にあって手に入りやすく，かつhtmlタグが埋め込まれているので，コンピュータで処理するには都合がよい。抄録ページには引用文献も載っていることがあるが，ことわりがなければ引用文献は除いて処理を行った。

判定のためのスコアは，ベイズ統計学を基に計算する。ベイズ統計学は既知の知識（学習）を基に確率を計算する手段を与えてくれる。計算過程は次のようなもので[5]，もともとは電子メールフィルタリング（迷惑メールフィルタリング）用に考えられたものである。

(1) 学習：抽出したい論文と抽出したくない論文（それぞれ正解論文，不正解論文と呼ぶことにする）を与えて，論文から語彙wを取り出し，その語彙が正解論文で登場する確率p(w)と，不正解論文で登場する確率q(w)を次式で近似計算する。

$$p(w) = g(w) \div [g(w) + b(w)] \quad \cdots\cdots(1)$$

$$q(w) = b(w) \div [g(w) + b(w)] \quad \cdots\cdots(2)$$

ここでg(w)は(wを含む正解論文の数)÷(正解論文の数), b(w)は(wを含む不正解論文の数)÷(不正解論文の数)である。ただし，登場頻度がまれな語彙に対しては，この手の確率計算は甚だ不正確になってしまう。例えばwが正解論文にのみ登場する（b(w)=0となる）語彙だった場合，その登場回数が100回でも，たった1回であっても等しくp(w)=1，q(w)=0となるが，1回のデータだけで確率100%と計算するのは理論的に間違っている。そこで，wが登場する論文の数をnとして，次式で補正を行う[5]。

$$f(p(w)) = [\, s/2 + n\, p(w)\,] \div (s + n) \quad \cdots\cdots(3)$$

sは重みパラメータで0ならば補正なしで，sを大きくするほど補正が入る。私たちはs=1とおいた。すると上述の例ではn = 1で，f(p(w)) = 0.75と補正される。

(2) スコア計算：判定したい論文を与えて，同じように語彙に分解し，すべての語彙wについて確率f(p(w))を次式で結合して，この論文が正解論文である確率Pを計算する。

$$P = \chi^{-1}(-2 \ln \Pi_W f(p(w)), 2N) \quad \cdots\cdots(4)$$

χ^{-1} はカイ二乗関数の逆関数であり，Nはwの総数である。式(4)の計算にあたって，もしもwが

学習されていなかったら（p(w)が不明だったら），中立的な立場をとってf = 0.5とする。同様に確率q(w)についても同様の方法で結合確率を計算して，この論文が不正解である確率Qを計算する。最終的に，Pと1 − Qの平均でスコアSを計算する。

$$S = (P + 1 - Q)/2 \quad \cdots\cdots\cdots(5)$$

Sには正解論文と不正解論文の両方の情報が入っており，正解論文と不正解論文の情報量が不均等であってもスコアの計算が適切にできるようになっている。

(3) 判定：計算したスコアSが，ある閾値（しきいち）を超えれば正解論文として抽出する。

以上のアルゴリズムを実装した後，評価実験を行った。評価のために用意した学術論文は合計16,394件で（**表1**），半導体の研究論文が掲載される物理学・工学分野の110誌から選出した。これを専門家（今回は執筆者の梅田准教授）が正解論文と不正解論文に選（え）り分けた。この論文集合を雑誌や発行年になるべく偏りが生じないように2つに分割し，片方の8,196件の集合で語彙の学習を行った。次にもう片方の8,198件の集合で正解論文の抽出実験を行い，専門家の抽出結果との比較を行った。

さらに，学術論文が他の一般文書とどのように違うのかを探るために，合計77,150件の電子メールの集合に対しても同様の学習・抽出実験を行った。この集合は，インターネット上で公開されている研究用データ集合2005 TREC Public Spam Corpus[6]から英文メールだけを取り出したもので，スパムメールが42,188件，そうでないメールが34,962件含まれている。これを2分割して片方で学習を行い，もう片方で抽出実験を行った。一般的に，電子メールフィルタリングは高い性能を発揮することが知られている。私たちのアルゴリズム実装が正しければ，高性能なフィルタリング機能が電子メールに対する実験で見られるはずである。

