Vol.2015-CH-106 No.13
2015/5/16

クラウドソーシングってどうですか？ Crowd4U × NDL データの事例

森嶋 厚行[1,a)]　川島 隆徳[2,b)]　原田 隆史[3,2,c)]　宇陀 則彦[1,d)]

概要：近年，クラウドソーシングは問題解決の新しいアプローチとして注目を集めている． 本講演では，クラウドソーシングの応用事例として，クラウドソーシング プラットフォーム Crowd4U を用いた NDL データ利用プロジェクトについて説明する． Crowd4U は，非営利・公益・学術目的のクラウドソーシングプラットフォームであり， 公益と学術のタスクが稼働している．Crowd4U は大学によって開発が行われており， プロジェクトの要望に応じて様々な機能追加が日々行われている． L-Crowd プロジェクトは，Crowd4U 上で NDL データを用いて行われているプロジェクトの 一つであり，ISBN による書誌同定における誤り (書誌誤同定) の判定をマイクロタスクで行おうというものである． 本講演では，本事例の紹介を通じて，クラウドソーシングを利用した公益・学術 プロジェクトの可能性を議論したい．

1. はじめに

近年，クラウドソーシングは問題解決の新しいアプローチとして注目を集めている． 本講演では，クラウドソーシングの応用事例として，クラウドソーシング プラットフォーム Crowd4U を用いた NDL データ利用プロジェクトについて説明する．本講演では，本事例の紹介を通じて，クラウドソーシングを利用した公益・学術 プロジェクトの可能性を議論したい．

2. Crowd4U

Crowd4U[1][5][6] は，非営利・公益・学術目的のクラウドソーシングプラットフォームであり，マイクロタスク型クラウドソーシングプラットフォームに分類される．マイクロタスク型クラウドソーシングとは，問題を解決するためのタスクを，短時間で作業が出来る小さなタスク (マイクロタスク) の集合に分割し，不特定多数の人々に委託するものである．商用のプラットフォームとしては，Amazon Mechanical Turk や国内では Yahoo!クラウドソーシング等がある．Crowd4U は，2011 年より 11 月より公開されており，現在では複数の学術・公益プロジェクトに利用され，2015 年 4 月時点で 29 ヶ国からの登録貢献者がいる (図 1)．また，最近は，週平均平日タスク数 600〜1000 程度の作業が行われている．Crowd4U は次のような特徴を持つ．

公益と学術のためのクラウドソーシングプラットフォーム
Crowd4U は公益と学術の利用に限定したクラウドソーシングプラットフォームである．図書館領域 [8]，自然災害領域 [9]，情報検索応用 [10] をはじめとした様々なクラウドソーシングプロジェクトが稼働している．

学術コミュニティの協力で開発され，運用されている
Crowd4U は国内外の研究者の要望に応じて大学で開発が行われており，様々な機能追加が日々行われている． 例えば，タスクの柔軟な生成・表示機能や，便利なタスク管理画面などがこれまで実装されて来た．

あらゆるタスクが可能　Crowd4U はクラウドソーシングのための宣言型プログラミング言語 CyLog[2][4] を提供しており，複雑なクラウドソーシングが得意である．これにより，クラウドソーシングのタスク結果に応じて柔軟にタスク内容を切り替えると行った高度な処理が可能になっている．この特徴は，タスクの文面を翻訳する別のタスクを生成するといった様々な形で活用されている．

様々なインセンティブの提供　Crowd4U は, 様々なインセンティブを提供する．最も基本的なものは，タスクへの説明文埋め込み (タスクの意義を説明する)，プロジェクトメンバとしての貢献者の記載，タスク数のラ

1 筑波大学
Tsukuba-city, Ibaraki 305-8577, Japan
2 国立国会図書館
Nagata-cho Chiyoda-ku Tokyo 100-8924, Japan
3 同志社大学
Kamigyo-ku, Kyoto-city, Kyoto 602-8580, Japan
a) mori@slis.tsukuba.ac.jp
b) t-kawash@ndl.go.jp
c) ushi@slis.doshisha.ac.jp
d) uda@slis.tsukuba.ac.jp

登録タスク数	78,995
登録貢献者数	602
登録者の国数	29
登録プロジェクト数	10

図 1 2015 年 4 月 16 日現在の統計量．Crowd4U で作業を行うのに登録は不要であるため，匿名を含む貢献者の数は 2,000 人以上と推測している．

