Epidemiological Modeling of Knowledge Propagation Represented by Scientific Publications

Daniel Moritz Marutschke

(Degree)
博士（学術）

(Date of Degree)
2014-03-25

(Date of Publication)
2015-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第6155号

(URL)
https://hdl.handle.net/20.500.14094/D1006155

Epidemiological Modeling of Knowledge Propagation Represented by Scientific Publications

氏名：Marutschke, Daniel Moritz
専攻：情報コミュニケーション
指導教員氏名：村尾　元

論文要旨

This doctoral thesis investigates the propagation of complex knowledge---such as scientific research methodologies, algorithms, etc.---represented by the number of scientific paper publications each year. As investigative tools, deterministic differential equation models from epidemiological fields are used. Epidemiology is a well founded field of researching the spread of diseases and has been successfully applied to tracking epidemics, endemics, and pandemics.

Epidemiological models using differential equations have matured over the last half century in fields such as medicine and biology. Viruses, parasites, and various kinds of contagion agents in a variety of communities, cross-community, vaccination adapted, delayed vaccination campaign, and many more attributes have been research in-depth. Although the adoption of tracking information diffusion with models from epidemiology dates back to the 1960s, the propagation of more complex knowledge is still insufficiently explored. One of the key challenges is the human factor, as information is transferred from one individual to another with a highly complex set of properties. Information propagation---such as rumors, expansion of economic fields, opinion, email messages, twitter news, and the like---has been successfully modeled using compartmental models from medical and biological epidemiology. The compartmental models that have proven to be able to track information diffusion are the SIR (Susceptible, Infective, Recovered) and the SEIR (Susceptible, Exposed, Infective, Recovered) models and slight variations thereof. In this paper, the author starts with modeling the number of topic-related keyword propagation in scientific publications using the generic SEIR model. It is apparent that the transfer of knowledge has unique characteristics that cannot be explained by generic epidemiology. Several causally extended models are proposed and tested for their performance to better represent the propagation of knowledge.

The propagation of knowledge represented by scientific publications has been carried out using four data sets. The first data set with 138,303 papers was gathered on the IEEE Xplore Digital Library. Information was accumulated about the paper title, abstract, authors, affiliations, year of publication, and citation count. This allowed for a detailed exploratory analysis of propagation attributes and additional room for refining any algorithms to track the growth of paper numbers.

Three further data sets were acquired using the help of online search engines inside major scientific publication databases. This allowed for extended data-points and some other gathering of semantic information such as cross-referencing the author's affiliation. Another advantage was the ability to effectively conduct full text searches which was previously prohibited due to data size as well as access restrictions. One of the data sets was gathered from CiNii, a search engine for academic information and articles in Japanese with more than 15 million articles. Another data source was CNKI.NET, China's largest online database of scholarly articles with a total of over 47 million articles. The third source was from Scirus, a database of more than 60 million English publications worldwide.

From a set of prominent science and engineering textbooks, 39 initial keywords---extended to 88 with the Scirus database---from information science and engineering are used to examine the suitability of this approach. Many of the authors' published papers, however, use a set of five keywords from Soft Computing to ensure comparability of the studies. The author will show that categorization of keywords into three source databases is possible using SEIR model parameters and discuss the limits and further exploration of epidemiological models. A modified SEIRE model is proposed to more robustly track knowledge propagation in scientific publications. Two more models, an SEIRK and an SEIREK model have also been proposed and were compared by performance, causality, and complexity.

As a classification and knowledge discovery process, all keywords were categorized using parameter-wise k-means clustering. The classification has proven to be difficult due to the limited number of data points, resulting in a limited number of keywords with sufficient accuracy. This knowledge discovery process could help in determining relationships between topics that have an evolutionary connection.

As suggested by previous research, diffusion of information is also influenced by cultural features. Using a classification process with k-means to be able to assign a propagation feature to a specific culture, the author analyzed five keywords from Soft Computing in the three data sets---CiNii, CNKI.NET, and Scirus.

As indicated by these three data sets---CiNii, CNKI.NET, and Scirus---culture has an influence on knowledge distribution. Qualitative research of ICT adoption in cultural settings is an active field. Quantitative measures of knowledge propagation connected to cultural settings, however, are still infrequent. From the Scirus database, a new data set was built with 32 country-affiliation, each with 22 fields of knowledge. Using Principal Component Analysis, further trends of culture and knowledge was investigated.

The Scirus database was also used to compare the SEIR and SEIRE model with regards to their coefficient of determination (adjusted R^2), their Basic Reproductive Rate (R_0), as well as the residuals distribution of the model fits.