3.2 学術論文と一般文書の違い

まず学術論文と電子メール文書がどのように違うのかの比較を**図5**に示した。コンピュータによる抽出アルゴリズムがどのくらい高性能かを測るために，一般的に用いられる再現率（recall）と精度（accuracy）の2指標で比較を行った（定義については**図5**参照）。横軸は正解を判定する閾値であり，これを高くすればするほど「厳密な抽出」をコンピュータに課することになる。再現率，精度が同時に100%（＝人間と同じ結果）に近づくほど高性能と言える。図から分かることは，このアルゴリズムは電子メール文書に対しては非常に効果的に働くということである。閾値を60%に設定すれば再現率，精度ともに99%以上という値が容易に得られる。これは私たちのアルゴリズム実装が正しいことの証明にもなっている。一方で，学術論文の場合は，再現率と精度の間に非常に強い逆相関が見られ，閾値をどのように設定しても再現率と

表1　抽出実験に使った学術論文と電子メール文書の集合

それぞれの集合に含まれる語彙の数を，重複を除いた場合（種類）と総数（語）で示した。1件あたりに含まれる語彙の総数はどの集合もほぼ共通である。文書はすべて英文である。

	集合	正解	不正解
学術論文（英文）16,394件	学習用	889	7,307
	評価用	890	7,308
	合計	1,779件 22,061種類 518,750語 292語/件	14,615件 124,982種類 4,123,892語 282語/件
電子メール（英文）77,150件	学習用	17,481	21,094
	評価用	17,481	21,094
	合計	34,962件 330,260種類 8,852,813語 253語/件	42,188件 388,590種類 9,964,045語 236語/件

精度の両方を同時に100%近くまで上げることは困難であることが分かった。このように学術論文の抽出は，電子メールなどの一般文書に比べて技術的に難しいことが分かる。その原因は，学術論文はどれも体裁が似通っており，また使用される語彙も似通っていることが挙げられる。

そのことを示す例として，抽出性能の学習量依存性がある。学習量がどのくらいあれば適正な抽出ができるのかを調べるために，学習に使ったデータ集合を分割して，学習量を段階的に増やして評価実験を行ってみた（図6）。すると，電子メール文書の場合は800件程度の学習で早くも最終的な性能が得られたのに対し，学術論文の場合は4,000件程度の学習を必要とした。これは多くの学習を行わないと正解論文と不正解論文の語彙の違いが明確にならない，つまり語彙が似通っていることを示唆している。

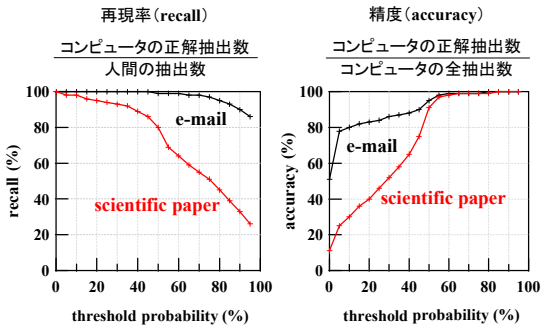

図5　学術論文の抽出実験結果（電子メール文書と比較して）

正解論文と不正解論文が混ざった論文集合（**表1**）からコンピュータに正解論文を抽出させる。その結果を専門家による判断と比較した。同様の実験を一般文書（電子メール文書）に対しても行った。横軸は，コンピュータが「正解」と判定するための確率閾値を表し，(5)式で計算したスコアSがこの値を超えれば正解と判定する。

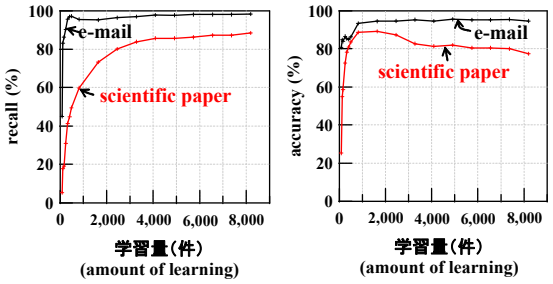

図6　抽出アルゴリズムの学習量依存性

学習量を段階的に増やして抽出実験を行った。左が再現率，右が精度。判定閾値は50%で行った。

3.3 語彙の抽出方法の検討

したがって，学術論文の抽出性能を上げようと思えば，もっと技術的な工夫が必要である。1つには語彙の抽出方法をもっと高度にすることが考えられる。そこで本研究では以下のような4つの方法を比較検討してみた。(i)が最も単純な方法で，(ii)はそのオプション的処理，(iii)と(iv)は形態素解析を使った処理と言える。本稿に述べられている結果は，特にことわりが無ければ，(iii)の抽出方法で得られたものとなっている。

