科学技術文献の解析・可視化における辞書の活用について

Usage of a Scientific Dictionary for Analysis and Visualization for Scientific and Technological Document

甲田彰^{1*}, 坂内悟²

Akira KOUDA, Satoru BANNAI

*1 独立行政法人 科学技術振興機構 情報提供部 営業推進課

Sales Promotion Division, Department of Service, Japan Science and Technology Agency

E-mail:kouda@jst.go.jp

2 独立行政法人 科学技術振興機構 情報提供部 普及企画課

Marketing Division, Department of Service, Japan Science and Technology Agency

E-mail:sbannai@jst.go.jp

独立行政法人科学技術振興機構(JST)では、科学技術文献検索サービス「JDream II」の検索結果を出現頻度の経年変化や類似度の分析を行い、利用者に提供する「解析・可視化機能」の開発を進めている。JDream IIには出典がはっきりした文献データ約 4,000 万件を搭載しており、そのうちの 2,300 万件については、JST 自らが抄録、索引語を付与している。今回は、現在の JDream II に搭載されている分析機能である「頻度分析機能」を説明した上で、解析・可視化機能の概要や機関名辞書を用いた解析可視化機能の分析精度向上の工夫や今後の課題について報告する。

At Japan Science and Technology Agency (JST), we are developing a new function "Analysis and Visualization" for our document retrieval system "JDream II". In this report, I describe the frequency analysis function on the current JDream, and Analysis and Visualization using our organization dictionary. Future developments are also described.

キーワード: データベース, 検索システム, 解析, 可視化, 辞書

1.はじめに

科学技術振興機構(以下「JST」)は
2006 年 4 月から、新たな科学技術文献

検索サービス「JDream II」の提供を開始した。これは、1976 年にサービスを開始した JOIS と、2003 年に主に大学や病院等の利用者向けに開発しサービスを提

供してきた JDream を統合し、機能を飛躍的に向上させた新しいシステムである。

本稿では、まず、JDream II の概要と、現在 JDream II に搭載している「頻度分析機能」を説明する。また、頻度分析機能に基づき開発を進めている解析・可視化機能について、機関名辞書の活用による精度向上と今後の課題を説明する。

2 JDream II の概要

2.1 JDream II の搭載データ

JST では、国内外の科学技術分野の資料として、科学技術系のジャーナルを初め、学会誌、協会誌、企業・大学・独立行政法人・公設試験場等の技術報告、業界誌、臨床報告等の科学技術文献を収集し、科学技術文献検索サービス「JDream II」で提供している。JDream II に搭載されている文献データは約 4,000 万件で、そのうち、JST が収集した約 2,300 万件については、各文献データに、主題分析に基づく抄録

と統制された索引語を付与している。

2.2 検索における辞書の活用

科学技術文献中には様々な異表記語や同義語が存在しているため、利用者が思いついた単語のみで検索した場合、必ずしも網羅的に検索することは困難である。JDream II では、網羅的な検索を可能とするため、同義語辞書、異表記語辞書を検索支援機能として搭載している[1]。また、19年度中には、機関名辞書も搭載する予定である。

3 JDream II の頻度分析機能

3.1 頻度分析機能の概要

JDream II では検索結果の文献データ中の索引語や著者名、機関名等をカウントする「頻度分析機能」を提供している。図 1 に検索結果を著者所属機関で頻度分析した結果の画面例を記載する。

JDream II

■ L1 に対して頻度分析を見る
対象の項目を選択し、実行ボタンをクリックしてください。

所属機関名/団体著者名(CS) [実行] ☒ これまでの抽出結果を削除する
1ページあたり 10件 表示

■ 頻度分析結果一覧
333件が抽出されました(※1-#333)

頻度分析結果一覧をすべて表示
← この抽出分をすべて選択

#	L1	コード	カウント数	抽出結果
#1	L1	CS	6	NHK 放送技研
#2	L1	CS	5	京大 大学院情報学研究科
#3	L1	CS	4	NATIONAL INST. INFORMATION AND COMMUNICATIONS TECHNOL., TOKYO, JPN.
#4	L1	CS	4	NEC
#5	L1	CS	4	国立情報学研
#6	L1	CS	3	KDDI 研
#7	L1	CS	3	KDDI 研究所
#8	L1	CS	3	UNIV. TOKYO
#9	L1	CS	2	DELFT UNIV. TECHNOL., DELFT, NLD.
#10	L1	CS	2	DOCOMO COMMUNICATIONS LAB. EUROPE GMBH

