大規模文献集を用いた学問分野の発見とその活用に向けて

田中　紘○（東京工業大学），脇田　建（東京工業大学）

# Discovering academic fields using large-scale literature collection

# and their utilization

## Hiro TANAKA and Ken WAKITA

##### **ABSTRACT**

Given a collection of scientific literature, we successfully identified academic fields and visualized their structure. To deal with the large number of articles and topics, we employed a method to build a hierarchy of research topics from sampled articles. The visual interface supported by interactive topic modeler allows the user to interactively manipulate the research area hierarchy.

**Keywords:** Visualization, Topic Modeling, Text Analysis, SS-NMF

# 序　論

科学技術に関する文献数が増えているため、効率的な分析のための高度な分析ツールが求められている。科学的技術に関する文献の検索と分析をサポートするインタラクティブな可視化アプローチが複数なされている (Federico et al. 2017)。時系列分析によって、研究のトレンドを発見すること (Dou et al. 2013)、引用を元に、研究分野間の関連性を概観すること (Nakazawa, Itoh, and Saito 2015)は重要なテーマである。文献集から、学問分野の抽出することはこれらのアプローチに重要な役割をもつ。

実際に、文献集から学問分野を抽出するタスクを考えてみる。文献集は電子化され、大量に貯蔵されている。計算機科学のみで考えても、IEEE Xplore[[1]](#footnote-1)には4,845,040の文献、ACM Digital Library[[2]](#footnote-2) には407,367の文献がある。加えて、抽出すべき学問分野も大量にある。ACM’s 2012 Computing Classification Scheme (CCS)[[3]](#footnote-3)には79分野が存在し、情報処理学会の2018年の取り扱い分野[[4]](#footnote-4)には43の分野についての研究会が存在する。

学問分野を抽出する際には、文書集合から潜在的な話題（トピック）を抽出する**トピックモデリング**がよく用いられる。トピックモデリングは、文書集合から、潜在的な話題(トピック)を発見することを目的としている。トピックは、そのトピックに特徴的な単語群を持ち、文書は、それらのトピックによって構成されるという仮定のもと、各文書に出現する単語の頻度から、各トピックに特徴的な単語群と、各文書がどのトピックから構成されるかを推定する事によって行われる。

トピックモデルはトピックを単語の分布として出力する。学問分野の抽出にトピックモデリングを利用する場合、単語群から、そのトピックがさす概念を人間が解釈し、その概念と対応する学問分野として扱う必要がある。しかし、トピックが学問分野として解釈することが困難な場合がある。さらに言えば、人間に解釈が困難なトピックを出力する場合すらある (Mimno et al. 2011)。トピックは人間によって解釈することが必要なため、トピックが解釈できない場合、人間によってトピックモデルを修正することが必要である。

学問分野の抽出の際には、大規模な文献集を扱えること、大量のトピックを持つ文書を扱えること、人間によってトピックモデルを修正できること、の三点が必要である。本研究は、これらの三点を達成することを目的としている。

人間にとって解釈しやすいトピックを得ることを目的として、ユーザーによる直接的なトピックモデルの容易な修正を可能にすることで、ユーザーとの対話を繰り返し、人間の解釈しやすいトピックをえる**人間参加型トピックモデリング**が研究されている。人間参加型トピックモデリングは、トピックモデルによって得られたトピックをユーザーに可視化し、修正したい点を発見したら、トピックモデルの修正を与える。その後、再びトピックモデルがトピックを出力し、そのトピックを可視化し、というループを繰り返す事によっておこなわれる。しかし、学問分野の抽出に用いる場合、二つの問題がある。

一つは、大規模化に関する問題である。人間参加型トピックモデリングでは、ユーザーのトピックモデルの修正の度にトピックを再び求める。そのため、ユーザーがトピックモデルの修正を行う度に、トピックが出力されるまでユーザーは待つ必要がある。大規模文書を扱う際には、この待ち時間が実用的でないほど長くなってしまう。例えば、人間参加型トピックモデリングのための一システムであるUTOPIAN (Choo et al. 2013)は、515の文書でを扱う場合でも、再びトピックを出力するまで48秒もの時間が必要であり、2,211の文書に対しては、対話的にモデルを操作することは、不可能であると報告されている (Lund et al. 2017)。

二つは、ユーザーの認知的負荷と作業量が増大する問題である。人間参加型トピックモデリングでは、ユーザーは、トピックの修正したい点を発見するために、トピックモデルの出力するトピックたちの把握を繰り返す必要がある。扱うトピック数が増加した場合、ユーザーは膨大な数のトピックの把握を繰り返す必要がある。トピックモデリングを利用したシステムの一つであるNSF Portfolio Explorer が作成される際、作者が1000個のトピックから分析と可視化のために30のトピックを選択するために数日を要した (Council 2014)ようにトピックの把握は労力を要する作業である。また、トピック数が増加した場合、ユーザーはトピックモデルの修正に対するトピックの変化を把握しきれないことが予想される。そのため、大量のトピックを扱うことはユーザーの負荷の面から現実的でないと言える。