(i) すべてのテキストを小文字に正規化し，さらにhtmlタグや制御コードも空白に置き換えて，空白区切りで語彙を切り出す。語彙は1ワード単位の小文字となる。

(ii) 学術論文では大文字が略語として重要な意味をもつことがよくあるので（例えば Electron Paramagnetic Resonance を略してEPRとするなど），(i)において大文字・小文字を区別するように変更する。

(iii) (i)に加えて，Defect dat@baseに蓄えられているタグを語彙抽出に活用する。タグは正解論文に含まれている典型的な語彙を表現しているので，正解論文からの語彙抽出を正確にする効果が期待できる。Defect dat@baseには，約700のタグに対して平均7個の同義語を登録した約5,000項目からなる専用の同義語辞書がある。例えば，「EPR」というタグに対してはelectron paramagnetic resonanceやepr spectrumといった同義語が登録されている。論文の中に同義語辞書に合う語句があればそのまま語彙として抽出する。

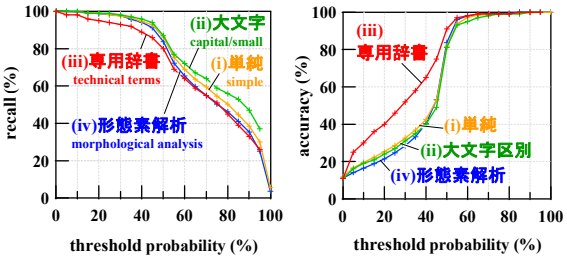

図7　語彙抽出方法を変えて論文抽出実験を行った結果

(i)～(iv)の４つの語彙抽出方法を比較。(i)が最も単純なワード単位の語彙抽出法で，(iv)は形態素解析によって名詞と名詞句を語彙として抽出した場合であるが，結果はほとんど変わらなかった。(ii)は大文字を区別する方法，(iii)は専用辞書を使った抽出方法であり，特に(iii)は大きな効果を上げた。

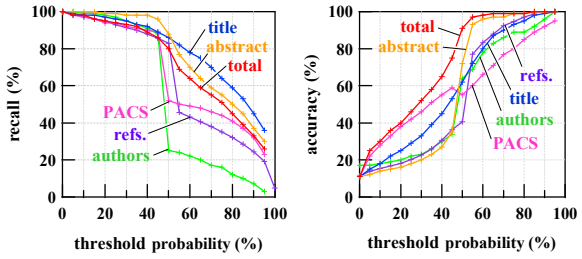

図8　学術論文のパーツごとに論文抽出実験を行った結果

学術論文のタイトル(title)，著者リスト(authors)，抄録(abstract)，PACSコード，引用文献（references）ごとに抽出結果を調べた。totalはreferencesを除くパーツすべてを使って抽出した結果で，精度は最も高い。それぞれのパーツから抽出された語彙数（重複を除く）は，正解論文1,779件/不正解論文14,615件から，title＝3,158/17,799，authors=3,638/49,584，abstract=17,257/69,226，PACS=595/2,470，references=12,303/113,519であった。

(iv) 自然言語処理で一般的に用いられる形態素解析を使って，名詞と名詞句のみを語彙として抽出する。例えば文中にelectron paramagnetic resonanceといった語句があれば，1つの名詞句として抽出し，さらにelectron，resonanceといった名詞も抽出する。大文字の区別はしない。形態素解析にはスタンフォード大学で開発されたパッケージ[7]を用いた。これは113,195語の英単語辞書を使用している。

図7は，以上の4つの方法による抽出結果を比較したものである。図を見て分かる通り，語彙の抽出方法は結果に大きな影響を与えなかった。唯一の例外が(iii)で，精度向上に対して大きな効果が見られた。語数としては約5,000語と少なくても，抽出したい分野の専門用語を定義しておくことがとても重要だということが分かる。しかし，この場合でも精度と再現率の間の強い逆相関の関係は克服できてはいない。大文字を区別する方法には再現率が若干向上するという効果が見られた。3.5節で述べるようにDefect dat@baseでは精度を最も重視しているので，(iii)の抽出法を標準的に採用している。

