1C5-1

ウェブを用いたサイエンスマップのためのクエリ拡張の研究

Query expansion method for creating Science Map using search engine

関 喜史*1　　森 純一郎*2
Yoshifumi Seki　　Junichiro Mori

東京大学
University of Tokyo

Science Map aim to a bird's eye view of the research area by analyzing the citation network of papers. Several studies have reported how analyze citation network, but little is known about the method how collect research papers. A number of studies has collected papers from typical keyword in the research area, but it's difficult for new research area to identify typical keyword. In this paper, we provide query expansion method identifying typical keyword in new research area for creating Science Map. In result, Our method succeed createing science map in new reseach area.

1. 背景と目的

学術論文を分析することで対象とする学術分野における知識を定量化・可視化しようという試みは書誌計量学とよばれ古くから様々な研究が行われている．そして複雑ネットワーク研究の発達，電子ジャーナルデータベースの整備，大規模な計算機資源が容易に利用可能になったことなどを背景にして，近年大量の論文集合における論文間の引用関係を元にネットワークを構築し分析する研究が盛んに行われており，学術領域全体の俯瞰 [伊神 09, Small 06, Leydesdorff 09] や特定学術領域の構成要素を明らかにする試み [Kajikawa 07, Porter 09] が行われている．本研究ではこのような引用ネットワーク分析による学術領域の俯瞰の試みをサイエンスマップと呼ぶことにする．

これらの研究は引用ネットワークの解析の結果によって学術領域に対する評価を客観的に行うことが可能になることから，国家の科学技術政策や，企業の科学技術戦略に関する意思決への活用が期待されている．

引用ネットワークを構築するためにはまず構築対象となる論文集合を構築する必要がある．学術領域全体の俯瞰を行うためには被引用数が高い論文を抽出する方法や [Small 06, 伊神 09]，代表的な論文誌から収集する方法 [Leydesdorff 09] が過去行われている．特定分野に対する分析の際はその分野の代表的な論文誌から収集する方法 [Porter 09]，分野名の分野における代表的な学術用語をクエリとして Web of Science 等の電子ジャーナルデータベースから論文を獲得する方法 [Kajikawa 07] がとられている．この手法は既に確立している領域には有効に働く．しかしまだ学術領域として確立していないような新しい分野についてはもちろん論文誌は存在していない．そして分野名をクエリにしても電子ジャーナルデータベースからは極めて少ない論文しか得ることはできない．そしてそういった分野では専門家の間でも定義や見解が別れることが多く，関連する論文を獲得することが極めて困難である．

連絡先: 関 喜史, 東京大学大学院工学系研究科技術経営戦略学専攻,seki@weblab.t.u-tokyo.ac.jp

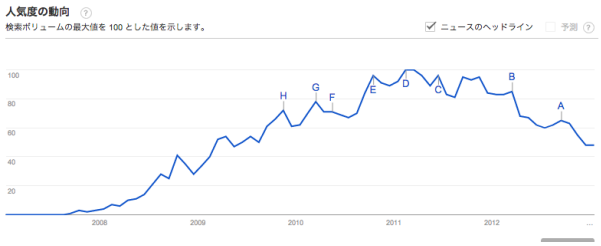

図 1: Cloud Computing の Google Trends における人気度 (http://www.google.co.jp/trends)

図 1,2 に示すのは「クラウドコンピューティング」における Google Trend の人気度と Web of Science から得られる 2011 年までの論文数の推移である．ウェブ上における人気度は 2008 年から高まりはじめているのに対して，2008 年に論文数はほとんどなく，現在でも合わせて 529 件とサイエンスマップを構築するには不十分な量である．

このようなキーワードはバズワードともよばれ批判の対象にもなるが，注目度が高い分野であることは確かでありこれらの領域に対する研究開発の必要性は高い．そして情報が不十分な領域に対してサイエンスマップを構築することが出来れば，科学技術政策や研究開発戦略の意思決定に大いに役立つ．しかしこれまでに提案されている手法では，これらのキーワードに対してサイエンスマップを構築するために有効な量の論文集合を獲得することができないのが実情である．

