Vol.2011-CH-91 No.3
2011/7/30

日本語版 Wikipedia の専門分野比較

山村あずさ[†]　　芳鐘冬樹[††]

本研究では，日本語版 Wikipedia において，図書館情報学と結晶成長学の専門用語の釈義を観察し，分野による違いを探る．そのために，Wikipedia 記事の文字，画像，表，数式，出典，リンク，見出しを計量し，分野ごとに因子分析を行い比較した．その結果，図書館情報学は，文章による説明の多い傾向と，数式による説明の多い傾向が分かれていることなどが明らかになった．これは，分野の特色が Wikipedia の記事にも表れていることを示唆している．

Comparative analysis of the descriptions of technical terms in Wikipedia between different domains

Azusa Yamamura[†]　and Fuyuki Yoshikane[††]

In this paper, we observe the descriptions of technical terms in Wikipedia for two different domains, ‘‘library and information science’’ and ‘‘crystal growth physics’’. The two domains are compared by applying factor analysis for each domain on seven variables, i.e., the numbers of characters, images, tables, formulae, references, links, and headings. The results show that the characteristics of domains, such as multidisciplinarity of library and information science, appear also in Wikipedia.

1. はじめに

一般の人が確かな専門知識を得ることは，その知識が生かされる場は様々であるにせよ近年求められてきている．科学技術が社会に与える影響は大きく，科学者は社会に対して説明を行っていく必要がある[1]．我々の社会は科学技術に基づくものであるため，生活や仕事をしていくうえで専門的な情報が必要になることがある．また将来を担う人材が生まれ育つのも社会であり，専門知識を一般社会の人が知ることは科学の発展に結びつく．

また，学術界では，そのコミュニティに属する人々が用語を共通の意味内容で認識していないと，さらに複雑な概念を正確に理解し，相手に情報を伝えることに支障が生じる．そのため，学問を学ぶうえで用語の意味内容が厳密に記されていることは非常に重要である．

専門用語の習得はこれまで，専門書で行われてきた．しかしインターネットの普及により，こういった様々な場面で必要とされる知識は誰もが知ることができるようになってきている．専門用語と一般語の境目というのは元々曖昧であるとも言われる[2]が，専門用語の公共性が高くなってくると，専門用語が一般語のように流通したり，釈義のされ方が多様になったりする．では，どのような媒体でそれが観察できるかと言えば，たとえば誰もが無料で閲覧し，編集できる Wikipedia が挙げられる．Medipedia (http://medipedia.jp)や Citizendium (http://en.citizendium.org/wiki/Main_Page)といった専門知識に特化したオンライン辞典は存在しているが，Wikipedia は知名度が高く，多くの人の目に触れやすい知識の集合体である．また，特定の分野に限らず様々な知識が集積していく点が非常に特徴的である．そのため様々な分野で専門性の高い記事が生起してくることも期待できる．これまで，Wikipedia を対象にした研究は，信頼性や特性の調査（たとえば山崎らの調査[3]）などが行われてきた．ブリタニカ百科事典と Wikipedia の科学記事の正確性を比較した Giles の研究では，Wikipedia は百科事典より多少信頼性に欠けるという結果が示されている[4]．

信頼性の問題が度々指摘されてきたが，依然として Wikipedia は全体にわたって査読されている状態ではなく，記事の責任者の所在もない．しかし，利用者を限定せずに広がり続ける Wikipedia の特徴は伝統的な冊子体の専門用語辞書にはないものである．Wikipedia における専門用語の釈義を観察し比較することにより，一般社会の人に対する専門知識の表現について，分野の特性を探ることができる．

[†] 筑波大学　情報学群
School of Informatics, University of Tsukuba
[††] 筑波大学大学院　図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

Vol.2011-CH-91 No.3
2011/7/30

2. データおよび分析方法

2.1 日本語版 Wikipedia の概況

日本語版，英語版それぞれについて，2011 年 6 月 1 日現在の Wikipedia の概況[5]を示す基本的な数量を表 1 に挙げた．記事数は英語版で 300 万を超え，日本語版でも約 75 万もの記事が存在しており，膨大な量の知識が蓄積されていることが確認できる．

