様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 5 月 23 日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：19700147
研究課題名（和文）　情報学的アプローチによる仮説発見：大規模文献解析に基づく遺伝病原因遺伝子の推定

研究課題名（英文）　Informatic approach to hypothesis discovery: Finding implicit gene-disease associations

研究代表者
関　和広（SEKI KAZUHIRO）
神戸大学・自然科学系先端融合研究環重点研究部・助教
研究者番号：30444566

研究成果の概要：

本研究では，疾病と遺伝子の関係を遺伝子機能と表現型を介してモデル化し，文献解析による原因遺伝子の予測を試みた．その結果，従来一般的であった抄録だけを使った場合と比較し，全文データを用いた場合は 5%程度の予測性能の向上が見られた．また，オントロジーに記述された概念間の関係を利用して確率パラメタを伝播したとき，システム性能の向上が見られた．代表的な先行研究と比較した場合も，提案手法の予測性能が最大で 20%程度高いことが分かった．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	2,600,000	0	2,600,000
2008 年度	700,000	210,000	910,000
年度			
年度			
年度			
総　計	3,300,000	210,000	3,510,000

研究分野：知能情報処理
科研費の分科・細目：情報学・知能情報学
キーワード：テキストマイニング，仮説生成

１．研究開始当初の背景

近年のコンピュータ関連技術の進歩に後押しされ，利用可能なテキスト情報が増大している．それにつれ，大量のテキストに埋没する情報を効果的に利用することが一層難しくなってきている．例えば生命科学の分野を見ると，Medline（当該分野最大の書誌情報データベース）に収録されている総論文数は 1700 万件以上に達し，一日に 1500～3000 件の新しい論文が追加され続けている．これだけの規模の情報を個人が処理するためには，情報検索や情報抽出，仮説発見といった知的な情報処理技術が重要になってくる．

情報検索や情報抽出が文書中に明示的に記述された「既知の情報」を対象とした処理であるのに対し，仮説発見は，文書を自動的に解析することで，これまで知られていない「未知の情報」を発見することを目的とする．仮説発見の先駆的な研究は今から 20 年前，シカゴ大学の Swanson によって行われている．

Swanson は論理的に関連している知識がその関連に誰も気付かないままに孤立して存在していると考え，この関係の同定が新たな知識となり得る仮説の発見につながると主張した．以降，Swanson を含めたいくつかの研究グループが，この仮説発見のプロセスを手続き的に整理し，これを支援するためのコンピュータプログラムを開発した．しかしながら，これらの研究で提案された手法は大部分においてヒューリスティックであり，未だ一般的なモデル・手法は確立していない．

２．研究の目的

上述のような背景から，本研究では，情報学的アプローチによる知識発見について研究を進めている．具体的には，情報検索モデルのひとつである推論ネットワークを用い，病気 d と遺伝子 g の関係を遺伝子機能 f と表現型 p を介してモデル化し，遺伝病の原因遺伝子の予測を試みる．その上で，Gene Ontology，MeSH といったオントロジー（あるいはシソーラス）の意味的階層構造の効果的な利用法を探る．また，提案モデルの有効性を相対的および客観的に評価するため，先行研究や中間ノードを一層しか持たない他の検索モデルとの比較実験を行う．さらに，大規模な全文データを用いて原因遺伝子の予測を行い，テキストマイニングにおける全文データの有用性を実験的に評価する．

３．研究の方法

本研究では，特に以下の点に焦点を当てて研究を進めた．

(1) データ過疎問題に対するオントロジーの利用．概念（遺伝子機能 f と表現型 p）間の関係を推定する際，学習に利用できるデータが限られていることからデータ過疎の問題が生じる．これを解決するため，オントロジーの意味的階層関係を利用する手法を提案・評価する．

(2) テキストからの知識発見における全文データの有用性．全文データとは，論文のタイトルから結論，参考文献および図表のキャプションまで全てのテキスト情報を含むデータを指す．本研究では，全文データが知識発見に有効であるかどうかを種々の実験によって調査する．

(3) 提案手法の相対的評価．これまでの類似研究が定性的な評価を中心としていたのと対照的に，本研究では，実データに基づくベンチマークを用い，提案手法の有用性・妥当性を先行研究や他のモデルとの比較によって定量的に評価する．これにより，情報検索モデルを用いた知識発見の枠組みの確立を目指す．

４．研究成果

(1) 研究着手 1 年目は，未知の遺伝的関連予測（仮説発見）の枠組みの中で，特に遺伝子機能と表現型（phenotype）の因果関係推定における「遺伝子機能オントロジーおよび表現型シソーラスの有用性」，「全文データの有用性」を調査した．前者について具体的には，オントロジーあるいはシソーラス中の概念の階層関係に注目し，訓練事例から推定された因果関係を隣接概念に伝播することで，より頑健なパラメタ推定を試みた．実データによる評価実験を行ったところ，遺伝子機能オントロジーのみ，かつ上位から下位への関係のみを用いたときに，仮説発見の性能向上が見られた．この結果は，上位概念の性質（具体的には遺伝子機能が引き起こす表現型）が下位概念に継承されるということを意味する．パラメタ推定に利用できる事例は限られているため，本手法によって，既知の事例からは獲得できないような因果関係に対しても，より信頼性の高い推定が可能になった．