そこで本研究ではここまでに述べたような新規の領域に対するサイエンスマップを構築することを目的とし，その領域に関連する論文集合を獲得するための方法を提案する．

2. 提案手法のコンセプト

本節では本研究の提案手法のコンセプトを述べる．本研究の課題を情報検索におけるクエリ拡張の問題と捉えると，「ク

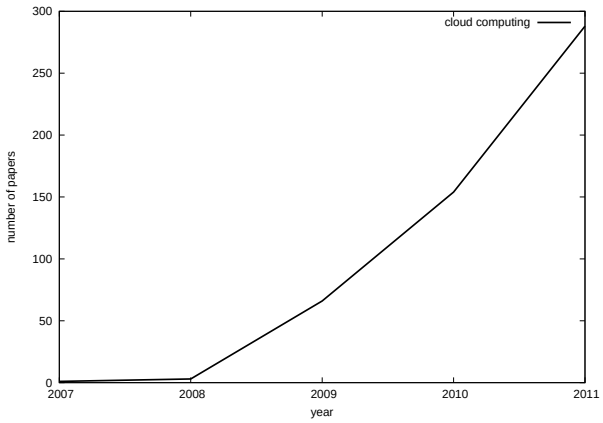

図 2: Cloud Computing の論文数の推移

ラウドコンピューティング」等をクエリとした時に有効な検索結果を得ることが出来ないため，代替クエリを構築するという問題と捉えることが出来る．これらの対象領域を表すキーワードを本研究では「シードワード」と呼ぶこととする．

本研究では新規の研究領域の論文集合を獲得するために，ウェブのデータを用いてシードワードの代替クエリを構築し，論文集合を獲得する手法を提案する．その手法は 2 段階にに分けられる．ひとつはウェブを用いてシードワードの代替クエリを構築する手法であり，3. 章で述べる．そしてその代替クエリから論文集合を獲得する手法を 4. 章で述べる．

3. ウェブを用いた代替クエリの獲得

本研究の目的は学術論文を検索するための代替クエリを構築することである．代替クエリを構築するためにはシードワードと対象語との学術領域における類似性を評価しなくてはならない．

まず学術用語辞書を用いてウェブから集める語を限定する．学術用語辞書は Web of Science にある論文の著者キーワードの内 150 件以上の論文に紐づくキーワードを用いて構築した．

代替クエリを獲得する際シードワードを用いて検索エンジンで検索し，その検索結果として得た URL の HTML ファイルから学術用語辞書を用いて学術語を得ることで，シードワードと関係する学術用語を獲得する．そして検索エンジンの共起検索件数を用いて，獲得した語とシードワードとの類似性を評価する．検索エンジンの共起件数を用いた類似性評価には Danushka らが提案している $WebPMI$ を用いた [Bollegala 07]．

検索エンジンを用いる学術用語の獲得と類似性の評価では，一般的な類似性の評価しか得られない．本研究で得たい類似性はあくまえで学術領域における語の類似性である．そこで本研究では検索クエリに対してクエリを付加し検索結果を絞り込むことで，学術領域における類似性を評価することを試みる．

今回対象としているようなシードワードはビジネス領域側で作られたキーワードであることが多く，ウェブのデータとしてもビジネスに関するものが多い．そのためまず $-business$ というクエリを付加し，学術領域に絞るために $research$ というクエリを付加，そしてより専門領域に絞るために $computerscience$ というクエリを付加する．これらのクエリを付加した上で，検索エンジンから学術用語を収集し，類似性を評価することとする．

4. 代替クエリを用いた論文集合の獲得

3. 章の手法により類似性をスコアとしてもつ代替クエリ候補を得た．本章ではこの代替クエリ候補から論文集合を獲得する手法について議論する．

代替クエリ候補はシードワードとの類似度スコアがついた学術用語から構成される．本研究ではスコアの高い順に学術用語を代替クエリとして処理し，獲得した論文数が一定数を超えた段階で論文の獲得を打ち切る．引用ネットワーク構築には論文数が多くなればなるほど多くの計算量がかかるため，論文数を一定の範囲に押さえるためである．

ここで大きな学術領域に紐づく代替クエリから論文集合を獲得した場合に，その代替クエリにより構築されるサイエンスマップが大きな影響を受けてしまうことが考えられる．このような単一の代替クエリの影響を軽減するために，複数の代替クエリから得られた論文のみを獲得することにする．

この手法について，2 つ以上の代替クエリから獲得された論文を取得する場合を例に説明する．代替クエリ候補 $Q = q_1, q_2, q_3, ..., q_n$ が与えられたとする．ここで代替クエリのスコア $score(q)$ において $score(q_i) \geq score(q_j)(i \leq j)$ であるとする．まず代替クエリ q_1 を用いて論文集合 P_1 を取得する．次に代替クエリ q_2 を用いて論文集合 P_2 を取得する．ここで $P_1 \cap P_2$ を取得する学術論文集合とする．続いて代替クエリ q_3 を用いて論文集合 P_3 を取得し，$P_1 \cap P_3$ と $P_2 \cap P_3$ を学術論文集合に加える．これを論文数が一定数を超えるまで繰り返すことで学術論文集合を獲得する．同じように 3 つ以上インデックスされている論文を取得する際は $P_1 \cap P_2 \cap P_3$ を獲得していくことを繰り返す．