表 1　日本語版 Wikipedia と英語版 Wikipedia の概況

	総記事数	総項目数	総編集数	管理者数	登録者数	活動中の登録者数	ファイル数
日本語	751,953	2,001,521	38,554,122	61	523,147	10,813	76,659
英語	3,648,422	24,081,865	465,149,155	1,789	14,657,322	144,171	844,976

英語版 Wikipedia は，全項目において日本語版 Wikipedia の 4 倍を超える値を示している．特に，管理者数と登録者数は日本語版 Wikipedia の約 30 倍にもなる．管理者とは通常の利用者には制限されている操作を行う権限を持つ利用者である．具体的には記事の保護や削除，投稿のブロックなどを行うことができる．管理者になるには他の利用者からの信任投票で一定の信任を受けなければならない．登録者とは管理者も含めた登録済みの利用者である．

総記事数は約 5 倍にすぎない，したがって，1 記事あたりで見ても，日本語版 Wikipedia の約 6 倍もの管理者・登録者が存在することになる．日本語版 Wikipedia は英語版 Wikipedia に比べて非常に少ないマンパワーで運営されていることが分かる．

2.2 比較する分野と辞書の選定

本研究では，次の 2 つの観点から，分野の特性を観察する．まず，(1) 当該分野の専門用語辞書に載録されている用語のうち，どれだけが Wikipedia にも載録されているか，すなわち，Wikipedia による専門用語のカバー率を調べ，さらに，(2) Wikipedia による専門用語釈義の量的特性（説明に用いられる文章，図表などの数量に関する特性）を明らかにする．(1)によって，一般社会に対する，その分野の専門知識の可視性についての示唆が得られ，(2)によって，一般社会に見える範囲での，専門用語集合の特徴を把握できると考える．

(1)では，図書館情報学，結晶成長学，食品微生物学の 3 分野を比較の対象とし，それらのうち，図書館情報学および結晶成長学の 2 分野に焦点を当てて，(2)について詳細な分析を行った．図書館情報学は用語の統制に積極的な分野，結晶成長学と食品微生物学は，それぞれ一般の人々の生活になじみのない分野と比較的なじみのある分野の例として選出した．図書館情報学の辞書は数あるが，結晶成長学の辞書と出版年が近いものを選んだ．

分析に使用したデータの取得方法を記す．図書館情報学用語辞典第 2 版（日本図書館情報学会用語辞典編集委員会編，丸善，2002）の 1,801 語，結晶成長学辞典（結晶成長学辞典編集委員会編，共立出版，2001）の 2,035 語，食品微生物学辞典（日本食品微生物学会監修，中央法規，2010）1,490 語のリストをそれぞれ作成し，Wikipedia にその語が存在するか検索した．その結果，図書館情報学用語辞典第 2 版 1,801 語のうち 433 語，結晶成長学辞典 2,035 語のうち 176 語，食品微生物学辞典 1,490 語のうち 292 語が存在することが分かった（図 1）．一般の人々の生活になじみのうすい分野と考えられる結晶成長学では，専門用語の Wikipedia 載録率が特に低いことを確認することができる．

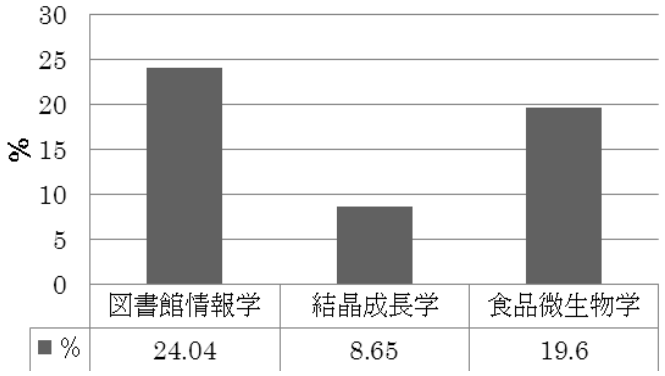


図 1　各用語辞典の Wikipedia での載録率

2.3 Wikipedia に載録されていた用語の概況

次に，図書館情報学と結晶成長学について，Wikipedia に存在する語の記事内容をダウンロードした．記事のダウンロードは 2010 年 6 月から 2011 年の 4 月にかけて行った．ダウンロードした記事それぞれの，文字数，画像数，表数，数式数，出典数，外部リンク数，見出し数を計量した．

