様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年　６月１１日現在

機関番号：３２７１４
研究種目：若手研究（Ｂ）
研究期間：2009～2011
課題番号：２１７００１２４
研究課題名（和文）　物質・材料に特化したウェブ検索システム構築に関する研究

研究課題名（英文）　Study on web search system specialized in materials

研究代表者
大塚　真吾　（OTSUKA SHINGO）
神奈川工科大学・情報学部・准教授
研究者番号：７０５０９７３６

研究成果の概要（和文）：物質・材料の検索は一般的に化学式を利用するため，従来のウェブ検索システムでは，ユーザが望む解を得ることが難しい状況である．そこで，本研究ではこれらの問題点を解決すべく，物質・材料に特化したウェブ検索システムの構築を行った．さらに，構築したシステムのアクセス履歴から利用状況の把握を行った．

研究成果の概要（英文）：It is difficult to obtain the good results about materials using conventional web search systems because these systems use a chemical formula generally. Therefore, we establish the system specialized in materials in order to solve these problems. Moreover, we grasp the usage with the access logs of our system.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	2,000,000	600,000	2,600,000
2010 年度	800,000	240,000	1,040,000
2011 年度	600,000	180,000	780,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
科研費の分科・細目：情報学・メディア情報学・データベース
キーワード：情報検索，ウェブ情報検索，物質検索，材料検索

１．研究開始当初の背景

物質・材料に関する研究分野においては物質・材料の発見から実用化までに長期間かかることも多い．例えば，炭素繊維は 1960 年代に発明された技術であるが，2006 年になりこれを旅客機のエンジンの一部分に利用すべく研究が開始されている．このような例では研究過程で数十年前の文献やデータが必要になることが予想される．IT 化社会以前に書かれた論文や実験データの大部分は紙媒体のまま図書館や研究者個人で管理されていたが，現在では NII による最先端学術情報基盤の整備により，学術論文の一般公開や紙媒体を電子化すべく様々な取り組みが行われている．

一方，物質・材料に関する研究成果はインターネット上で広く公開されており，研究者が Google や Yahoo!のような，既存の検索エ

ンジンを利用することで物質・材料に関する情報の検索を行うことができる．しかし，例えば検索語「nano」の結果の上位には iPod に関連するページが多くあり，研究者が目的とする物質・材料分野に関連のある nano テクノロジーに関するページを探し出すことが難しく，研究者は最適な検索語を思いつくために労力を費やしている．同様にウラン，プルトニウムのような原子力関連の物質やイリジウムやラドンのように携帯サービスや温泉の成分として一般に広く知られている元素に関してもその検索は容易ではない．

また，物質・材料検索では「CO_2」のような化学式を用いた検索ニーズも高いが，現状の検索エンジンにおいては同一原子の個数を示す数字部分のワイルドカード検索は不可能である．ポリマーなどの研究分野では，$CH_3CO[OCH_2CH_2]_nOCOCH_3$ のように反復の個数を n や m などを用いてワイルドカード的に示す場合があり，化学式を用いたウェブ検索の問題をより複雑にしている．

さらに，既存の検索エンジンでは検索語をダブルクォーテーションで囲むことで，例えば”P_2O_5” のように指定した化学式そのものを含むページを検索することが可能である．しかし化学式中の元素記号の順番は研究者によって異なる物質も存在するため，例えば化学式「○△$_2$□×$_3$」は研究者によっては「□×$_3$○△$_2$」と表現されることがある．このため，既存の検索エンジンではウェブ空間から特定の物質や材料を含むページを検索することは難しい．

国内における化学式での検索サービスとして NIST が提供する Chemistry Web Book がある．このサービスでは「C_4H_*CL」のように原子の個数部分をワイルドカードにすることで，それに該当する物質が化学式の一覧と共に実験データなどの情報を得ることができる．しかし，ウェブ空間に対する検索はサービスの対象となっていない．一方，オランダの Elsevier 社が提供する科学情報専用のインターネット検索エンジンである SCIRUS は，多くの科学関連 Web ページと論文情報に対して包括的な検索が可能だが，前述した元素記号の入替えには対応していない．また，このシステムで使われている技術は一般に公開されていない．

