テキスト解析による社会と学術的な関心の推移に関する研究～防災分野を対象として～

公共システム研究室　河野夏樹

1. はじめに

わが国では自然災害が多く，近年では極端な災害の頻発による被害の拡大なども問題となっている．また，急速な人口減少と高齢化が進行しており，地域の活力の低下が懸念されている．これらの背景のもとで，防災に関する社会的な関心は高まっており，社会的な関心に応えるための防災研究の推進が期待されている．しかし，学術的な関心の状態や，それらがどのような変遷をたどり，その時々の社会的な関心とどのような関係にあったのかを知ることは容易ではない．このため，現在までの学術的な活動を批判的・客観的に概観することが困難であり，これに伴い今後に向けた戦略的な改善も困難である．

双方の関係が明らかにされてこなかった理由として，学術的な関心にせよ社会的な関心にせよ，それらを定量化するための手法が十分に開発されていなかったことがあげられる．しかし，近年では，テキスト解析手法が開発され，これらの分析が可能になっている．加えて，分析するためのデータについても入手が容易になっている．そこで，本研究では，防災分野を対象に，学術的な関心と社会的な関心の距離の推移をテキスト解析により明らかにする方法論を開発する．

2. 本研究の考え方

学術的な関心については学術論文，社会的な関心については新聞記事を用いて定量化する．その際，トピックモデルを用いて複数の特徴的なトピックを抽出することで，学術的な関心と社会的な関心を明らかにする．その上で，これらのトピック間の距離をジェンセン・シャノン情報量を用いて算出し，その時系列的な推移を明らかにする．

一方で，論文や新聞記事には，それらの内容を特徴付ける領域（キーワード）が存在することから，それらの単語に着目して，学術的な関心と社会的な関心の距離を分析する．さらに，これらの距離が論文数と記事数のどちらに影響を受けているのかを統計的に明らかにすることで，学術研究が社会の関心を先取りしていたのか，後追いをしていたのかを判別し，学術的な研究がどのように社会の要請に応えているのかを評価する．

3. 分析手法

3.1　トピックモデル

論文や新聞のトピックを抽出するにあたり，トピックモデルを用いる．文書 d の n 番目の単語にトピック k が割り当てられる可能性（負担率）は次式のように求められる．

$$q_{dnk} = \frac{\theta_{dk}\phi_{kw_{dn}}}{\sum_{k'=1}^{K}\theta_{dk'}\phi_{k'w_{dn}}} \tag{1}$$

負担率が与えられたもとで，文書 d のトピック k の確率 θ_{dk} は次式で表される．

$$\theta_{dk} = \frac{\sum_{n=1}^{N_d} q_{dnk}}{\sum_{k'=1}^{K}\sum_{n=1}^{N_d} q_{dnk'}} \tag{2}$$

また，トピック k で語彙 v が出現する確率 ϕ_{kv} は次式で表される．

$$\phi_{kv} = \frac{\sum_{d=1}^{D}\sum_{n:w_{dn}=v} q_{dnk}}{\sum_{v'=1}^{V}\sum_{d=1}^{D}\sum_{n:w_{dn}=v'} q_{dnk}} \tag{3}$$

以上の変数を EM アルゴリズムを用いて求める．以下では，論文をデータとして導出されるトピックを学術的な関心，新聞記事をデータとして導出されるそれを社会的な関心とし，論文をデータとして導出されるトピックを「論文トピック」，新聞記事をデータとして導出されるトピックを「新聞トピック」と呼び，これらを区別する．

3.2　関心の距離の測定

学術的な関心と社会的な関心はトピックモデルで導出されるトピックとして表されるため，これらの関心の距離はトピックの距離である．ここで任意のトピックは確率分布 ϕ_{kv} で特徴づけられることから，二つのトピックの距離とは確率分布の距離である．確率分布の距離を測定する手法として，本研究ではジェンセン・シャノン情報量 (Jensen-Shannon divergence)を用いる．任意の二つの確率分布 Q_1, Q_2 に関する情報量 D_{JS} は次式で表される．ただし，D_{KL} はカルバック・ライブラーの情報量である．

$$D_{JS}(Q_1 \| Q_2) = \frac{1}{2}D_{KL}(Q_1 \| Q) + \frac{1}{2}D_{KL}(Q_2 \| Q)$$
$$Q = \frac{1}{2}(Q_1 + Q_2) \tag{4}$$

