O_020

Hit数とSimpson係数を用いた47都道府県の解析

Analysis using Hits and Simpson coefficient in 47 prefectures

吉村真弥
Shinya Yoshimura

1．はじめに

World Wide Web（WWW）には膨大なデータが蓄積されており、それらは常に変化している。検索エンジンはWWWのコンテンツを全てではないが、概ね全体を包括的に網羅しているため、検索エンジンの検索結果から全体を知ることはできるはずである。近年、WWWのリンク構造や、友人関係などを解析する複雑ネットワーク解析の研究が注目されているが、巨大で複雑なネットワークの場合は計算量が膨大となる。

本報告では、大まかではあってもネットワークの関係を知ることができないかと考え、特定のキーワードが任意のコンテンツに含まれる量、すなわちヒット数で大まかな関係を知ることを目的とする。さらにヒット数から関係の強さを示すシンプソン係数[1]を用いることで全体の関係を示し考察する。

2．ヒット数と人口の関係

本報告では最初に特定のキーワードとヒット数の関係を調査する。調査方法としてGoogle、Yahoo、gooの各検索エンジンに対し、特定の意味を持ったキーワードで検索を実行し、各検索エンジンでのキーワードのヒット数の平均値を取り、統計情報として人口[2]と比較検証を行った。

最初にキーワードとして日本全国47都道府県で実行した結果を表1に示す。簡単のため東京のヒット数、人口を100%として表示している。47都道府県のヒット数で高いのは東京(100%)、北海道と大阪(45%)、京都(33%)、神奈川(32%)、沖縄(28%)と続く。最も少ないのは島根(8%)である。人口では多い順に東京(100%)、大阪(70%)、神奈川(70%)、愛知(58%)であり、最も少ない県は鳥取(5%)となる。

ここでヒット数と人口の差の絶対値を取ると平均で9%程度となり、概ね相関が取れていることがわかる。例えば北海道、青森では人口とヒット数の割合が東京とほぼ等しい。逆に大都市近郊の都道府県ではブレが大きく、千葉(28%)、埼玉(41%)、神奈川(38%)と大きな差が生じている。これらは愛知(44%)、大阪(25%)、福岡(17%)などの他の主要都市でも見られる現象である。さらに絶対値を取らない場合は値がマイナスとなることから、人口ほどには地域のコンテンツが少ないことを指し、WWWでは人口ほどに地域の話題になっていない（他に関心がある）地域と言えるだろう。逆に絶対値無しの差がプラスの地域も存在しており、沖縄(17%)、京都(12%)、和歌山(10%)、鹿児島(8%)と話題になりやすい観光地であり、WWWには、その地域以外でも話題になりやすい地域が存在していることがわかる。

なお東京を100%とした人口(%)とヒット数(%)の差を図1に示す。基準値の東京を除き、話題になるような地域はプラスになり、人口が多いときヒット数との差がマイナスになり、差も大きくなっていく。これらは人口が高くなれ

	ヒット数(A)	人口(B)	|(A)-(B)|	(A)-(B)
北海道	45%	45%	0%	0%
青森	11%	11%	0%	0%
岩手	9%	11%	2%	-2%
宮城	10%	19%	8%	-8%
秋田	12%	9%	3%	3%
山形	12%	10%	2%	2%
福島	13%	17%	4%	-4%
茨城	10%	24%	13%	-13%
栃木	9%	16%	7%	-7%
群馬	10%	16%	6%	-6%
埼玉	15%	56%	41%	-41%
千葉	20%	48%	28%	-28%
東京	100%	100%	0%	0%
神奈川	32%	70%	38%	-38%
新潟	15%	19%	4%	-4%
富山	11%	9%	2%	2%
石川	14%	9%	4%	4%
福井	11%	7%	4%	4%
山梨	9%	7%	2%	2%
長野	13%	17%	5%	-5%
岐阜	11%	17%	6%	-6%
静岡	13%	30%	17%	-17%
愛知	14%	58%	44%	-44%
三重	12%	15%	3%	-3%
滋賀	10%	11%	1%	-1%
京都	33%	21%	12%	12%
大阪	45%	70%	25%	-25%
兵庫	10%	44%	34%	-34%
奈良	13%	11%	2%	2%
和歌山	18%	8%	10%	10%
鳥取	9%	5%	4%	4%
島根	8%	6%	2%	2%
岡山	13%	16%	2%	-2%
広島	18%	23%	5%	-5%
山口	17%	12%	6%	6%
徳島	10%	6%	4%	4%
香川	9%	8%	1%	1%
愛媛	9%	12%	3%	-3%
高知	10%	6%	4%	4%
福岡	24%	40%	17%	-17%
佐賀	9%	7%	3%	3%
長崎	13%	12%	1%	1%
熊本	12%	15%	3%	-3%
大分	13%	10%	4%	4%
宮崎	13%	9%	4%	4%
鹿児島	22%	14%	8%	8%
沖縄	28%	11%	17%	17%
		avg	9%	

表1　47都道府県の人口とヒット数(2006.4)

