科学技術文献を対象とした自動索引システムの開発

Development of Automatic Indexing System on scientific and technological literature

独立行政法人科学技術振興機構　知識基盤情報部主査　**関根　基樹**

PROFILE　平成 26 年 4 月より現職

1　はじめに

独立行政法人科学技術振興機構（以下「JST」）は、現在、JDreamIII[1] にて提供している科学技術文献データベースの作成工程の改善に取り組んでおり、その一環として、科学技術文献の索引を自動付与する「自動索引システム」を構築し、索引作業の効率化、品質の維持向上を目指している。

これまでも、JST では、米国国立医学図書館（National Library of Medicine）が開発し運用している自動索引システム [2] に関して調査するなどの取り組みを行ってきたが、この度、科学技術全分野の文献を対象とし、JST の実際の索引ルールに適合させながら日本語テキストから自動索引を行うシステムを構築することとなった。そのため、まずフィージビリティスタディを行って課題抽出と対応策を検討し、大学、企業等の研究者や専門家に意見を伺いながら検証を進め、その結果をもとにシステム開発の仕様を作成して 2013 年 2 月にプロトタイプシステムの開発を開始した。現在は実運用を見据えて本番システムの開発を実施中である。

本稿では、自動索引システムの概要と評価結果を紹介し、今後の展望について述べる。

2　自動索引システムの概要

2.1　自動索引の対象

（1）索引事項

JDreamIII の索引事項には以下の種類があり、現在は全て人手により索引を行っている。図 1 に、JDreamIII の回答表示例を示す。

・分類コード
・シソーラス用語
・準シソーラス用語
（大規模辞書に既登録の語、および未登録の語）
・物質索引語
（大規模辞書に既登録の語、および未登録の語）

※索引語（シソーラス用語、準シソーラス用語、物質索引語）には主題語指定を行う。
また、医学薬学等の分野ではサブヘディング索引を行うことがある。

人手による索引では、「科学技術用語シソーラス」、「大規模辞書」などの用語辞書を活用しているが、これらの辞書に登録がない用語を、準シソーラス用語としてフリー入力する場合がある。本稿ではフリー入力する語のことを「自由語」と呼ぶ。また、辞書に登録がない有機低分子化合物の新規物質などについては、別途、専用のシステムを用いて物質索引語として索引する。

自動索引システムは、上記のうち下線部分を索引対象

とし、JDreamIII の和文標題・和文抄録を入力データとすることにより自動索引を行う。なお、プロトタイプシステムでは和文本文からの自動索引についても試行している。また、科学技術用語シソーラス、大規模辞書などの用語辞書を自動索引においても使用することとした。

(2) 索引ルールの適用

JST は、索引作業に様々なルールを制定しており、索引基本方針の先頭に「必ず原文献に目を通したうえで、文献の内容について主題分析を行い、キー概念に相当する用語を適切かつ最も特定的なシソーラスに変換して索引する」ことを掲げている。

この索引基本方針を自動索引に適用するにあたり、まず「原文献」に対して自動索引を行う場合は電子データを利用できることが必須であるが、外国語文献・日本語文献ともに必ずしも原文献の全てが電子化され利用許諾が得られている訳ではないのに対し、和文標題・和文抄録は JST 作成分を含めて電子データが揃っているため、当面は和文標題・和文抄録からの自動索引を目指すこととした。ただ、電子ジャーナルなどの電子データは今後普及していくことが確実といえるため、プロトタイプシステムでは原文献の電子データからの自動索引に関しても検討した。

次に「主題分析」については、人手による索引が付与されている過去文献は主題分析の教師データとも見なせるため、対象文献と過去文献における索引語を数値化して索引語一つ一つにスコアを付けることにより、可能な限り人手作業の再現を試みた。