２．研究の目的

物質・材料に関連する検索システムを構築しウェブ上に公開することで，この分野の研究者に材料に関する情報を提供する．また，このシステムを利用したユーザのアクセス状況を解析することで，利用者のニーズを把握する．

ナノテクノロジーなど物質・材料が関わる研究は日本において重要な研究課題であり，今後さらなる飛躍が望まれている．この分野の研究者に対して IT 技術を用いて支援を行うことは非常に意味深いことだと考えている．また，本研究成果により，個々の研究者が国内外の物質・材料に関する情報をいち早く収集することができれば，サーベイに費やす期間を大幅に短縮できるため，研究効率の向上に貢献できる．

３．研究の方法

研究目的を達成するためには，以下の問題を解決する必要がある．

(1) ウランやラドンといった一般に広く知られている元素についての検索が難しい．
(2) 同一原子の個数を示す数字部分をワイルドカードにした場合の検索が難しい．
(3) 化学式中の元素記号の順番は研究により異なるため，複雑な化学式に関する検索が難しい．

本研究ではこれらの問題を解決することを目標に研究を進め，得られた成果や知見をもとに物質・材料に特化したウェブ検索システムを構築する．

上記(1)を解決するための取組みとして，物質名や材料名のみを検索語として入力したユーザ（研究者）に対しては，検索結果をより専門的な情報にするための絞込み語や関連性の高い専門的な語の提示を行う連想検索の検討を行う．また，上記(2)，(3)に対しては XML を用いてウェブページなどに含まれる化学式を構造的に格納する手法を用いることで化学式検索の精度向上を目指す．

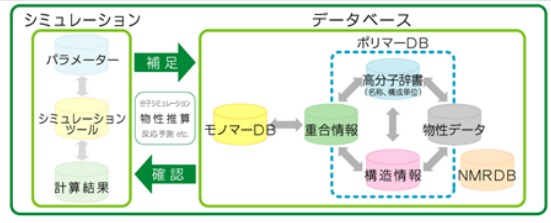

図１：PoLyInfo システム構成

４．研究成果

本研究課題の期間中に，材料の検索が可能なシステムについて検討を行い，現在，(独)物質・材料研究機構において稼働中の高分子データベース(PoLyInfo)への応用を試みた．また，PoLyInfoを含む物質・材料データベースに関するポータルサイトであるMatNaviのアクセスログを解析し，システムを利用したユーザ（主に材料研究者）の利用傾向について考察を行った．本報告書では，代表的な研究成果である，高分子検索システムの構成，および，利用状況の考察について述べる．

(1) 高分子データベース(PoLyInfo)のシステム構成

高分子データベース(PoLyInfo)は図1に示すように，シミュレーション部分(物性推算)とデータベース部分から構成されている．シミュレーション部分ではデータベースに格納された実データを元にまだ登録されていないデータや仮想的な材料の物性推算を行うアプリケーションである．

物性推算については，ポリマーの構成繰り返し単位(CRU)の化学構造から物性を推算する手法の1つに原子団寄与法がある．PoLyInfoではその中のVan Krevelenの物性推算を用いており，予測対象となる物性をいくつかの因子の関係式として表し，それぞれの因子を構成繰り返し単位(CRU)中に含まれる原子団から の寄与(原子団パラメーター)の和として計算を行う．関係式や原子団パラメーターは実測データの解析により求められることから，化学構造と物性値を関連づけたデータベースが効果的に利用できる手法である．

データベース部分については，PostgreSQLを用いてシステム構築を行っている．データモデルを内部的には保持しながら，一般のユーザがこれを意識する必要がないよう工夫した結果，高分子が保持する情報を物性データ，構造情報，重合情報の3つに分け，重合情報については高分子を構成するもとになるモノマーデータベースと連携をとる設計とした．また，各情報における名称や構成単位などについては，高分子辞書を用いることで同じ組成式で構造の異なる高分子を違うものとして扱うことが可能となる．