3.4 学術論文のパーツごとの分析

次に考えられる高度な処理は，語彙の位置に着目して解析を行うことである。学術論文の抄録ページ（図3）は，その中に埋め込まれたhtmlタグを分析することで，タイトル，著者，抄録，キーワード，引用文献などのパーツに分解することが可能である。これらのパーツごとに評価実験を行い，どのパーツがどのように抽出に役に立っているのかを評価した。図8が各パーツごとに調べた再現率と精度である。前節の語彙抽出法が似たり寄ったりだったのとは対照的に，パーツ間ではより大きな差異が現れた。

最も極端なのは著者名（authors）のパーツであり，再現率と精度が最も極端に逆相関した。研究者はそう頻繁に研究分野を変えることがないので，登場する研究者名によって論文の分野を精度よく判定できるのではないかと期待をしていたのだが，実際には効果的ではなかった。それは抽出された著者が多く（正解論文から3,638人，不正解論文から49,584人，重複を除く），しかもその約80%が1回のみしか登場しない著者だったためである。

最も効果的だったのは抄録のパーツで，これは論文の内容が最も説明的に書かれていることを反

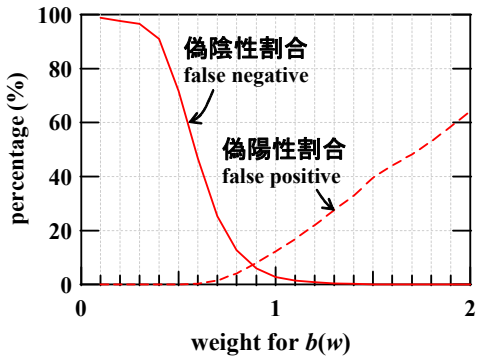

図9　論文抽出アルゴリズムにウェイトを与えて性能を調節する

不正解論文の語彙に与えるウェイトを変えて抽出性能の変化を調べた。ウェイトが1以上では不正解論文の語彙を重視することになり，逆に1未満の場合は正解論文の語彙を重視することになる。偽陽性割合は（正解論文を不正解と誤判定した数）÷（正解論文数）で，正解論文の取りこぼしを表す。偽陰性割合は（不正解論文を正解と誤判定した数）÷（不正解論文数）で，不正解論文の過抽出を表す。2つの指標は明らかに逆相関しており，両方を同時にゼロにすることはできない。精度100%を達成する（偽陰性割合を0%にする）ためには，ウェイトを1よりも大きくする必要がある。

映したものと考えられる。他のパーツと比べると，精度の向上に大きな効果が見られた。再現率に対しても同様に効果的で，最も優秀なパーツであると言える。タイトルのパーツは再現率に対してさらに効果的で，再現率を重視したい時は最も有効なパーツであることが分かる。

意外なことに，APSやAIPが採用し，チューニングを施しているPhysics and Astronomy Classification Scheme（PACS）の性能はかなり低く評価された。PACSは5階層にわたって細かく研究対象を特定する仕組みであり，著者自身の手で1〜4個のコードが論文に付与される。したがって，その論文の内容を正確に表現しているはずなのであるが，再現率，精度ともに他のパーツに比べて振るわなかった。その原因は分野がピンポイントであってもPACSは分散する可能性があるためである。実際に今回の実験でPACSがいくつ抽出されたのかを見てみると，14,615の不正解論文から2,470種類，1,779の正解論文から595種類が抽出されている（重複を除く）。論文数の差8.2倍に対し，PACS数の差は4.1倍にまで縮まっており，正解論文のPACSが狭い分野にもかかわらず多岐にわたっていることがうかがえる。

同様に，引用文献（references）もあまり振るわなかった。関連のある論文間では引用文献もよく似るのではないかと期待をしていたのだが，予想以上に引用文献がばらけていて効果が現れなかった。それを端的に表すデータとして，引用文献を取り出せた正解論文のうち38%はDefect dat@abase内の論文を「1つも」引用していなかった。

学術論文には上述のパーツ以外に，本文という最大のパーツが存在する。しかし抄録ページと違って，本文は有料ダウンロードが通常であること，大学図書館のように閲覧可能な場所であってもプログラムによる大量ダウンロードは禁じられていること，処理量が大幅に増えること等の理由から，本文を使った抽出アルゴリズムは現実的ではないと判断している。