各変数の基本統計量（最大値，中央値，平均値，標準偏差）を表に示した．また，比較しやすくするために最大値で各データの値を割ることにより規格化を行ったうえで，7 変数の状況を次のボックスプロット（箱ひげ図）で記述した．表 2 と図 2 は図書館情報学，表 3 と図 3 は結晶成長学である．図 2・3 において，箱の内部の太線は中央値，箱の右辺および左辺はそれぞれ上側および下側四分位数を表す．データの 50%が箱の中に含まれ，箱を挟む左右の線は四分位範囲の 1.5 倍以内の最小値と最大値を表す．その範囲の外の点は外れ値である．

表 2　図書館情報学の基本統計量

	最大値	中央値	平均値	標準偏差
A　文字	31824	891.5	1758	2754.75
B　画像	62	1.0	1.96	4.13
C　表	36	0.0	1.22	2.44
D　数式	44	0.0	0.26	2.72
E　出典	44	0.0	2.24	5.25
F　リンク	39	0.0	1.47	3.36
G　見出し	64	5.0	6.44	8.11

表 3　結晶成長学の基本統計量

	最大値	中央値	平均値	標準偏差
A　文字	8142	751.0	1247	1286.18
B　画像	34	2.0	3.31	4.15
C　表	15	0.0	1.10	2.17
D　数式	46	0.0	1.46	5.20
E　出典	47	0.0	1.27	3.93
F　リンク	8	0.0	0.45	1.25
G　見出し	29	4.0	4.51	5.82