我々はPoLyInfoに登録する高分子をポリマーサンプルと呼んでおり，1つのポリマーサンプルを効率良く表現するために階層的なデータ構造を採用している．高分子辞書に収録されている一次構造情報，成形方法などの各ポリマーサンプルに関する情報は，サンプルごとに対応づけてられている．また，重合情報は実際の重合に関する詳細な情報が収録されている．

(2) 高分子情報の分析

高分子の情報として代表的なものに，ポリマー名称(IUPAC準拠の構造基礎名，原料基礎

名，慣用名，略称など)，ポリマー種別(ポリマーにはホモポリマー，コポリマー，ポリマーブレンドなど)，ポリマー物性(高分子の熱的物性，電気的物性，機械的物性など)などがあげられる．

高分子は一般的に同一の名称でも，立体規則性や重合度などの差異，成形方法，測定法によって物性値は異なる．そのため，ある高分子の試料(実験サンプル)が持つ特性を網羅的に扱うためには，モノマーからポリマー鎖，ポリマー集合体を経て高分子材料にいたるまでの化合物や物質に関する情報を全て考慮しなければならない．

モノマーの情報としては名称，分子式，構造図などがある．ポリマーに関する情報としては，構成単位(名称，化学構造，分子量など)，一次構造(立体規則性，分子量など)などの情報がある．また，モノマーとポリマー鎖を結ぶ情報として重合情報があり，ポリマー集合体としては，高次構造(成形加工情報，結晶構造など)，物性などの情報がある．さらに，高分子材料のもつ特徴として名称(材料名，商品名など)などの情報がある．PoLyInfoでは，これら全ての情報をデータモデル構築の際の分析対象としている．

PoLyInfoに登録されているデータの情報は，Chemical Abstracts Service(CAS)に登録された文献の中から化学構造が明確で各種物性の実測値を有する文献としている．データは原則として文献中の記述から実測値であることが判断され，かつ測定方法，測定条件の記載があるものについて採択している．実測値を元に計算により求めた値(誘導値)も対象とし，単なる引用のみの値は採択対象外としている．CASの検索から高分子の専門家が手作業で年間約2800文献からスクリーニングを行い，年間約700文献からポリマーサンプルを抽出し，その物性に関する情報をデータシートに書込み，これをデータベース(PostgreSQL)に登録している．

(3) PoLyInfoを含む物質・材料データベースに関するポータルサイト(MatNavi)のアクセス動向

MatNaviにある各データベースを利用するためには，ユーザ登録が必要となる．ユーザ登録を1度行うことで，PoLyInfo以外にも結晶基礎データベースや拡散データベースなど全てのデータベースを利用することができる．2012年5月時点の登録ユーザ数は67,499人であり，毎月100万件を超えるアクセスがある．また，登録ユーザは世界141ケ国，18,121機関におよび，材料分野のデータベースとして世界でも有数のユーザ登録数を誇る．

(4) 詳細なアクセス動向の調査

MatNaviのアクセス状況をより詳細に分析するため，まず，アクセスログに対して，以下に示すクリーニング処理を行った．

① 画像ファイルとcssファイルへのアクセス情報を削除
② Webクローラによるアクセスの削除
③ エラーページに対するアクセスの削除

次に，アクセスのIPアドレスを元に約110万セッションを抽出し解析を行った．全てのデータベース，および，MatNaviについて，月ごとのアクセス数の推移を図2に示す．各月で乱高下するものの，緩やかであるが年々アクセス数が増加している．また，どの年も8 月はアクセスが低いことが分かる．これは，夏季休暇があるため稼働日が他の月より少ないことが原因だと考えられる．次にPoLyInfoの結果を図3に示す．こちらも，乱高下するものの月に4-6万のアクセスがあることがわかる．

