Vol.2015-CH-105 No.9
2015/1/31

知識・言葉の情報学と地域研究

亀田 尭宙[†1]

地域研究はそれ自体が学際的な学問で，異なるディシプリンで異なる地域の異なる現象を読み解いてきた学問と言える．そこで，地域情報学の名のもとに情報資源の統合や共有化が図られ，空間や時間といった切り口で地域やディシプリンを超えて研究を俯瞰する試みが行われてきた．本発表では，Linked Open Data と自然言語処理の技術で新たな切り口を提供しようという試みについて紹介し，課題や将来展望について語る．

1. 地域研究と言葉の情報学

地域研究は，比較的新しい学問であり，世界の各地域を政治・文化・宗教といった様々な切り口から包括的に理解しようと試みる学際的な学問である．各研究者はフィールドと呼ばれる研究対象地域を定め，フィールドワークを通して得た情報と公的な情報，他の学術的な情報を組み合わせて，その地域に生じていることを読み解いていく．一方，そのような切り口の幅広さは，個々の地域研究の関連を複雑なものにしており，研究成果をただ並べるだけでは，比較することさえ困難である．

そこで，情報学な手法を用いて，地域に存在する様々な「地域の知」を比較可能な方法で構造や意味を理解するという試みが行われてきた[1]．経度・緯度・標高で構成される３次元空間に，時間を加えた４次元空間の中で情報を整理することで，地域研究に関わる幅広い情報を同じ空間のなかに位置づけることで，比較や俯瞰による知の発見が可能になったわけである．

具体的には，地域ごとの仏教僧の修業期間を時間軸でそろえてグラフ化し比較することで，修業期間のパターンの地域差が見えてくるという例や（図１）[2]，ハノイの都市の変容過程を，史資料から，地勢の情報から，衛星から観る情報から，地下構造の情報から，読み解いた例[3]がある．

そして，近年トピックという切り口を加えて，地域資料を統合する試みが始まっており，その代表的なものが高谷好一氏のフィールドノートを Google Earth 上にマッピングした上で，フィールドノートのトピック構造を Latent Dirichlet Allocation（以下 LDA） を用いてモデル化し色分けした研究である4（図２）．コショウとコーヒーが同じ農作物という意味でフィールドノート内の多様な単語の中で近接性を持っており，コショウの記述が現れるノートが時空間的にも近接しているといった構造が見て取れるようになっている．地域研究の資料には写真や装飾品等の非文字資料も多くあるが，それらのメタデータは文字で表現されて管理されること，フィールドノートや雑誌，新聞といった文字資料も多く扱われていることから，言葉の情報学による地域情報学の発展が期待される．