これ以降いくつの代替クエリから得られたかは重複度 n と表すことにする．このように複数の代替クエリから得られる論文のみを得ることで，分野の偏りを軽減すると共に，ネットワークを密にすることが期待できる．

5. サイエンスマップ構築のフロー

本章ではサイエンスマップの構築フローを示す．まずは以下に全体の流れを示す

1. ウェブからシードワードの代替クエリ集合を獲得する．
2. 代替クエリから論文集合を獲得する．
3. 論文集合を用いて引用ネットワークを構築する．
4. 引用ネットワークの最大連結成分を取り出しクラスタリングを行い，各クラスタの特徴キーワードを算出する．
5. クラスタリングに基づき引用ネットワークの可視化を行う

引用ネットワーク構築には書誌結合とよばれる指標を用いる．書誌結合は他の指標に比べてリンクが形成されやすいた

め，リサーチフロントを特定しやすい性質を持つとされている [Boyack 10].

クラスタリングには引用ネットワーク解析の研究でこれまでにも用いられている Newman 法 [Newman 04] を高速化した手法を用いる [Blondel 08]．クラスタ内の特徴キーワードとしては各クラスタを文書，著者キーワードを語とみなして TF-IDF[Manning 08] による特徴付けを行ったものを用いる．

引用ネットワークの可視化の際は描画における計算量を削減するために重み 1 のエッジを除去し，最大連結成分のみを描画する．描画の際はグラフ可視化ツールの Gephi を用いて，可視化アルゴリズムには Fruchterman-Reingold 法を用いた．

6. 評価実験

本章では提案手法である代替クエリ構築法を検証するための実験について述べる．シードワードとしてはクラウドコンピューティングを用いる．1. 章でも述べたとおりクラウドコンピューティングに関する論文は 2011 年末時点で 529 件存在する．サイエンスマップを構築するには不十分であるが，この論文集合の著者キーワードを正解データとして提案手法の評価を行うこととする．正解データにはクラウドコンピューティングの著者キーワードのうち，2 つ以上の論文に出現するものの扱う．

ここで代替クエリを構築する際の検索エンジンでは，検索エンジンを機能である期間指定検索を用いる．図 2 で示した通り，クラウドコンピューティングが学術領域で普及し始めたのは 2009 年である．そこで検索エンジンの期間を 2008 年 12 月 31 日以前に設定し，構築した代替クエリから得られた引用ネットワークのクラスタリングの結果を，クエリを何も付加しなかったケース (*plain*) と $-business$"$research$""$computerscience$" クエリを付加したケース (*brm*) において比較し比較し，どちらの引用ネットワークのクラスタにおける特徴キーワードが正解データを予測できたかを評価する．論文獲得のための重複度は 2 と 3 を用いる．

評価基準としてはクラスタサイズが大きい上位 10 個のクラスタの TF-IDF における特徴著者キーワード上位 10 件の計 100 語のうち，正解データに含まれるものの割合と，正解データが 2 つ以上含まれたクラスタの数を評価する．

7. 結果

評価結果を表 1 に示す．*plain* のケースと比較しクエリを付加した *brm* のケースのほうが正確性，一致したクラスタ数共に高いことからクエリを付加する事によって適切なクエリ候補を獲得できていると言える．

表 1: cloud computing による手法の評価

base query	similarity type	num	accuracy	matched cluster
plain	pmi	2	0.3	5
plain	pmi	3	0.11	1
brm	pmi	2	0.31	8
brm	pmi	3	0.24	6

クエリを負荷しない *plain* から得られたクラスタの特徴著者キーワードを見ると business process, supply chain management, invasive species といったクラウドコンピューティング研究とは一見関係が薄そうなキーワードと持つクラスタが存在する．これらのキーワードはビジネス面から使われることの多いキーワードであり，クエリを付加しない場合，代替クエリの生成にビジネス面のシードワードに対する類似性が強く働いていることが分かる．

表 1 にあるように重複度 2 のほうが正確性，一致したクラスタ数の両方が重複度 3 の結果を上回っている．しかし重複度 2 よりも重複度 3 のケースのほうが分散処理においてより専門的なクラスタが獲得できていた．クラウドコンピューティングの根幹をなす技術は分散処理であり，データマイニングや情報検索はその応用範囲である．この結果を定性的に判断すると，重複度 3 のケースのほうがよりクラウドコンピューティング研究を反映しているようにも見える．