3.5 自動論文抽出システムの稼動

以上の評価実験の結果を踏まえて，Defect dat@baseでは2007年9月より自動論文抽出システムを稼動させた（図4参照）。当面の対象は図1にも登場した主要4誌で，APSやAIPが提供しているE-mail Alert Service（最新刊の目次を無料で電子メール送信してくれるサービス）からのメール受信をトリガーとして，最新論文の抄録ページをスキャンし，正解論文と判定できれば自動的にDefect dat@baseに登録を行う。

この自動登録では精度を最も重要視している。間違った論文の登録は人間による余計な修正作業を発生させるうえ，データベースの価値を下げるからである。抽出精度を上げるには図5，7，8で見てきたように，閾値を上げて「厳しい判定」をするやり方がある。しかし，精度を100%とするためには閾値を80%よりも上に設定しなければならず，再現率は著しく犠牲になる。もっとよい方法

として，不正解論文の語彙にウェイトをかけるやり方があり，こちらの方が性能の調節の範囲が広い。ウェイトをどこに与えるかは議論の余地があるが，今回は(1)・(2)式のb(w)にウェイトをかけた。図9がウェイトを導入した時の性能の変化を調べたグラフである。b(w)に1.5倍のウェイトを与えると，判定閾値50%でも精度100%を達成することができた。実際には万全を期して1.8倍のウェイト（予想される性能は精度100%，再現率40～50%）を使用してシステムを稼動させ，評価を行った。2008年2月までの半年間の稼動で計62,846件の論文をスキャンし，そのうち561件（0.9%）を正解論文として抽出した。専門家（梅田）が確認したところ，精度は予想通り100%であった。

Defect dat@baseはRSS（RDF Site Summary）をサポートしているので，RSSリーダーがあれば半導体の結晶欠陥に関する最新の論文を無料で自動チェックできるようになる。これはとても便利な機能で，e-コマース分野で使われる推薦（recommendation）サービスの学術版と言える。同様のサービスはすでに一部の学術雑誌で有料で始まっている。例えばAIPが運営に関わるScitation®（科学技術に関する情報全般の電子化サービスを担う組織）では，あらかじめ設定された約200のトピックスに合致する最新論文を電子メールで知らせてくれるScitation Research Alertが数年前から始まっている（http://www.scitationalerts.org/）。情報量としては最大50件/週である。このようなサービスは学術情報が爆発的に増加する時代ではとても重要になるサービスだろう。今回の私たちの評価結果は，学術雑誌から論文を抽出する技術に関する貴重な実践的基礎データになるのではないだろうか。

4. 問題点と今後の方向性

Defect dat@baseの運用を開始して2年半が経過したが，当然のことながらさまざまな問題が発生している。今後の方向性を交えて最も大きな課題を簡単に述べて本稿を終わりたい。

最大の課題は，いかにDefect dat@baseを多くの研究者に使ってもらうかという点であろう。継続的に使ってもらうためには，まずデータベースの中身を充実させなければならないし，インターフェースも重要である。中身を充実させるには専門家によるタグ付けが一番重要で，現在はその多くを執筆者たちのグループで行っている。また，データベースの作業だけでなく，執筆者たち自身が本業（半導体の研究）で活躍することもDefect dat@baseの運用にとって非常に重要な要素である。幸い，この分野は過去の論文が引用されやすい分野なので，過去の論文に付けられたタグは貴重な財産として残る。まだ開発途上のため，大した宣伝ができていないのにもかかわらず，海外の研究者が使ってくれている現状もあるので，こつこつと育てていきたいと考えている。

第2の課題は，タグの一貫性維持の問題である。タグは放っておくとどんどん増え続け，同義語が多数発生してしまう。この問題はどのソーシャルブックマーク・システムにも共通する問題で，多くのシステムは数千から数十万種類のタグを有している[4]。このようにタグの数が増えると管理することはほとんど不可能になり，Defect dat@baseのように「バンドル」するのも非常に困難となる。また別の大きな問題として，新たにタグが作り出された場合，それよりも過去に登録された論文にはそのタグが使用されていないという一貫性矛盾の問題も発生してしまう。Defect dat@baseではタグの統廃合を執筆者のグループで管理しており，これまでに約300のタグを統廃合して（図4参照），タグの数をむやみに増やさないように試みている。しかし，一貫性維持のような煩雑な作業にはコンピュータによる補助が欠かせないと感じている。