図 2 床タスクシステム，図書館に設置された Crowd4U タブレット，スマートフォンスクリーンロックシステム

ンキングである．また，様々なデバイスを利用したインセンティブの提供も行っている (図 2)．床タスクシステムは，現在，筑波大学，同志社大学，明治大学の 3 大学キャンパスに設置されており，平成 27 年度には筑波大学附属図書館への設置も行われる予定である．

タスク開発支援 Crowd4U では，近日中に，エンドユーザが容易に直接タスクを登録するためのツールを公開予定である．これにより，簡単なタスクであればデータの入ったテキストファイルをアップロードすることによりタスクが誰でも簡単に追加可能になる．

オープンである 公益と学術目的であれば誰でも利用でき，ソースコードの提供も可能である．また，Crowd4U の API 等を通じて，他アプリケーションや他のクラウドソーシングプラットフォームとの連携なども容易である．典型的な利用方法は二つある．第一に，ボランティアベースのクラウドソーシングプロジェクトである．この場合は，ボランティアのリクルートにもご協力いただき，他のプロジェクトにも参加いただけるような Crowd4U ネットワークの構築にご協力いただいている．この利用方法は，長期プロジェクトに向いている．第二に，商用プラットフォームを用いて作業者をリクルートし，Crowd4U の高度機能を用いた複雑なクラウドソーシングを行う事である．この利用方法は，短期間に多くの貢献者を確保するのに向いている．

3. L-Crowd プロジェクト

L-Crowd プロジェクト [7] は，Crowd4U 上で NDL データを用いて行われているプロジェクトの一つであり，図書館領域に関する問題に対するクラウドソーシングの適用を試みるものである．最初の試みとして，ISBN による書誌同定における誤り (書誌誤同定) の判定をマイクロタスクで行うというものである．国内の複数の大学からの協力者によって進められている [3][8]．本プロジェクトは，NDL が持つユニオンカタログのデータを対象として進めている．近年，異なる組織が持つ書誌レコードを統合する試みが数多く行われている．そこでの問題の一つは，同一のソースが，しばしば複数の書誌レコードで表現されていると言う事である．したがって，同一のソースを指すレコードを同定する作業が必要とされ得る．

書誌が機械可読名場合に，アルゴリズムによる機械的な書誌同定がしばしば行われる．NDL でもそのような書誌同定作業を行っている．一般的なアプローチは，各レコードが含む値から同一のソースを指すキーを求め，それらを比較することである．ISBN や MARC 番号等が利用ことがあるが，現時点では，全ての書誌レコードに存在するのは ISBN だけであるため，MARC 番号などは補助的に利用されることが多い．

しかし，機械的な書誌同定はうまく行かないことが多い．それにはいくつかの理由がある．第一に，同じソースを指す書誌レコードでも，入力する人が異なれば異なる書誌レコードになる事である．例えば，図 3 は，同一のソースを指す異なる書誌レコードの例である．NDL では他にも，同一のソースを指すレコードが，片方は英語，片方は日本語で入力されている場合もある．第二に，入力されているデータが正しい場合にでも，ISBN が完全なソースの識別子として働かない場合が多々あるからである．よく見かけられる例は，ある書籍の改訂版に同じ ISBN を付与している例である．旅行ガイドブックなど，毎年出版される書籍で見受けられる．また，シリーズものの書籍に同じ ISBN が付与されている場合もある．更には，同じ出版社の全く異なる書籍に同じ ISBN が付与されている場合も存在する．

以上のことから，ISBN による機械的な同定を行うと，異なる資料が誤同定され，検索できなくなってしまう場合がある．一方で，タイトルなどを同定条件として追加すると，書誌の取り方の違いから，同じ資料が同定されないという事象が起きうる．機械的な同定においてはこれはトレードオフとなっているが，問題は，そのような ISBN による誤同定がどの程度存在するのかが明らかになっていない，ということである．そこで，本プロジェクトでは「明らかな誤同定」，すなわち全く異なる本が何らかの理由で同じ ISBN を割り振られている場合がどの程度の規模で存在するのかを明らかにしたいと考えている．このような作業は，完全に機械化することは前述の理由により不可能であるが，人間であれば容易に判定が可能であり，マイクロタスクに適していると言える．