後者，「全文データの有用性」については，遺伝子機能と表現型の因果関係推定に際して，論文タイトルと要約に加えて全文データを用いた場合，遺伝的関連の予測性能が向上するかを調査した．要約と比較し，全文データは実験等に関してより完全な情報を提供するため，本研究のような学術文献解析に基づく知識システムでは，全文データを利用することでシステムの性能向上が期待できる．しかし，著作権の関係などから，これまで仮説生成における全文データの有用性を定量的に調査した例はない．評価実験の結果，小規模な実験ながら，全文データを用いた場合には全体で 5.1%の性能向上が見られ，その潜在的な有用性が示された．

(2) 研究 2 年目は，提案した未知の遺伝的関連予測（仮説発見）の枠組みと先行研究との比較に焦点を当て，提案手法の有効性を評価した．比較対象としては，類似の著名な先行研究として，Freudenberg と Propping（2002）およ

び Perez-Iratxeta ら（2002，2005）の手法を用いた．前者については，Freudenberg と Propping が用いた Online Mendelian Inheritance in Man（OMIM）のデータを利用し，同様の条件で 878 の遺伝病の原因遺伝子の予測を行った．その結果，我々の提案手法はより多くの遺伝病について，その原因遺伝子をより正確に予測できることが分かった．より具体的には，所与の遺伝病について，human の全遺伝子をその遺伝病の原因遺伝子としての尤度に基づいて順位づけたところ，Freudenberg らの手法では上位 3%以内に順位づけられる真の原因遺伝子が 33%であったのに対し，我々の手法では 40%へと向上した．

また，Perez-Iratxeta らの手法に関して類似の比較実験を行ったところ，同様に，より多くの遺伝病について原因遺伝子のより正確な予測が可能であることが分かった．具体的には，Perez-Iratxeta らの手法では上位 8 位以上に順位づけられる真の原因遺伝子が 47，上位 30 位以上に順位づけられる真の原因遺伝子が 62 であったのに対し，我々の手法では 49，75 の真の原因遺伝子がそれぞれ上位 8 位，30 位に順位づけられた．これらの結果から，類似の研究に対する我々の提案手法の優位性・有用性が示された．

以上の成果から，このモデルを応用することにより，遺伝的関連研究（genetic association study）に要する時間を短縮することが可能だと考えられる．これは，遺伝病のメカニズムの解明や疾病の予防，予測，治療等の促進につながる．また，将来的に知識発見の提案モデルを他の対象にも一般化できれば，テキストマイニング研究のさらなる発展・活性化につながるものと期待される．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 4 件）

①. Kazuhiro Seki and Javed Mostafa. Discovering Implicit Associations among Critical Biological Entities. International Journal of Data Mining and Bioinformatics, Vol. 3, No. 2, pp. 105-123, 2009. 査読有り
②. Kazuhiro Seki and Javed Mostafa. Gene Ontology Annotation as Text Categorization: An Empirical Study. Information Processing & Management, Vol. 44, No. 5, pp. 1754-1770, 2008. 査読有り
③. 関和広，モスタファジャビド．多様な遺伝子名認識と文書分類を用いたGene Ontologyアノテーション．電子情報通信学会論文誌, Vol. J91-D, No. 04, pp. 1033-1041, April 2008. 査読有り
④. Kazuhiro Seki and Javed Mostafa. Literature-Based Discovery by an Enhanced Information Retrieval Model. In Proceedings of the 10th International Conference on Discovery Science (DS 2007), pp. 185-196. October 2007. 査読有り

〔学会発表〕（計 3 件）

①. 宮西大樹，関和広，上原邦昭．生物医学文献からの知識抽出とイベント間のつながりを考慮した発見性を伴う仮説の提示. 第 1 回データ工学と情報マネジメントに関するフォーラム／第 7 回日本データベース学会年次大会．2009 年 3 月 8～10 日．静岡．
②. 木野嘉祐，関和広，上原邦昭．相同分子種を考慮した遺伝子機能アノテーションへの多階層分類の適用. 情報処理学会研究報告 2008-MPS-72, 2008 年 12 月 17 日．大阪．
③. 木野嘉祐，関和広，上原邦昭．相同分子種を利用した動的多階層分類による遺伝子機能アノテーション．iDBフォーラム 2008， pp. 205-210．2008 年 9 月 21～23 日．福島．

〔図書〕（計 2 件）

①. Kazuhiro Seki, Javed Mostafa, Kuniaki Uehara. Finding Explicit and Implicit Knowledge: Biomedical Text Data Mining. In Leon S.L. Wang and Tzung-Pei Hong (Eds.), Intelligent Soft Computation and Evolving Data Mining: Integrating Advanced Technology. IGI Global.（採録決定）
②. Javed Mostafa, Kazuhiro Seki, and Weimao Ke. Beyond Information Retrieval: Literature Mining for Biomedical Knowledge Discovery. In Jake Chen, Stefano Lonardi, and Randi Cohen (Eds.), Biological Data Mining. Chapman & Hall/CRC Press.（採録決定）

６．研究組織

(1)研究代表者

関 和広（SEKI KAZUHIRO）
神戸大学・自然科学系先端融合研究環重点研究部・助教
研究者番号：30444566

(2)研究分担者

なし

(3)連携研究者

なし