さらに、上位下位の階層関係にある語のスコアを調節することにより、「最も特定的なシソーラス」が選定されやすい仕組みにした。

このように、同システムはルールベースに基づいているが、検証過程では、機械学習による分類コード付与の実験 [3] も行っている。

2.2 自動索引の処理方法

自動索引の方法を処理フロー（図 2）に従って説明する。

(1) シソーラス用語、準シソーラス用語、大規模辞書に既登録の物質索引語の自動索引

①切り出し語を抽出

和文標題・和文抄録などのテキストを形態素解析により品詞分解し、切り出し語を生成する。

整理番号　12A0754331
和文標題　Xanthium italicum 地上部分からの数種の化合物の構成成分
英文標題　Constituents of several compounds from Xanthium italicum aerial part
著者名　●● ●●
資料名　一関工業高等専門学校研究紀要
JST資料番号 L0489A ISSN0913-8668 CODENJANEEC
巻号ページ(発行年月日)　Vol.25 No.2 Page.282-285 (2011) 写図表参参20
資料種別　逐次刊行物(A)
記事区分　原著論文(a1)
発行国　日本(JPN) 言語英語(EN)
抄録　Xanthium italicum のメタノール抽出物は Xanthomonous oryzaeおよびCochiliobolus miyabeanusに対する試験によって阻害活性を示すことを見出し，その活性物質の探索研究を開始した。この植物の地上部分のメタノール抽出物を水とクロロホルムに分配し，さらにクロロホルム相をメタノール水溶液と石油エーテルに分配した。メタノール水溶液可溶分をシリカゲルTLCで展開し，40のフラクションに分画した。フラクション中の成分は，再クロマトグラフィーあるいは再結晶化し，IR，1HNMR，13CNMRによって，パルミチン酸メチル，β - シトステロール，パルミチン酸，γ - ラクトンを同定し，さらに，セスキテルペン - ラクトン骨格を有するキサンタノール，カルペシオリン，イバリン，カラブロンおよびカルペシオリンの存在を推定した。
分類コード FB06020U, EH01050J (632.951, 581.192)
シソーラス用語　*オナモミ属，溶媒抽出，薄層クロマトグラフィー，分取，再結晶，NMR【磁気共鳴】，赤外スペクトル，*生体成分分析，*薬物，構造解析，ステロール，脂環式アルコール，脂肪酸
準シソーラス用語　Cochiliobolus miyabeanus，Xanthium italicum，*植物成分分析，*阻害剤
物質索引　β - シトステロール (J4.633G, 83-46-5, 5779-62-4)，パルミチン酸 (J1.378A, 57-10-3)，パルミチン酸メチル (J1.994A, 112-39-0)
DOI情報　doi : 10.1007/s00540-010-1085-0