すると，任意の論文トピック k，新聞トピック k' の距離は，$D_{JS}(\phi_k \| \phi_{k'})$ で表される．これらのトピッ

クの生起確率は，それぞれ次式のように求めることができる．ただし，D_1, D_2 はそれぞれ論文と新聞記事の集合である．

$$\lambda_{1k} = \sum_{d \in D_1} \frac{\theta_{dk}}{|D_1|} \tag{5}$$

$$\lambda_{2k'} = \sum_{d \in D_2} \frac{\theta_{dk'}}{|D_2|} \tag{6}$$

すると，論文トピックの重心 ϕ_{pap} と新聞トピックの重心 ϕ_{news} はそれぞれ，式(7), (8)で表される．

$$\phi_{pap} = \sum_k \lambda_{1k} \phi_{kv} \tag{7}$$

$$\phi_{news} = \sum_{k'} \lambda_{2k'} \phi_{k'v} \tag{8}$$

これらより，論文と新聞の総合的な距離は，次式で求めることができる．

$$D_{JS}(\phi_{pap} \| \phi_{news}) \tag{9}$$

例えば，避難に関して論文と新聞の距離を求めたいというように，ある特定の領域（キーワード）に関する距離を明らかにしたい場面も考えられる．このとき，その領域に関するトピックの生起確率を推計し，それを式(5), (6)に代用することが考えられる．具体的には，ベイズの公式に基づき，任意の領域 v に関する論文トピック k，新聞トピック k' の生起確率は次式で表される．

$$\bar{\lambda}_{1k|v} = P(k|v) = \frac{P(v|k)P(k)}{\sum_k P(v|k)P(k)} = \frac{\phi_{kv}\lambda_{1k}}{\sum_k \phi_{kv}\lambda_{1k}} \tag{10}$$

$$\bar{\lambda}_{2k'|v} = P(k'|v) = \frac{P(v|k')P(k')}{\sum_k P(v|k)P(k)} = \frac{\phi_{k'v}\lambda_{2k'}}{\sum_k \phi_{kv}\lambda_{2k}} \tag{11}$$

すると，ある特定の領域 v に関する論文トピックの重心 ϕ'_{pap} と新聞トピックの重心 ϕ'_{news} は，それぞれ式(7)，(8)と同様に求めることができる．

4. 分析結果

領域に関する論文と新聞の距離の時間的な推移を表 1 に示す．この表より，自然災害（風水害）や予防（建造物・公共施設）に関する領域の多くは，時間の経過に伴って，距離が大きくなっていることがわかる．また，予防（地域）や復旧・復興に関しては，距離が小さくなっていることが見てとれる．しかし，これの距離の大小が学術的な活動の貢献によるものかは明らかではない．そこで，この点を明らかにするために，距離を被説明変数，論文と新聞の数を説明変数として回帰分析を行った．図 1 は，表 1 に示す領域を回帰分析した結果を示している．図の横軸は新聞，縦軸は論文の数に関する回帰係数に−1 を乗じた値である．したがって，上方に位置する領域であれば，それに関連する学術的な活動が多い（論文が多い）ほど距離は短くなることを表しており，学術的な活動が社会の関心を後追うことで貢献している．一方，右方に位置するほど，それに関連する学術的な活動は社会の関心が高まる（新聞記事が多い）ほど距離が短くなることを表しており，学術的な活動が社会の関心を先取りすることで貢献している．このように，その領域に関する学術的な活動の位置づけを把握することができる．加えて，下方，左方に位置する活動は，社会との関心という観点においては低調であることがわかる．具体的には，シミュレーション，地下，警報に関する活動は先取り，高齢，交通，水道に関する活動は後追い的な貢献が認められるものの，洪水や豪雪は必ずしも良好な状態にはない．

表 1　領域に関する距離の時間的な推移

区分	領域	推移	区分	領域	推移
自然現象	地震	−	自然現象（風水害）	豪雨	+
	津波	−		豪雪	+
	火山	+		洪水	+
予防（建造物・公共施設）	学校	−	予防（地域）	教育	−
	耐震	+		自主	−
	堤防	+		計画	−
	地下	+		シミュレーション	−
被災（二次災害）	火災	−		想定	−
	原子力	−	応急（避難）	避難	−
	倒壊	−		高齢	−
	電力	+		救助	+
	水道	−		交通	−
応急（情報）	警報	−	復旧・復興	ボランティア	−
	速報	−		再建	−
	電話	−		心	−
	ネットワーク	−		住宅	−
	インターネット	+			

※推移が+(−)とは，時間の経過に伴って距離が長く（短く）なることを表す．

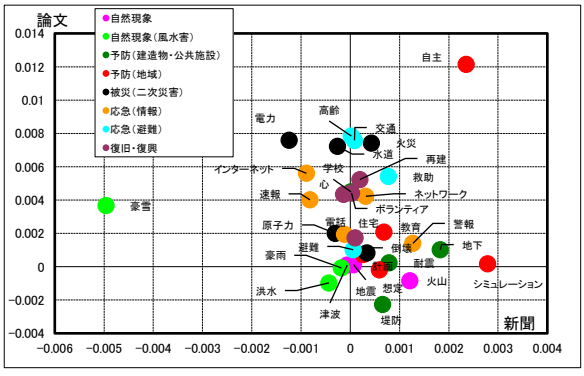

図 1　社会の関心に対する学術的活動の貢献