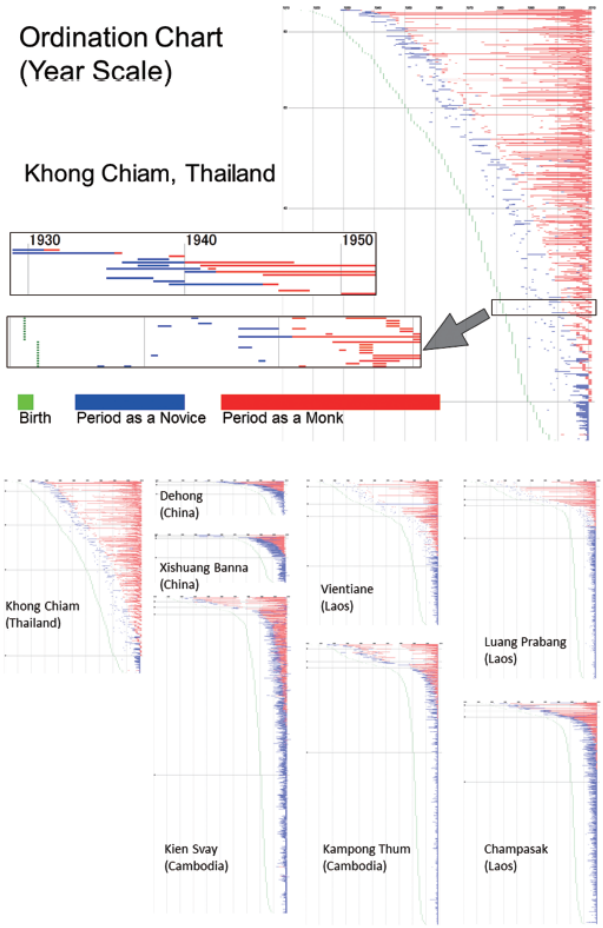

図 1　得度チャート

Figure 1　Ordination Chart

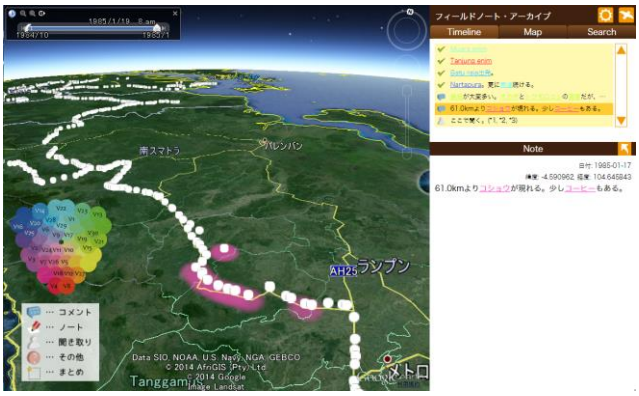

図 2　高谷フィールドノートのマッピング

Figure 2　Mappings of Takaya Field Notes

†1 京都大学
Kyoto University

ⓒ 2015 Information Processing Society of Japan

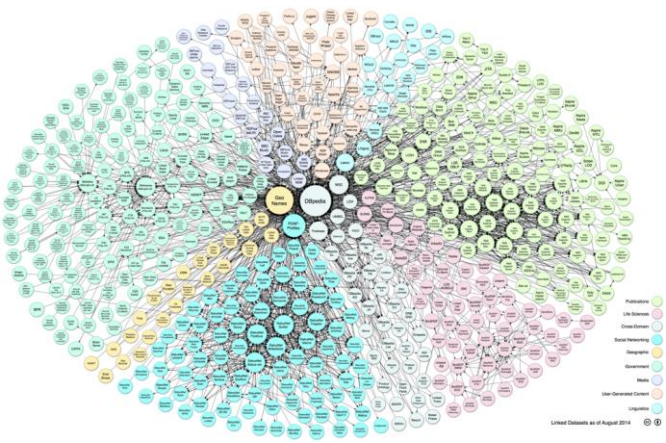

図 3　LOD クラウド
Figure 3　LOD Cloud

2. 知識体系との関連付け

LDA のような手法による知の構造化は，文書集合から自動的に生成できるために，ドメイン知識を要さずに可能なことが一つの大きなメリットであるが，それと同時に人間の理解と隔たりがあるという欠点を孕みやすい．それぞれの文書や語が割り当てられる先のトピックは，語の集合で以って表現されることが一般的であるが，その語の集合を眺めても何を表現した集合なのかが分からないこともしばしばである．

人間の知識を直接的に表現してコンピュータに扱えるようにしようとしてきた学問の流れとしては，知識工学があり，より具体的には，セマンティックウェブの研究，オントロジ工学，それらを引き継ぎつつ近年急速に発展している Linked Open Data（以下 LOD）の研究が挙げられる．

Linked Data の原則は

1. あらゆる事物を URI で記述すること
2. HTTP URI を用いることでアクセス可能にすること
3. 記述に RDF を用いることで有用な情報を提供すること
4. さらなる情報を発見するための外部リンクを提供すること

の 4 つである．LOD においては，それに加えてオープンライセンスによる情報共有という実践が結合している [5][6].