書誌レコードが入力されると，次の手順でタスクが生成される．まず，同じ ISBN を持つ書誌レコードのグループ化を行う．次に，各グループに含まれる書誌レコード毎に比較する組合せを作る．この組合せは複数のアプローチが考えられる [8]．最後に，各組合せに関してタスクが生成される．図 4 はタスクの例である．ここでは，上下の書誌レコードを比較し，異なる場合にはチェックを行うという作業を行う．

Title	Series	Publisher
Towards the e-society : e-commerce, e-business, and e-government : the first IFIP Conference on E-Commerce, E-Business, E-Government (13E 2001), October 3-5, 2001, Zurich, Switzerland / edited by Beat Schmid, Katarina Stanoevska-Slabeva, Volker Tschammer	The International Federation for Information Processing ; 74	Kluwer Academic Publishers
Towards the e-society: e-commerce, e-business, and e-government : the first IFIP conference on e-commerce, e-business, e-government (13E 2001) October 3-5, 2001, Zurich, Switzerland. : Oct 2001, Zurich, Switzerland	IFIP ; 74	Kluwer Academic Publishers

図 3 同じ ISBN を持つ書誌レコードの例

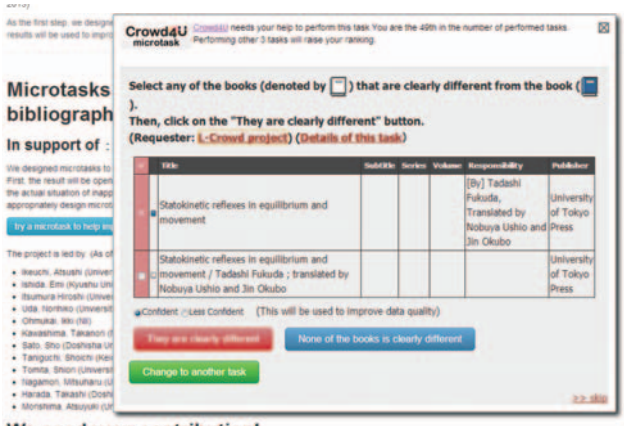

図 4 Crowd4U タスクの画面例

本実験では，ISBN で同定された書誌レコードグループのうち，他の書誌事項が異なる 12277 組のグループから，22764 組のタスクが生成されている．なお，タスク数が書誌レコードグループ数より多いのは，同じ ISBN を持つ 3 つ以上の書誌を含む書誌レコードグループが存在するためである．各タスクについて 3 回の判定が行われており，うち 15915 タスク（全体の約 69%）については，3 回の試行において結果が一致している．この結果が一致しているタスクのうちの 5519 タスク（全体の約 24%）については，そのタスク中に他と異なる書誌が含まれている，すなわち誤同定であると判断されたタスクとなっていることが判っている．今後，サンプル調査によるタスクの信頼性の評価や，最終的な誤同定書誌の規模を分析していく．

4. L-Crowd で利用する Crowd4U の機能

L-Crowd では次の Crowd4U の機能を利用している．

- データに基づく自動的なタスクの生成: 元の書誌レコードを組み合わせてタスクを生成する作業は，CyLog プログラムとして記述され Crowd4U 上で実行される．したがって，書誌レコードを追加するだけでタスクが自動生成される．
- 様々なインセンティブ構造: L-Crowd では，書誌誤同定タスクの意義の説明文の埋め込みと，ランキングの機能を利用している．また，PC 上および Crowd4U 端末上でタスクを提供している．
- タスクの順序制御: 連続してタスクを行うときに同じようなタスクが並んで飽きないよう，タスクの出現順を制御している．
- 品質管理のための情報提供: 九州大学櫻井祐子先生の研究に基づき，タスクの作業時に自信の有る無しの情報を入手している．この情報と品質との関連を今後調査予定である．

