3E3-4

リスク検知にむけたコミュニティ発見手法のシステム化

Community Mining System for Risk Detection

市瀬 龍太郎[*1] 武田 英明[*1] 村木 太一[*2] 太田 正幸[*3]
Ryutaro Ichise Hideaki Takeda Taichi Muraki Masayuki Ohta

[*1]国立情報学研究所 [*2]トライアックス
National Institute of Informatics TRIAX Inc.

[*3]産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

The research community plays a very important role in keeping scientific knowledge. The authors propose a community mining system that helps to find communities of researchers by using bibliography data. The key feature of our method is a network model of researchers and a word assignment technique for the communities obtained. We implemented proposed method in a visualized system.

1. はじめに

私たちの日々の生活は，さまざまな技術によって支えられている．しかし，技術を使用する人が，その技術に関する知識を深く理解していないと，大きなリスクをもたらすことになる．そのために，科学的な知識をいかにして，技術者の間で共有するかということは，大きな課題となっている．一般的に，科学的な知識は論文などの文書で蓄積され，共有されていくことが多い．しかし，このような知識全てを文書化することは，非常に難しい．本研究では，文書化が難しい知識の一つとして，研究者のコミュニティで暗黙的に伝承される知識に焦点を当て，研究者のコミュニティを発見することで，このような知識の所在場所を同定するシステムを提案する．そして，構築したインタラクティブにコミュニティを発見するためのシステムについて考察を行う．

2. コミュニティの発見

本研究では，文書化されにくい科学的知識を取り出すために，特定の領域の知識を保持していると考えられる研究者のコミュニティを取り出すことを試みる．そのためには，コミュニティのモデル化手法と，それらの中から特定のトピックに関するコミュニティを同定する技術が必要となる．本研究では，市瀬らの提案した手法 [市瀬 06] と同様な手法を用いて，コミュニティを発見する．本章では，その概要について記述する．

研究者のコミュニティを発見するために，研究論文の書誌情報を利用する．ここでは，共著で論文を執筆した研究者は，その論文で取り扱っている同じ研究トピックに共同で取り組んでいると考える．そこで，研究者をノードとし，研究者が共同で取り組んでいるトピックをラベルとしたエッジで関係を表すとラベル付グラフによって研究者の関係を表すことが可能となる．これを図示すると，図 1 のようになる．

ここで，研究者のコミュニティを同じ研究の興味やトピックによって密につながれたクラスターと定義すると，目的となる研究者コミュニティは，同じトピックによってラベル付けされたエッジを持つクラスターであると考えられる．書誌情報を利用して上記の方法で構成したグラフでは，研究トピックがエッジのラベルとして表されるため，ユーザにとって興味の無いエッジを消去することで，研究者のコミュニティを得ることができる．つまり，ユーザがある特定の研究トピックを指定した時に，それに関係の無いエッジを消去することで，研究者のコミュニティを抽出することが可能になる．

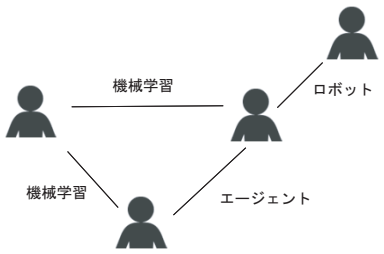

図 1: 研究者のネットワークモデル

しかし，このようにして発見されたコミュニティは，これだけでは個々の特性が表されないため，それぞれのコミュニティが実際にはどういうコミュニティを指しているのかが分からない．つまり，あるトピックに関して十分な知識を持っていないと，そのコミュニティがどういう知識を暗黙的に蓄えているのかが分からないという問題が生ずる．そこで，発見されたコミュニティに対して，そのコミュニティの特徴を抽出することで，どのような知識がそのコミュニティに保持されているのかの明示化を試みる．それぞれのコミュニティに属する著者によって書かれた論文のキーワードをここでは使う．もし，あるキーワードがコミュニティの構成員によって頻繁に使われるものであれば，そのキーワードはそのコミュニティを特徴づけていると考えることができる．しかし，単純にそのようなキーワードの数を数えるだけだと，キーワード間の関係性が失われてしまう．そこで，論文毎にキーワードを考えることとし，コミュニティの特徴の抽出を下記のアルゴリズムで行う．

1. コミュニティに属する研究者によって書かれた論文を取り出す．
2. 1．で取り出された論文に対して，論文をトランザクション，キーワードをアイテムとして，Apriori アルゴリズム [Agrawal 94] を適用し，頻出語の組み合わせをコミュニティの特徴とする．

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，Fax:03-3556-1916，E-mail:ichise@nii.ac.jp