また，2006年10月など急激にアクセス数が上がったものについては，特定のユーザによるデータの自動収集が行われた影響である．

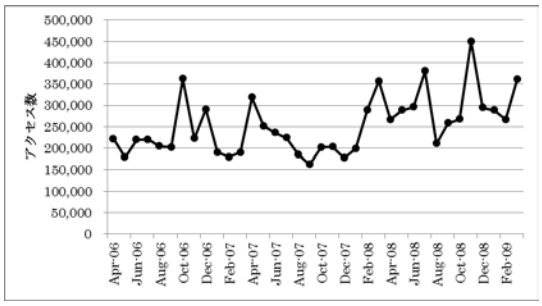

図２：物質・材料DBのアクセス数の推移

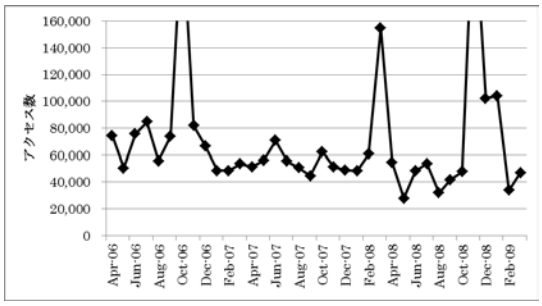

図３：PoLyInfoのアクセス数の推移

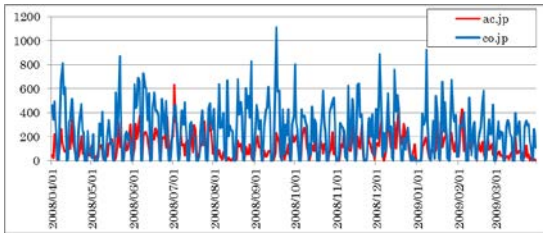

図４：大学と企業によるアクセス傾向の違い（PoLyInfo）

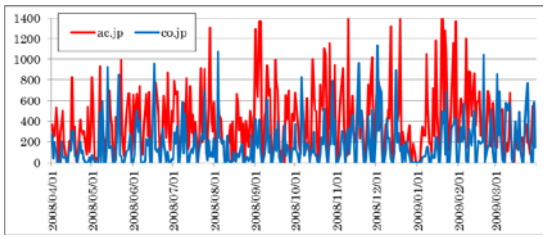

図５：大学と企業によるアクセス傾向の違い（結晶基礎データベース）

(5) データベース利用状況の相違

アクセスログにはアクセスしたユーザのアクセス元の情報(IPアドレス)が保存されているため，これを元に日本の大学と日本の企業によるPoLyInfoの利用状況を調べた．解析対象とした期間については2008年4月から2009年3月までの1年間とし，1日毎にアクセス数の集計を行った．PoLyInfoの解析結果を図４に，比較対象として結晶基礎データベースの解析結果を図５に示す．解析結果から結晶基礎データベースは大学からのアクセスが多いが，高分子データベースは企業からのアクセスが多く，ユーザの所属により利用されるデータベースが異なることが分かった．この結果から，高分子に関する調査や研究は企業の方が積極的であるという知見が得られた．

５．主な発表論文等

〔雑誌論文〕（計２件）

① 大塚真吾，宮崎収兄：女性向けフリーマガジンと連動するサイトにおけるユーザの行動分析，日本知能情報ファジィ学会誌「知能と情報」，査読有，Vol.3，2012，https://www.jstage.jst.go.jp/browse/jsoft/24/3/_contents/-char/ja/，再録決定

② Shingo Otsuka, Isao Kuwajima, Junko Hosoya, Yibin Xu and Masayoshi Yamazaki：PoLyInfo: Polymer Database for polymeric materials design, The 2-nd International Conference on Emerging Intelligent Data and Web Technologies (EIDWT-2011), 査読有，2011, DOI:10.1109/EIDWT.2011.13

〔学会発表〕（計１件）

① 大塚 真吾，細谷 順子，桑島 功，徐 一斌，喜連川 優，山崎 政義：材料データベースを利用するユーザの行動解析，第２回データ工学と情報マネジメントに関するフォーラム(DEIM2010)，B9-3，2010.3.2，兵庫

６．研究組織

(1)研究代表者

大塚　真吾（OTSUKA SHINGO）
神奈川工科大学・情報学部・准教授
研究者番号：70509736