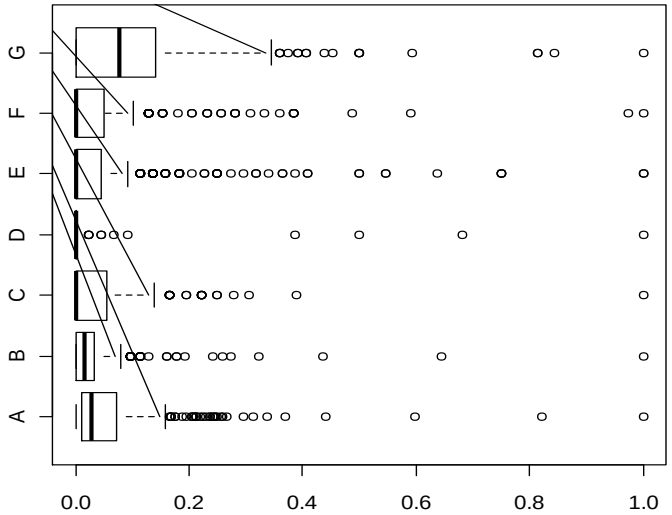


図 2　図書館情報学の文字(A)，画像(B)，表(C)，数式(D)，出典(E)，リンク(F)，見出し(G)のボックスプロット（最大値に対する比率の分布）

Vol.2011-CH-91 No.3
2011/7/30

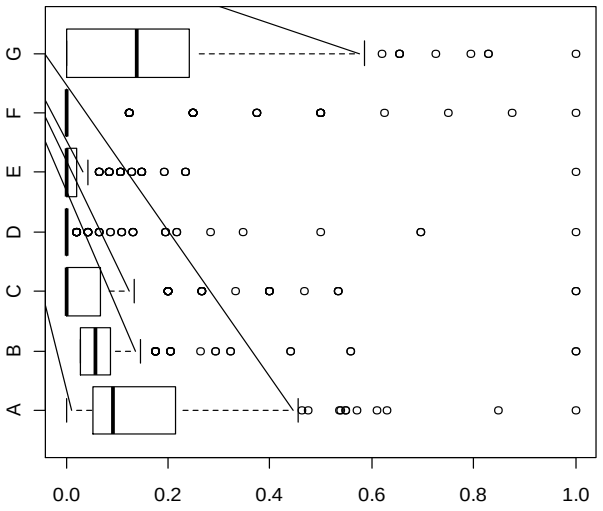


図 3　結晶成長学の文字(A)，画像(B)，表(C)，数式(D)，出典(E)，リンク(F)，見出し(G)のボックスプロット（最大値に対する比率の分布）

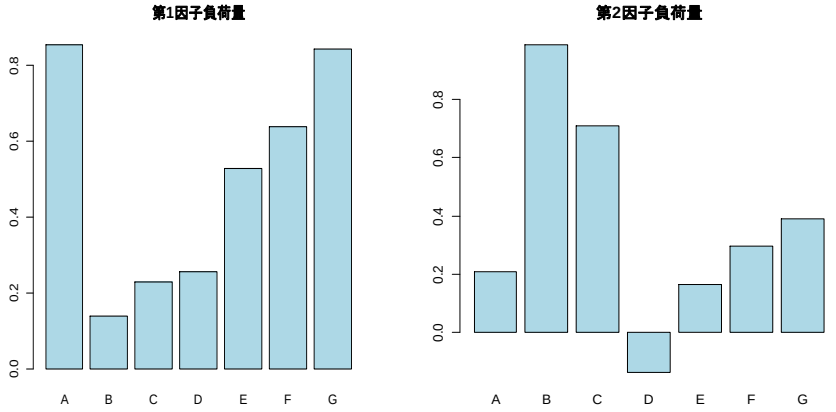
ボックスプロットによって，分布の様子を把握することができる．文字数(A)と画像数(B)については，図書館情報学よりも結晶成長学の方が箱の幅が広く，分布にばらつきがある．それに対し，リンク数(F)については結晶成長学よりも図書館情報学の方が箱の幅が広く，図書館情報学の用語は，リンクの多さの点で，より多様であることを示している．

2.4　因子分析

記事の記述に影響を及ぼしている因子を探索的に求めるために因子分析を行った．7 つの変数，すなわち，文字数，画像数，表数，数式数，出典数，リンク数，見出し数を観測変数として，それらに共通する因子を仮定し，複数の変数間の相関関係を説明する．

3.　分析結果

図書館情報学の因子分析結果を表 4 と文字(A)，画像(B)，表(C)，数式(D)，出典(E)，リンク(F)，見出し(G)
図 4 に示し，結晶成長学の因子分析結果を表 5 と



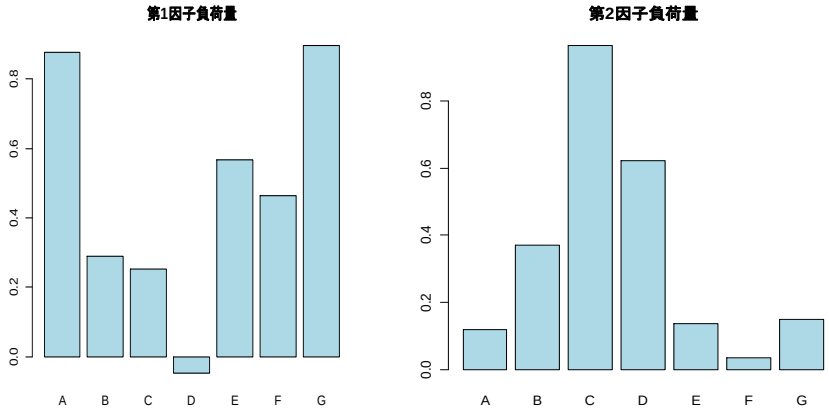
文字(A)，画像(B)，表(C)，数式(D)，出典(E)，リンク(F)，見出し(G)
図 5 に示した．図書館情報学の第 1 因子は文章と見出しに大きく依存し，第 2 因子は表，数式の順に依存する傾向がはっきりと見られる．つまり，図書館情報学の記事は，文章量が多い傾向と，数式が多い傾向が明確に分けられる．因子に意味づけをするならば，第 1 因子は文章での説明が多い文系的な因子であり，第 2 因子は数式での説明が多い理系的な因子であるといえる．