ば主要都市の地域性が減少し、話題がある場合は、地域性が高まっていくことを示していると考えられる。

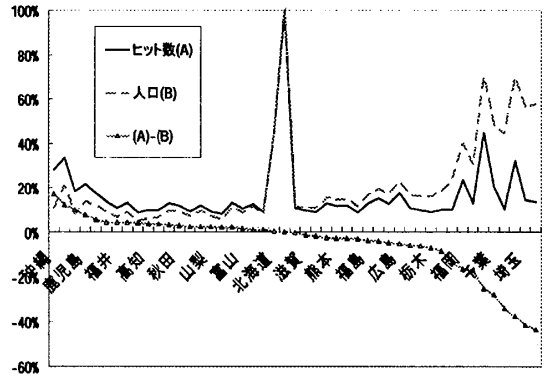

図1. ヒット数と人口の差

3. Simpson係数による地域相関

Simpson係数はWWWの関係を示す指標の一つである。計算方法は、2つのキーワードの∩集合を2つのキーワードの小さい方のヒット数で割ることで求められる。そこで各都道府県のネットワークに対してSimpson係数を求め、視覚化のため、値が0.8以上のネットワークを図示する。

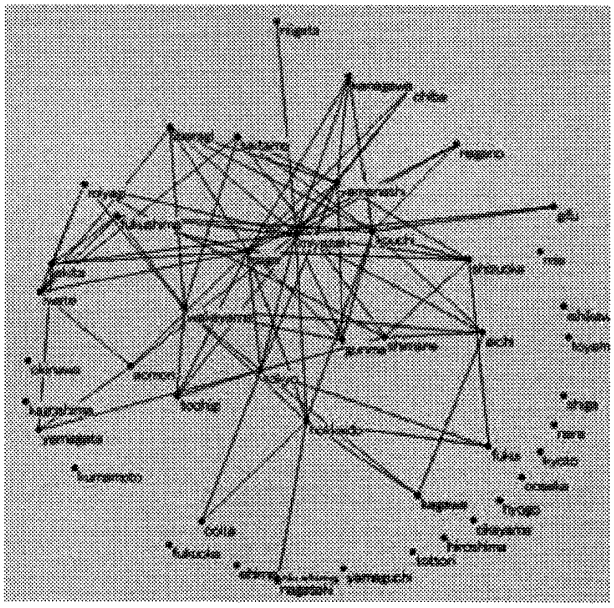

図2. Simpson係数による都道府県ネットワーク図

Simpson係数ではヒット数の少ないものでも全体の中で占める割合が高ければ係数が高くなる。したがってこの中でハブを示すものは意外なことにMiyazaki（宮崎県）である。宮崎県は人口の比べて他の県と繋がるコンテンツを多く持っていると考えられる。この他、gunma(群馬)、kanagawa(神奈川)、kouchi(高知)等、大都市周辺部分にこのようなハブ化する傾向あると言えるだろう。

また、小さな都道府県は東京のような大都市と直接繋がろうとしているのではなく、その周辺の都市（例えば神奈川）と繋がろうとしているように見える。距離についても東北なら東北、九州なら九州と大きな区分けの中では平均して係数が高い傾向があり、距離の近いものと仲良くなりやすいという傾向は明らかである。またヒット数では人口以上に話題のあるという特徴のあった沖縄や京都のSimpson係数は高くない。つまり他県と自らがくっつこうとする力が弱いことがわかる。これらは、特定の都道府県だけと特別に繋がろうとしない。図1と合わせて考えると右側の人口とヒット数のブレの大きいところが、遠隔の都道府県と繋がっている傾向があり、逆に左側では高い係数は見られない。まとめると表2、3のようになる。このようにWWWから都道府県を見ることで新しい洞察を得ることができる。

他県への対応	ヒット数と人口比率	Sympson係数
受動的	ヒット数＞人口の差	小さい
普通	ヒット数＝人口	普通
積極的	ヒット数＜人口	大きい

表2　Sympson係数とヒット数の関係

他県への対応	情報公開性	接続性
受動的	多い	周辺地域が主
普通	普通	なし
積極的	少ない	遠隔接続もある

表3　Sympson係数ネットワークのまとめ

4. まとめと今後

本報告では都道府県別のヒット数と人口の関係を検証し、そしてヒット数からシンプソン係数を用いて関係を簡潔に描いて考察した。このように簡単な実験でも把握できることは多く、大まかな関係を把握する手法の応用は広い。評判などが群集心理を反映する以上、WWWにその傾向は現れるはずならば、マーケティング、また金融や株取引ならば、行動ファイナンスの分野への応用が期待できる。したがって研究をよりよくするために動的なプロセス必要であり、現在、様々な時系列データを計測中である。

5. 参考文献

[1] 安田雪, 松尾豊, 人工知能学会における研究者ネットワークの分析, 2A3-02, 人工知能学会全国大会, 2005年

[2]統計局, 平成17年国勢調査全国･都道府県･市区町村別人口, 総務省,2005年