表 1: システムで使ったデータのレコード数

	レコード数
論文	131,000
研究者	93,000
著者	358,000
共著	519,000
トピック	40,000

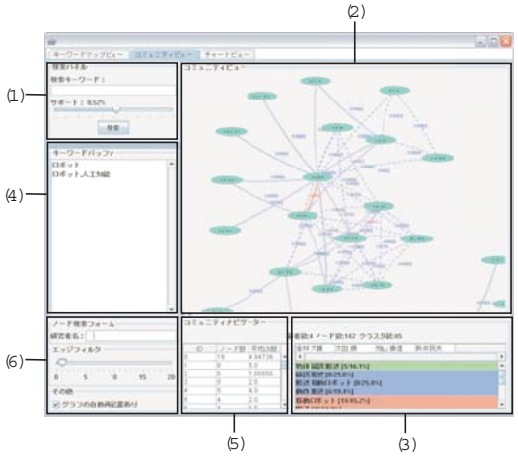

図 2: システムの動作画面

3. コミュニティマイニングシステム

本研究では，前章で述べた手法を用いて，コミュニティマイニングシステムの実装を行った．実装システムは，クライアントサーバ方式で実装されている．サーバ側は，データベースシステム MySQL のコンポーネントを含む Perl で実装されたプログラムが使われ，Web サーバを通して，クライアントとデータのやりとりがされる．クライアントの方はブラウザを通して，サーバから送られてきた JAVA プログラムを実行し，必要に応じて，サーバ側に情報を問い合わせるような形式になっている．JAVA のプログラミングには，JUNG を利用した．

本研究では，文献情報として，CiNii [CiNii 06] のデータベースの一部を用いた．システムに使われたデータのレコード数は，表 1 のようになっている．表中の研究者は，データベース中に出てくる研究者の数を，著者は，執筆者の延べ人数を，共著は，共著者の組合せの数を，トピックは，含まれるトピックの種類の数を表す．本研究では，論文のタイトルに出現する語をトピックとして用いた．

図 2 に，システムのスクリーンショットを載せる．(1) の部分は，ユーザが調べたいトピックを入力する場所である．ここに，トピックを入力すると，(2) にそのトピックで発見されたコミュニティが表示される．表示されたコミュニティの中からどれかを選択すると，第 2 章で説明したアルゴリズムに従って，そのコミュニティの特徴を表すキーワードが計算され，(3) の部分に表示される．

これらの基本機能の他に，用いたトピックの履歴を表示 (4)，コミュニティの統計量（コミュニティに含まれるノード数など）を表示 (5)，表示されているノードの検索機能および表示されるエッジの重みの変更機能（表示されるノードの数が変化する）(6) がシステムに含まれている．

4. 考察

提案システムを用いると，容易に研究者のコミュニティを発見することができる．このように研究者のコミュニティを発見するシステムとして，CiteSeer [CiteSeer 06] や Google Scholar [Google 06] がある．しかし，このようなシステムでは，着目した特定の研究者が含まれるコミュニティを発見するのに適してはいるが，その近傍にあるコミュニティを発見することが難しい．しかし，本システムでは，近傍にあるコミュニティも表示されるため，発見が容易である．Newman の研究 [Newman 04] では，本研究と同様に，共著ネットワークを用いて，コミュニティを分析している．しかし，この研究では，分野間の違いなどの分析に共著ネットワークを用いているに過ぎず，本システムのように，特定のコミュニティを発見することはできない．市瀬らの研究 [Ichise 05] では，本研究と同様に，ユーザのインタラクションを通してコミュニティを発見する手法の提案をしている．しかし，本研究で提案したシステムでは，ユーザにあらかじめ探したいコミュニティのトピックを入力させることで，巨大なコミュニティをトピック毎に切り分けることが可能である．

5. むすび

本論文では，リスクを検知するためには，コミュニティで保持される暗黙知を抽出することが重要であるという観点から，コミュニティを発見するシステムの提案を行った．提案したシステムでは，文献のデータを用いて，特定のトピックのコミュニティを発見することが可能となる．今後は，このシステムを使って，コミュニティで保持される暗黙知が，どこまで発見できるかを検証していく予定である．

参考文献

[Agrawal 94] Agrawal, R. and Srikant, R.:Fast algorithms for mining association rules, In Proceedings of the 20th International Conference on Very Large Data Bases, pp. 487–499, (1994).

[CiNii 06] 国立情報学研究所: NII 論文情報ナビゲータ, http://ci.nii.ac.jp/, (2006).

[CiteSeer 06] CiteSeer.IST, Scientific literature digital library, http://citeseer.ist.psu.edu/, (2006).

[Google 06] Google, Google scholar, http://scholar.google.com/, (2006).

[Ichise 05] Ichise, R., Takeda, H., and Ueyama, K.: Community mining tool using bibliography data, In Proceedings of the 9th International Conference on Information Visualization, pp. 953–958, (1994).

[市瀬 06] 市瀬龍太郎，武田英明，村木太一: リスクマイニングのための研究者コミュニティの発見方法, 人工知能学会研究会資料，SIG-KBS-A503, pp. 19-24, (2006).

[Newman 04] Newman, M. E. J.: Coauthorship networks and patterns of scientific collaboration, In Proceedings of the National Academy of Sciences of the USA, Vol. 101, No. suppl. 1, pp. 5200–5205, (2004).