LOD の世界では知識のカテゴリの表現などは DBpedia (a) のような幅広いものから，個々のドメインにおける知識カテゴリのデータまで多くの知識が情報資源として整備されており，それらが相互に関連付けられている（図 3 参照．図 3 はカテゴリ構造に限らずより多様な関係を以って色で分けて示された生命科学や行政など様々なドメインの知識が相互に結び付けられ発展した様子を表している(b)）．

たとえばコショウとコーヒーがそれぞれ作物の一種だということが明示でき，その地域の行政によって開示されている生産量の情報などと関連付けることができる．

さらに，LDA の結果と合わせて読み解くことで，一般には作物と思われていないものがフィールドノート内では作物と強く関連付けられているなどといったことも分かってくる．言い換えれば，言葉の情報学を用いて資料から立ち上がってくる知識構造と，知識の情報学における知識の網の目を接合することでフィールドワークから既存の知識の枠を越えたものを読み解いていくという地域研究的な色を帯びつつ，現実的な理解が可能な地域情報学が可能になると考えている．

3. 課題

今後，トルキスタン集成(c)やカラムデータベース(d)といった地域研究の情報資源を，知識・言葉の情報学を用いて読み解いていきたいと考えている．

既存の自然言語処理のツールが容易に使えないような言語が用いられていること，つなげる先の LOD も同時に整備する必要があるほどユニークなデータであること，データの多くが人手によるもので記号や表記ゆれがあり解析までの前処理に工夫が必要であるといったことが今後取り組むうえでの課題である．

参考文献

1) 柴山　守 : “〈特集〉地域情報学―地域研究と情報学の新たな地平―　序論”. 東南アジア研究 (京都大学東南アジア研究所), 2009.
2) 林 行夫，柴山　守，Julien Bourdon-Miyamoto，長谷川　清，小島　敬裕，小林　知，高橋　美和，笹川　秀夫，土佐　桂子，須羽　新二 : “宗教実践を可視化する　―大陸部東南アジア上座仏教徒の寺院と移動―”，CIAS Discussion Paper No.42，2014.
3) 柴山 守 : “地域情報学の創出プロジェクト”，地域情報学ニューズレター No. 1，京都大学東南アジア研究所，2006.
4) 高田 百合奈，渡邉 英徳，柳澤 雅之，山田 太造 : “位置情報とトピックモデルに基づくフィールドノートのビジュアライズ手法”，情報処理学会論文集「人文科学とコンピュータシンポジウム(じんもんこん)2014)」，2014.
5) Bizer, Christian; Heath, Tom; Berners-Lee, Tim: Linked Data—The Story So Far, International Journal on Semantic Web and Information Systems 5(3), pp. 122, Solving Semantic Interoperability Conflicts in CrossBorder E-government Services，2009.
6) Tom Heath, Christian Bizer: Linked Data: Evolving the Web into a Global Data Space, 2011. (邦訳:Linked Data: Web をグローバルなデータ空間にする仕組み，武田英明監訳，大向一輝，加藤文彦，嘉村哲郎，亀田尭宙，小出誠二，深見嘉明，松村冬子，南佳孝訳，2013).

a http://dbpedia.org/
b Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

c ロシア帝国の中央アジア征服の結果タシュケント（現在のウズベキスタンの首都）にトルキスタン総督府が設置された後，初代総督カウフマン K. P. fon-Kaufman の命により，19 世紀後半から 20 世紀初頭にかけて収集された，当時の中央アジアに関する文献資料コレクションである．全 594 巻および 4 種の索引から成る．
http://app.cias.kyoto-u.ac.jp/infolib/meta_pub/G0000003Turkestan
d 『カラム』は過去に発行されていたマレー語の月刊誌．1950 年代～60 年代のマレー世界におけるムスリム社会の動向を理解するうえで重要な史料で，ジャウィ（マレー語のアラビア文字表記）で表記されている．京都大学地域研究統合情報センターがローマ字化およびデータベース化をしている．
http://majalahqalam.kyoto.jp/