以上の二点から、人間参加型トピックモデリングにおいて、大量のトピックを持つような大規模な文書を扱うことは困難であると言える。既存の人間参加型トピックモデリング手法の適用事例は、Table 1 のように、文書数について数万文書程度、トピック数については20個程度のトピックを扱うにとどまっている。

学問分野の抽出の際には、大規模な文献集を扱えること、大量のトピックを持つ文書を扱えること、人間によってトピックモデルを修正できること、の三点が必要である。本研究では、人間参加型トピックモデリングの、大規模化に関する問題、ユーザーの認知的負荷と作業量が増大する問題を解決することで、学問分野の抽出のために必要な三点を達成することを目的とする。

ユーザーの認知的負荷と作業量が増大する問題に対処するため、全てのトピックを一度に扱うことはせず、一度に扱うトピックを少数に限定し、段階的に求める。まずは、全体を大雑把に分割するような少数のトピックを求める。それぞれ分割されたトピックから、もう少し詳細に分割するような少数のトピックを求める。この階層的な分割を繰り返すことで、一度に扱うトピックを少数に限定したままで、詳細なトピックを取得する。

大規模化に関する問題に対処するため、トピックの発見をサンプル文書から行う。提案手法では、大雑把に分割するような、少数のトピックのみを繰り返し求める。その際には、全体の文書ではなく、大規模化に関する無作為に取得した問題が起きない程度の大きさのサンプル文書を扱えば十分であると仮定する。サンプル文書から得られたトピックを全体のトピックとして扱い、全体の文書がどのトピックに属するか求めることで、全体のトピックモデリングを近似的に行う。

本論文の構成を以下に述べる。2章では本研究に関連する研究を述べ、3章では、提案手法と、その元となる半教師あり非負行列因子分解について述べる。4章では、提案手法を用いた分析システムを紹介し、5章で実際に大規模文献集の分析を行い、6章で考察を述べ、7章で今後の課題、8章でまとめを述べる。

# 関連研究

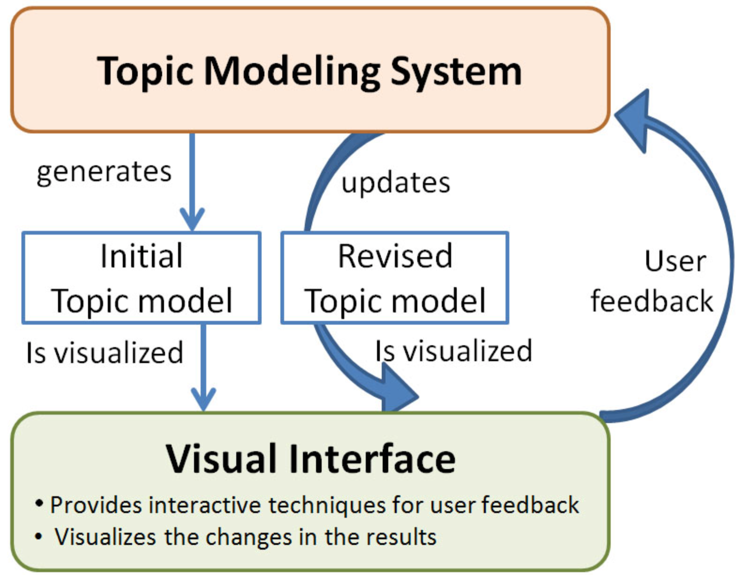
本章では、トピックモデリングと、メタデータを加えたトピックの可視化について触れる。

## トピックモデリング

トピックモデリングは、文書集合から、潜在的な話題(トピック)を発見することを目的としている。トピックは、そのトピックに特徴的な単語群を持ち、文書は、それらのトピックによって構成されるという仮定のもと、各トピックに特徴的な単語群と、各文書がどのトピックから構成されるかを各文書に出現する単語の頻度から推定する事によって行われる。 **LDA** (Blei, Ng, and Jordan 2003)等のトピックモデリング手法は、各文書に出現する単語の頻度からトピックを単語の分布として、各文書がどのトピックから構成されるかをトピックの分布として抽出する。これらによって、文書集合の要約を得ることができるため、文書集合の解析に広く用いられる。

学問分野の抽出にトピックモデリングを用いる場合、文献集を入力として、単語の分布として出力されたトピックを人間がどの概念を指すのか解釈し、対応する学問分野として扱う。また、各文書がどのトピックに属するかを見る事で、その分野に属する文書を得る事ができる。

しかし、トピックモデリングでは、解釈困難なトピックが出力される場合がある事 (Mimno et al. 2011)、分野の専門家の感覚とは一致しない場合がある事 (Saeidi et al. 2015)が知られている。これらのトピックに対しユーザーは、 NSF Portfolio Explorer を作成する際、1000個のトピックから分析と可視化のために30のトピックを選択し、970のトピックを捨てた (Council 2014)ように、トピックを採用するか、捨てるかの選択肢しかない。