そのためここでは重複度 2,3 のどちらが優れているという議論は避けることとする．

8. サイエンスマップの可視化

図 3 に可視化の結果を示す．

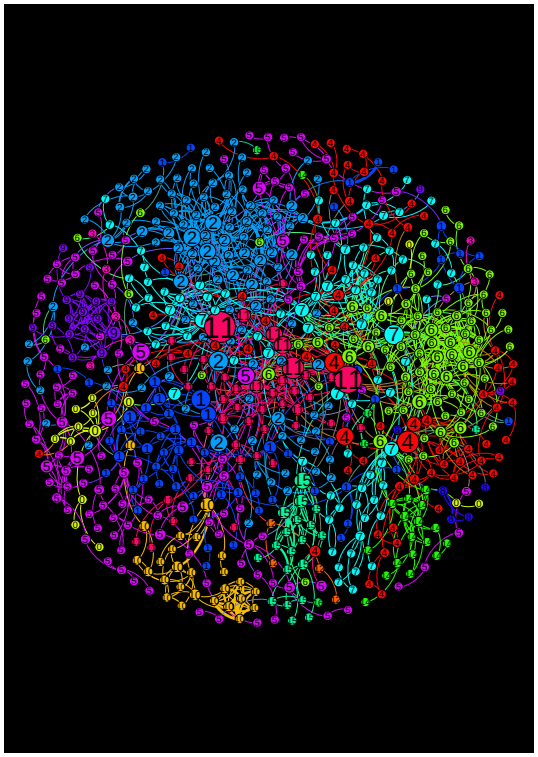

図 3: Cloud Computing のサイエンスマップ

左側にデータマイニング関係の論文が，右側に分散処理に関する論文が集まっている．データマイニングやバイオインフォマティクスに関してはもともとの学術領域が大きい分野であるため，クラスタのサイズとしては大きいものになっているが，クラウドコンピューティング領域から見た場合には応用研究という立ち位置である．そのためかデータマイニング，情報検索

のクラスタはネットワーク図の端の方に位置している．またバイオインフォマティクス分野には媒介中心性が高い論文が複数存在する．これは分散処理領域とデータマイニング領域双方に関わりを持っているためと推測される．

9. まとめ

本研究では新しくまた未確立な学術領域を俯瞰するためのサイエンスマップ構築のために，論文集合を集めるためのクエリ拡張手法について提案した．学術領域のトレンドは一般社会に対して一歩遅れているのが実情であり，そのギャップを埋めるためにウェブのデータを用いて代替クエリを生成するという方法を試みた．そこで検索エンジンからウェブのデータを得る際，一般的な類似性ではなく学術領域における類似性を抽出する必要がある．本研究ではそこでいくつかのクエリを付加することで学術領域における関係性を抽出する方法を提案し，実データを用いた実験を通して，提案するサイエンスマップ構築手法の有用性を示した．

しかし本研究のような領域には正解データセットが存在しないため正確な評価が困難である．今後は専門家などの評価を得ながら，現在の手法における問題点を探し手法の高度化を行なっていく必要がある．

本研究では従来議論されていなかったサイエンスマップ構築における論文収集という課題に対して一定の成果を示すことが出来た．この成果を元にこの問題に注目が集まり，よりよい手法が生み出されるきっかけとなれば幸いである．

参考文献

[Blondel 08] Blondel, V. D., Guilaume, J.-L., Lambiotte, R., and Lefebvre, E.: Fast unfolding of communities in large network, *Journal of Statistical Mechanics: Theory and Experiment* (2008)

[Bollegala 07] Bollegala, D., Matsuo, Y., and Ishizuka, M.: Measuring Semantic Similarity between Words Using Web Search Engine, in *In Proceeding of World Wide Web Conference 2007* (2007)

[Boyack 10] Boyack, K. W. and Klavans, R.: Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?, *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, Vol. 61, (2010)

[Kajikawa 07] Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K., and Komiyama., H.: Creating an academic landscape of sustainability science: an analysis of the citation network, *Sustatinability Science*, Vol. 2, (2007)

[Leydesdorff 09] Leydesdorff, L. and Rafols, I.: A Global Map of Science Based on the ISI Subject Categories, *Journal of the American Society for Information Science and Technology*, Vol. 60, (2009)

[Manning 08] Manning, C. D., Raghavan, P., and Schutze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008)

[Newman 04] Newman, M. E. J.: Fast algorithm for detecting community structure in networks, *Physical Review E* (2004)

[Porter 09] Porter, A. L. and Youtie, J.: How interdisciplinary is nanotechnology?, *Journal of Nanoparticle Research*, Vol. 11, (2009)

[Small 06] Small, H.: Tracking amd predicting growth areas in science, *Scientometric*, Vol. 68, (2006)

[伊神 09] 伊神正貫, 阪彩香：サイエンスマップによる科学研究の動的変化の計測, 情報管理, Vol. 52, (2009)