一方，結晶成長学の第 1 因子も文章と見出しに大きく依存している，図書館情報学よりも数式の因子負荷量が多いことが特徴的である．第 2 因子は画像，表の順に依存している．因子に意味づけをするならば，第 1 因子は文章で説明する傾向の因子であり，第 2 因子は画像や表での説明が多い図鑑的な因子であるといえる．

Vol.2011-CH-91 No.3
2011/7/30

表 4　図書館情報学の因子分析結果

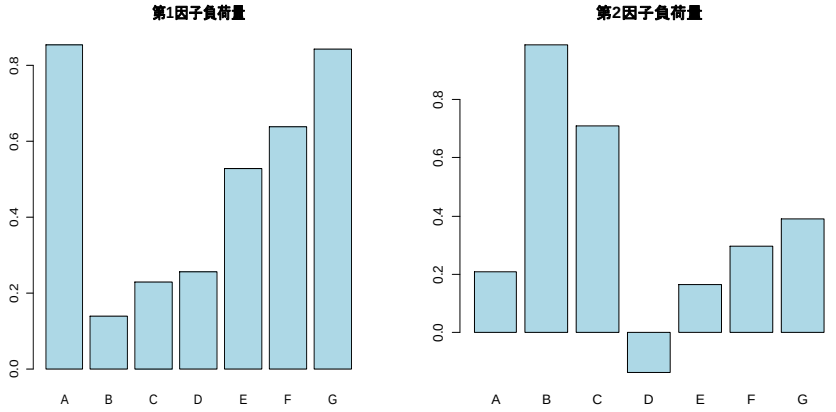
要素	第１因子	第２因子	共通性
見出し	0.896	0.148	0.825
文字	0.876	0.118	0.781
出典	0.567	0.136	0.340
リンク	0.465	0.034	0.217
画像	0.289	0.372	0.221
表	0.254	0.965	0.995
数式	-0.047	0.623	0.390
因子寄与	2.259	1.512	



文字(A)，画像(B)，表(C)，数式(D)，出典(E)，リンク(F)，見出し(G)
図 4　図書館情報学の因子負荷量

表 5　結晶成長学の因子分析結果

要素	第１因子	第２因子	共通性
文字	0.917	-0.088	0.772
見出し	0.872	0.111	0.864
リンク	0.660	0.086	0.495
出典	0.560	-0.017	0.305
数式	0.311	-0.240	0.084
表	0.129	0.677	0.556
画像	-0.022	1.008	0.610
因子寄与	2.464	1.559	



文字(A)，画像(B)，表(C)，数式(D)，出典(E)，リンク(F)，見出し(G)
図 5　結晶成長学の因子負荷量

Vol.2011-CH-91 No.3
2011/7/30

4. おわりに

因子分析の結果から，図書館情報学と結晶成長学では記事の記述に影響を与える因子が異なるといえる.図書館情報学は図書館学と情報学を融合させた分野であるため，図書館学の文章主体の説明と情報学の数式での説明に分かれているのではないかと考えられる．一方，結晶成長学は物理学から派生した分野であるが，Wikipedia における専門用語の記述でも，鉱物や結晶の写真での説明など，図鑑的な性質に関わる要因が確認された．記事に責任表示がなく伝統的な専門用語辞書とは異なるプロセスで編集される Wikipedia であるが，このように，それぞれの分野が持つ特色が記事にも反映されるような現象が起こっていることを示唆する結果となった.

参考文献

1) 藤垣裕子: 科学者の社会的責任の現代的課題(科学は今…), 日本物理學會誌, Vol.65, No.3, pp. 172-180 (2010).
2) 仲本秀四郎: 用語「情報」: ターミノロジー的考察, 情報の科学と技術, Vol.52, No.6, pp.339-342 (2002).
3) 山崎由佳, 伊藤貴一, 井庭崇, 熊坂賢次: Wikipedia の経年変化に関するカテゴリ間の比較分析, 情報処理学会研究報告. BIO, バイオ情報学, No.126, pp.183-186 (2007).
4) Giles, J: Internet encyclopaedias go head to head, NATURE, Vol.438, No.7070, pp.900-901 (2005).
5) Wikipedia：全言語版の統計, http://ja.wikipedia.org/wiki/Wikipedia:全言語版の統計