［課程博士用］

論文審査の結果の要旨

氏　名	Marutschke Daniel Moritz（マルチュケ・モリツ）		
論文題目	Epidemiological Modeling of Knowledge Propagation Represented by Scientific Publications		
判　定	合　格 ・ 不合格		
審査委員	区　分	職　名	氏　名
	委員長	教授	森下　淳也
	委　員	教授	KRYSSANOV, Victor（立命館大学情報理工学部）
	委　員	教授	村尾　元
	委　員		
	委　員		

要　旨

本論文は，知識の伝搬をモデル化する新しい数理モデルを提案するものである．ここでは特に工学分野における論文のキーワードの伝搬と拡散の様子のモデル化を試みている．これにより，知識伝搬の特徴を明らかにするとともに，知識伝搬の類型分類を試みている．

本論文では，この目的のために感染症の数理モデルを利用する．感染症数理モデルは感受性保持者(S)と感染者(I)，回復者(R)もしくはこれらに加えて潜伏期間の者(E)のポピュレーションダイナミクスを微分方程式の形で記述したものである．これらはSIRモデルやSEIRモデルと呼ばれ，感染症患者の増減といった振る舞いを分析し，対策を検討するために利用されている．

感染症数理モデルは感染症以外のモデル化にも利用されており，これまでにも物理学におけるアイデアの伝搬や，経済学における知識の伝搬，ツイッターなどネットワーク上のメッセージの伝搬や，噂の伝搬などの分析への感染症数理モデルの適用が試みられている．

本論文は，これまでの感染症数理モデルの適用に関する非常に多くの文献調査を行い，整理を試みている．その上で，それらで提案されたモデルが工学分野における論文のキーワードの伝搬と拡散のモデル化に利用できないことを示し，新しい数理モデルを提案している．

本論文で提案するのはSEIREモデルである．ここでSは論文のアイデアを保持している者，Eはアイデアに基づいて実験を行うなどし，論文を書く者，Iは論文を発表する者，Rは論文発表後にアイデアを暖めたりする者である．一般的な認識としては当然だと思われるが，数理モデルとしては，論文をいったん書いた者(R)が，論文を再び書く者(E)に変わるという点が新しいという．

SEIREモデルの妥当性を示すため，実際のデータを用いてフィッティングしている．ここでは，IEEE Xplore Digital Libraryから英語の論文約14万点，CiNiiから日本語の論文1,500万点，CNKI.NETから中国語の論文4,700万点，さらにScirusから英語の論文6,000万点のそれぞれについて，論文題目や要約などに含まれるキーワードを収集している．含まれる単一のキーワードについて，その個数の年変化に一致するようにSEIREモデルのパラメータを決定する．

ソフトコンピューティングに関するキーワードについてフィッティングを行い，得られたパラメータをK平均法でクラスタ化した結果，２つの特徴有るグループが見つかったという．１つは「ファジー論理」や「主成分分析」などのグループで，もう１つは「ニューラルネットワーク」や「遺伝的アルゴリズム」などのグループである．これらのグループ間には，論文数の減少が始まっているか，依然として多くの論文が発表されているかという違いが見られるという．

また，世界中の32カ国について，各国で発表された論文数について，SEIREモデルでフィッティングを行い，その結果得られたパラメータを分類している．その結果，欧州の国々は記号や発見的手法に関する論文が多く，日本を除くアジア各国はソフトコンピューティング，日本はブラジルやカナダに近く分析系の論文が多いという知見が得られたという．

本論文は７章から構成されている．第１章は研究の目的と背景，論文の構成について書かれている．

第２章は感染症数理モデルについて，第３章は感染症数理モデルを用いた知識伝搬のモデル化について，第４章は文化の伝搬について，十分な量の既発行の文献を参照しながら，関連研究の紹介と整理をし，本研究の位置づけについて説明を行っている．特に第２章と第３章では，非常に多くの文献に基づいて感染症数理モデルに関する研究を整理している．感染症数理モデルの関連研究について，これだけの文献を整理した例は少ない．

第５章は実験で利用するデータが紹介されており，第６章で数理モデルの紹介と実験データを用いたフィティングの結果，考察が行われている．第７章は結論である．

論文全体を通して，感染症数理モデルを用いた知識伝搬のモデル化という本研究の学術的な位置と新規性（第１章～第４章）とその有用性（第６章）が丁寧かつ明確に説明されている．

本論文の内容の一部は３件の国際論文誌に掲載されている．また，４件の査読付き国際会議での口頭発表が行われており，４件の国内会議での口頭発表とあわせて計１１件の発表が行われている．全て情報分野に関する学術論文及び学術会議である．

本論文は，工学分野の論文に含まれるキーワードの伝搬と拡散のモデル化の試みについて著述されたものである．本研究は情報分野において学術的に価値があるのみならず，国による論文テーマの違いという題材にも踏み込んでおり，社会学や文化学など学際的な学術分野にも少なからぬ貢献を行うものである．

以上のような理由から，本審査委員会は，マルチュケ・モリツ氏に博士（学術）の学位を得る資格があると判定する．