[前の見出し] [次の見出し]

◎ 全て ○ # [以降] [削除]

■ 検索語候補
語間のスペースを ○ AND ◎ OR とする

図 1 著者所属機関による頻度分析結果画面の例

3.2 頻度分析機能の活用方法

頻度分析機能を用いることにより、検索結果の文献データを一つ一つ参照することなく、全体の傾向を把握することが可能である。また、頻度分析の結果に対して、さらに頻度分析を実行することで、項目間の関連性を把握することが可能で

ある。図2に、著者所属機関で頻度分析した結果に対して、再度著者所属機関で頻度分析を行った場合の例を記載する。

こうした分析を行うことで、ある概念と別の概念の文献における関連性や、機関と科学技術分野との関連性などを把握することが可能である。

機関名	件数		件数	A 製作所	B 大学	C 研究所	D 機構
A 製作所	40		A 製作所	40	20	10	0
B 大学	30		B 大学		30	10	0
C 研究所	20		C 研究所			20	0
D 機構	10		D 機構				10

(頻度分析結果)

(頻度分析結果をさらに頻度分析)

図2 頻度分析結果に対する再分析の例

4 解析・可視化機能の開発

4.1 解析・可視化機能開発の背景

4.1.1 文献検索に対する環境の変化

これまで、JSTの文献検索サービスでは、「漏れなく検索する」ことを重視し、サービスを提供してきた。しかしながら、研究開発のスピードアップ、ボーダレス化が進む一方で、研究開発の現場で扱う情報の量は過去に比べ格段に増加している。こうした状況の変化により、研究開発の現場では、文献をひとつひとつ参照するのに加え、文献の全体的な傾向を迅速に把握する必要性が生じている。

4.1.1 JSTの文献データの特徴を活かした取り組み

JDream II に搭載した文献データは前記2.1の通り、

- a. 出典が明らかな
- b. 幅広い範囲を収集

c. 抄録と統制された索引語を人手で付与という特徴があり、全体的な傾向を正確に把握するのに適している。JDream IIの利用者からは「頻度分析機能を拡張し、解析し可視化する機能」を求める声が寄せられている。

4.2 解析・可視化機能の概要

以上の背景のもと、JSTでは、検索結果の文献データ中の索引語や機関名の出現頻度や共出現回数を分析し、グラフやマップの形式で提供する「解析・可視化機能」の開発を進めることとした。

a. 統計グラフ

機関や索引語の出現頻度を経年グラフの形で表現したもの。

b. 概念マップ

索引語の共出現や索引語と機関名の共出現をスケルトンマップで表現したもの。図3に索引語の共出現を表現したスケルトンマップの例を記載する。

図 3 スケルトンマップ(索引語)の例

4.3 精度向上のための方策

4.3.1 異表記による分析精度の低下

前記 2.1 の通り JDream II の文献データには統制された索引語が付加されているので、これを用いることで、科学技術用語については精度の高い頻度分析が可能である。

一方、著者所属機関の場合は、同一機関であっても様々な記述が存在するため、別の機関としてカウントしてしまい、出現頻度が正確に算出できない、という問題がある。

4.3.2 機関名辞書の利用による分析精度の向上

このため、19 年度に JDream II に検索支援機能として搭載する予定の機関名辞書を用いて、機関名の変換したうえで解析・可視化を行うことを予定している。JST の機関名辞書には、組織名(第 1 階層)に加え、部署名(第 2 階層)、部門名(第 3 階層)やそれらの履歴として、約 10 万個の機関名データが登録されている。

なお、日本語機関名の場合は大学名・学部名・学科名等のように、先頭に第 1 階層の機

関名の記載があるが、英語表記の場合、必ずしも先頭に第 1 階層の組織名が記載されていないため、機関名の表記を、(カンマ)や.(ピリオド)等で分解してから辞書を適用し分析する予定である。

5 今後の課題

企業の場合には資本提携のような名称からは判断できないような関係があり、辞書で分析精度の向上に対応することには限界がある。こうした情報も解析・可視化に反映するためには、将来的には利用者が機関間の関連性情報を登録し解析・可視化に反映できるような機能が必要と考えられる。

参考文献

[1] 甲田彰, 森田歌子. 科学技術文献検索システムにおける大規模用語辞書の活用について, 情報知識学会誌, vol. 16, no.2 P.45-48