図 1　JDreamIII の回答表示例

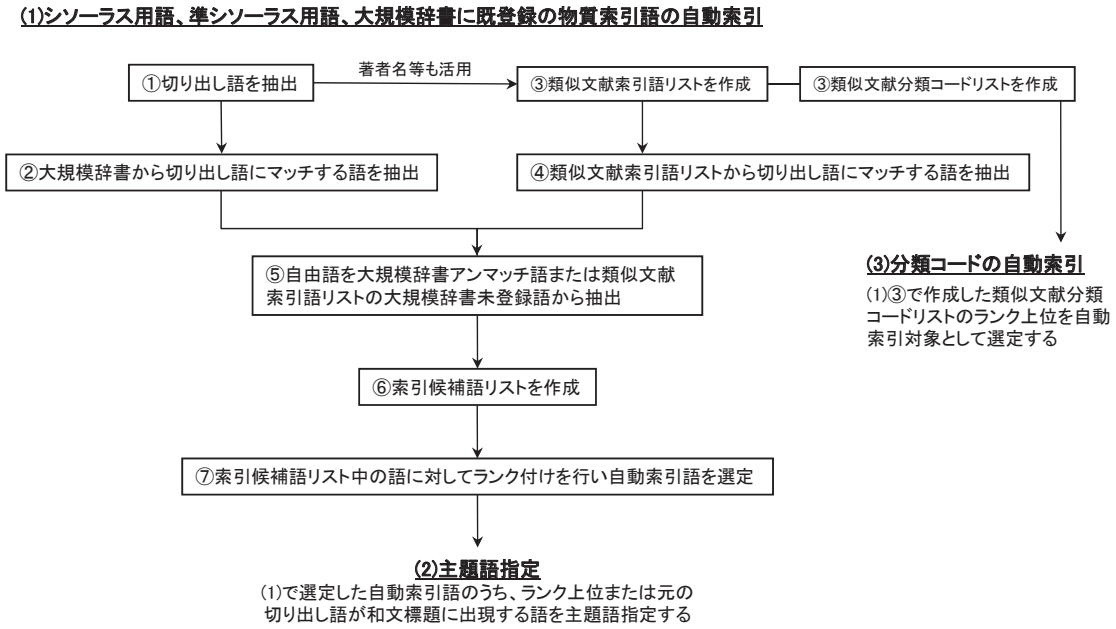

図２　自動索引の処理フロー

②大規模辞書から切り出し語にマッチする語を抽出

　切り出し語と大規模辞書のマッチングを行い、マッチする語を抽出する。

③類似文献から類似文献索引語リストと類似文献分類コードリストを作成

　切り出し語や名寄せした著者名を用いて、対象文献と類似する過去文献を選定し、索引語と分類コードをそれぞれリスト化する。

④類似文献索引語リストから切り出し語にマッチする語を抽出

　切り出し語と類似文献索引語リストのマッチングを行い、マッチする語を抽出する。

⑤自由語を大規模辞書アンマッチ語または類似文献索引語リストの大規模辞書未登録語から抽出

(a) ②において、大規模辞書とマッチしない切り出し語から自由語を抽出する。

(b) ④において、類似文献索引語リスト中の「準シソーラス索引された大規模辞書未登録語」から自由語を抽出する。

⑥索引候補語リストを作成

　②と④で抽出した語をマージして索引候補語リストを作成する。

⑦索引候補語リスト中の語に対してランク付けを行い自動索引語を選定

　索引候補語の特徴度などに基づきスコア計算を行い、閾値以上の語を自動索引語として選定する。

(2) 主題語指定

(1)で選定した自動索引語のうち、ランク上位または元の切り出し語が和文標題に出現する語を主題語指定する。

(3) 分類コードの自動索引

(1)③で作成した類似文献分類コードリストのランク上位を自動索引対象として選定する。

3　自動索引の性能評価

3.1　評価指標

人手索引されている過去文献から評価用文献を選び出し、それぞれの文献に自動索引を行って以下の値を計算し、その平均を評価指標とした。

・適合率

　＝自動索引と人手索引の一致語数／自動索引の総数

＝B／A

・再現率

＝自動索引と人手索引の一致語数／人手索引の総数

＝B／C

・F値

＝（2×適合率×再現率）／（適合率＋再現率）

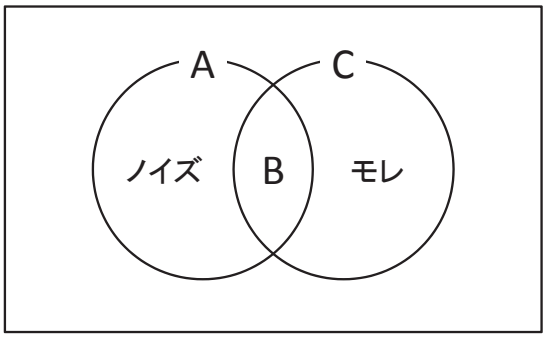

図3　適合率と再現率

なお、適合率と再現率は次のようなトレードオフの関係にある。

・適合率が高い（再現率が低い）＝ノイズは少ないがモレは多い

・再現率が高い（適合率が低い）＝モレは少ないがノイズは多い

3.2　評価結果

プロトタイプシステムでは、和文標題・和文抄録と和文本文のそれぞれに基づき、分類コードと索引語の自動索引を行い、F値を算出した。その結果を表1に示す。

表1　自動索引の結果

	F値	
	分類コード	索引語
和文標題・和文抄録	0.513	0.605
和文本文を含む（※）	（0.479）	（0.591）

（※）和文本文は方式の検討に主眼を置いたため参考値として算出

索引語は処理の一部が自動化できていないが、本番システムでは全自動が必須であるため、現在改修を進めている。ただし、分類コードは既に全自動化している。

分類コードは階層構造になっており、表1は最下位の階層での結果であるが、最上位階層でみるとF値は0.75を超えている。

JDreamIIIにおける索引は、分類コード、索引語ともに複数個の付与が必要であるが、人手でも個数の決定が難しいことなどにより、高いF値は出にくい傾向にある。しかし、今回の結果は、人手索引の支援システムとして用いることに関しては十分な可能性が示唆されていると考える。

なお、今回の評価方法は索引結果の評価であるが、今後は、自動索引された文献と人手索引された文献の検索結果を比較することによる評価も実施する予定である。

4　今後の展望

自動索引システムは現在、実運用に向けた改修を実施中であり、今年度下期からはJDreamIIIの中国文献データベースへの試験導入を計画している。

その先の本格導入にあたっては取り組むべき課題が幾つもあるため、外部の多様な知見を取り入れながら検証を進めたいと考えている。

また、JSTの科学技術用語シソーラス・大規模辞書は、日本で最大級の科学技術全分野の用語辞書であるが、自動索引のような試みに対しては登録語数が十分ではないため、自動索引の結果を辞書整備に活かすことができないか検討中である。

さらに、研究課題情報など文献情報以外への索引や分析業務への展開も視野に入れて、今後も改善を続けていきたい。

参考文献

[1] JDreamIII
http://jdream3.com/

[2] NLM Medical Text Indexer（MTI）
http://ii.nlm.nih.gov/MTI/

[3] 村脇有吾．階層的複数ラベル文書分類におけるラベル間依存の利用．自然言語処理．2014, Vol.21, No.1, p.41-60.