タグ付け支援システム（タグの候補を推薦したり，確度の高いタグについてはコンピュータで付けたりする。そのためにタグの同義語辞書を作成した。3.3節参照）や，タグの一貫性を記述する言語の開発，それによってタグ管理をプログラムすることのできる環境の開発と評価に取り掛かっているところである。

参考文献

1) Umeda, T; Hagiwara, S; Mizuochi, N; Isoya, J. “Open web-based databases for defects semiconductors”. The 2006 Gordon Research Conference on Defects in Semiconductors. New London, USA, 2006-07-2/7.

2) Umeda, T; Hagiwara, S; Mizuochi, N; Isoya, J. “Development of web-based database system for EPR centers in semiconductors”. The 23rd International Conference on Defects in Semiconductors (ICDS-23). Awaji-island, Japan, 2005-07-24/29. A web-based database system for EPR centers in Semiconductors. Physica B. 2006, vol.376-377, p.249-252.

3) Sze, S. M; “Nanoelectronic technology: Challenges in the 21st century”. International Conference on Solid State Devices and Materials (SSDM). Tsukuba, Japan, 2008-09-23/26.

4) Hammond, T; Hannay, T; Lund, B; Scott, J. Socail bookmarking tools (I): A general overview. D-Lib Magazine. 2005, vol.11, no.4.

5) Robinson, G. A statistical approach to the spam problem. LINUX journal. 2003, http://www.linuxjournal.com/article/6467/, (accessed 2008-10-01).

6) Cormack, G. V.; Lynam, T. R. “2005TREC Public Spam Corpus”. http://plg.uwaterloo.ca/~gvcormac/treccorpus/, (accessed 2008-10-01).

7) “Stanford Parser version 1.6”. The Stanford Natural Language Processing Group. http://nlp.stanford.edu/downloads/lex-parser.shtml, (accessed 2008-10-01).

著者抄録

インターネットの世界だけでなく，科学技術の世界でも情報量（学術論文）の急激な増加が問題となっている。そのような論文大量生産時代には，大量の論文の中から特定の論文をピンポイントで抽出したり，検索したりする技術が重要になってくる。本稿では，ソーシャルブックマーク技術を応用して，物理学・工学領域の中の「半導体の結晶欠陥」に関する重要な学術論文をピンポイントで検索するデータベースシステムDefect dat@baseについて紹介する。また，このデータベースに該当する重要な論文を専門家と同じ精度で学術雑誌から自動的に抽出するために，人間（専門家）とコンピュータの抽出アルゴリズムとの間で，約16,000件の学術論文に対する大規模かつ詳細な抽出比較実験を行い，さらに一般文書との違いについても比較検討した。その研究結果について詳しく述べる。

キーワード

学術論文, データベース, 理工学, タグ, ソーシャルブックマーク, ピンポイント検索, 論文抽出, 電子メールフィルタリング, 語彙統計解析

Author Abstract

Similar to the Internet, scientific and technological communications are rapidly accelerating in the 21st century. Among many papers in scientific journals, researchers must find out proper ones by using search engines of journal databases as well as other techniques that enables "pinpoint" search. We developed such a pinpoint search system based on the social bookmark technology, entitled "Defect dat@base" (http://www.kc.tsukuba.ac.jp/div-media/defect/). This database covers the specialized research area of "Defects in Semiconductors and Semiconductor Devices", and is collecting and precisely classifying important papers in this area in cooperation with specialist members. To extend and maintain the collection by not only human specialists but also computers, we studied statistical and morphological algorithms. As a result, we could learn how to choose important papers as accurate as human specialists do. Using over 16,000 papers of physics and engineering and over 77,000 e-mail texts, we carried out a large-scale comparative studies about differences between human and computer, and reached the following conclusion: Scientific papers are not as easily selectable as we can do for other types of texts, and for the better selection, we should focus on technical terms of the relevant area as well as abstracts and title words of the relevant papers.

Key words

scientific paper, database, science and technology, tag, social bookmark, pinpoint search, automated selection, e-mail filtering, statistical and morphological analysis