5. 今後に向けて

第 3 章に示すように，L-Crowd によつて ISBN をキーに機械的に書誌同定を行った場合，その中で書名などが一致しない組み合わせの少なくとも 24%が誤同定であることを確認することができた．また，同時に機械による ISBN をキーとした同定処理では正確な判定が困難であったもののうち，約 45%を正しく判定することができた．このように，機械だけでは同定が困難である組み合わせについても L-Crowd のような仕組みを使用することで正しい判定が行えることは，実用システムに対する寄与としても非常に大きいと考えられる．実際に国立国会図書館の NDL サーチにおける書誌同定では，誤同定を行った結果として見つからなくなる書誌が最小限となるように ISBN による同定だけではなく書名の一部の情報なども加味して判定を行っている．判定結果をさらに細かく分類し，どのように取り入れることが可能であるかなどの検討が期待される．

ただし, L-Crowd では全体の約三分の一にあたる約 31%について判定者の結果が一致しなかった．その原因などの分析はこれからであるが，いくつかの例を見ただけでも，たとえば毎年刊行される図書において「xx 年度版」が記載されていない例のように書誌事項の一部が欠けているものを判定する場合や，シリーズ名まで含めてタイトルとしている場合と各巻のタイトルのように書誌事項の記載レベルに違かある場合など，いくつかの典型的な例が散見される．このような誤同定の原因を分析することは，同定処理の精緻化にも貢献することが考えられる．今後，参加者の意欲を高める工夫とともに内容の分析も進め，今後とも書誌同定に対する効果的な手法を検討していきたい．

謝辞 Crowd4U 開発者, 協力者の皆様, L-Crowd プロジェクトの関係者の皆様, そして数多くの Crowd4U ボランティアの方に感謝申し上げます．彼らの貢献無しに Crowd4U

は成り立ちません．開発者・協力者・登録貢献者の皆様の一覧は http://crowd4u.org にあります．登録貢献者は実際の貢献者の方のごく一部です．また，L-Crowd プロジェクトの関係者は http://crowd4u.org/projects/lcrowd に有ります．本研究の一部は科研費基盤研究 (#25240012) および科学技術振興機構さきがけの支援による．

参考文献

[1] Crowd4U. http://crowd4u.org.
[2] Shun Fukusumi, Atsuyuki Morishima, Hiroyuki Kitagawa. Game Aspect: An Approach to Separation of Concerns in Crowdsourced Data Management. 27th International Conference on Advanced Information Systems Engineering (CAiSE 2015), June 8-12, 2015.
[3] L-Crowd project. http://crowd4u.org/projects/lcrowd.
[4] Atsuyuki Morishima. CyLog/Crowd4U: A Case Study of a Computing Platform for Cybernetic Dataspaces (Invited Chapter). Handbook of Human Computation, Springer, pp. 561-572, Nov. 2013.
[5] Atsuyuki Morishima, Sihem Amer-Yahia, Senjuti Basu Roy. Crowd4U: An Initiative for Constructing an Open Academic Crowdsourcing Network. Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2014) WorkInProgress, pp. 50-51, Pittsburgh, USA, November 2-4, 2014.
[6] Atsuyuki Morishima, Norihide Shinagawa, Tomomi Mitsuishi, Hideto Aoki, Shun Fukusumi. CyLog/Crowd4U: A Declarative Platform for Complex Data-centric Crowdsourcing, PVLDB 5(12): 1918-1921 (2012)
[7] Atsuyuki Morishima, Takanori Kawashima, Takashi Harada, Norihiko Uda, Ikki Ohmukai. L-Crowd: A Library Crowdsourcing Project by LIS and CS Researchers in Japan (Invited Talk and paper), International Conference on Digital Libraries (ICDL2013), pp. 40-47, November 2013.
[8] Atsuyuki Morishima, Shiori Tomita, Takanori Kawashima, Takashi Harada, Norihiko Uda, Sho Sato, Yukihiko Abematsu. A Crowdsourcing Approach for Finding Misidentifications of Bibliographic Records. iConference 2014, pp. 177-191, 2014.
[9] 丹治寛佳, 森嶋厚行, 井ノ口宗成, 北川博之,「Web 情報を用いた竜巻経路推定支援のためのクラウドソーシング技術開発の試み」情報処理学会論文誌 データベース（TOD60）, vol.6,No.5,pp95-106,2013 年 12 月 27 日.
[10] 渡辺知恵美, 中村聡史, オノマトペロリ：味覚や食感を表すオノマトペによる料理レシピのランキング, 人工知能学会論文誌, Vol.30, No.1, pp.340-352, 2015.