Fig1 Outline of human-in-the-loop topic modeling (Hoque and Carenini 2015)

そこで、ユーザーによる容易なトピックの修正を可能にし、ユーザーとの対話を繰り返すことで、ユーザーの知識を取り入れ、よりユーザーの意図に沿った解釈しやすいトピックを得ることを目的とした人間参加型トピックモデリングが研究されている。Interactive Topic Modeling (Hu et al. 2014)はユーザーが指定した単語同士を同一視することにより、指定された単語同士を同じトピックから出現しやすくする操作を可能にし、ITMVis (Saeidi et al. 2015)はこの手法をソースコードの解析等に活用した。Anchor method (Lund et al. 2017)は、単語間にリンク不可能、リンク必須のリンクを貼ることのみに限ることで語彙数にスケールする高速な手法を提案した。iVisClustering (Lee et al. 2012)は、入力する文書集合のフィルタリング、単語の重み変化によるトピックモデルの操作を、文書のクラスタリングに活用した。UTOPIAN (Choo et al. 2013)は、**非負行列因子分解**に制約を追加する簡単なアルゴリズムで、より直接的な操作が可能な半教師あり非負行列因子分解を提案し、視覚的分析を行なった。

しかし、人間参加型トピックモデリングには問題が二点ある。一つ目は、大規模化に関する問題、二つ目は、ユーザーの認知的負荷と作業量が増大する問題である。既存手法の適用事例を見てみると、Table 1 のように最大でも、文書数について数万文書程度、トピック数については20個程度である。これらの手法をそのまま用いて、大量のトピックを持つような大規模文献集を分析することは困難である。

本研究では、半教師あり非負行列因子分解を活用し、少数のトピックの発見の階層的繰り返しと、サンプル文書からのトピックの発見によって、人間参加型トピックモデリングの、大規模化に関する問題、ユーザーの認知的負荷と作業量が増大する問題を解決する。

Table 1 Application example of human-in-the-loop topic modeling

|  |  |  |
| --- | --- | --- |
| 論文とデータセット | 文書数 | トピック数 |
| (Hu et al. 2014) New York Times Article | 2,000 | 20 |
| (Hu et al. 2014) 20News | 18,846 | 20 |
| (Lund et al. 2017) 20News | 18,846 | 20 |
| (Saeidi et al. 2015) JEditのソースコード | 538 | 15 |
| (Smith et al. 2018) BBC news articles | 2,225 | 10 |
| (Smith et al. 2018) Twitter US Airline | 14,485 | 10 |
| (Choo et al. 2013) Infovis-VAST abstract | 515 | 11 |
| (Choo et al. 2013) Car Review | 231 | 11 |
| (Choo et al. 2013) TV Review | 110 | 6 |
| (Lee et al. 2012) VAST abstract | 454 | 10 |
| (Lee et al. 2012) 20News | 800 | 9 |

## メタデータを加えたトピックの可視化

科学技術に関する文献の効率的な分析のために、文献に付随する発表年数、引用情報等のメタデータと、各文献がどのトピックに属するかを用いて、有益な情報を可視化する研究がされている。

トピックの時系列分析は、トピックモデルの有効な活用法の一つである。TIARA (Wei et al. 2010)は、文献の出版年を用いて、トピックに含まれる文書数の時間的変化を可視化した。この手法は他の研究者にも活用され、TextFlow (Cui et al. 2011)はトピックの誕生、滅亡とそれに関連する重要なイベントを可視化した。HierarchicalTopics (Dou et al. 2013)は、より多くのトピックを扱い、活発な研究分野と、その原因となった研究資金プログラムの存在を発見した。中澤らは、文献の発表年数、引用情報を用いて、トピック間の引用、トピック内の引用を、その引用年を可視化し、分野間の関係性の分析に役立てた(Nakazawa, Itoh, and Saito 2015)。

これらは、LDA等の人間参加型でないトピックモデルの出力するトピックを基に分析を行っている。学問分野の抽出には、人間によるトピックモデルの修正が必要なため、人間参加型トピックモデルによって得られたトピックを基にこれらの分析を行うことで、よりユーザーの意図にそった分析が可能になると思われる。

# 手法

ここでは、半教師あり非負行列因子分解(Choo et al. 2013)とそれを活用した提案手法について述べる。

## 半教師あり非負行列因子分解

非負行列因子分解によるトピックモデリングでは、各文書に出現する単語の頻度を単語・文書行列で表現する。この行列の 要素は、番目の文書において番目の単語が出現した頻度を表す。そのため、この行列の行と列は、それぞれ、単語がどの文書で使われているか、文書がどの単語で構成されているかを表す。この巨大な単語・文書行列()を非負行列因子分解によって、比較的少数のトピックに関して単語・トピック行列()とトピック・文書行列()の積で近似することができる。 の各列を見る事でトピックの単語分布、 の各列見る事で文書のトピック分布を得ることができる。各文書は、トピックの単語分布とそのトピックが文書を構成する割合を掛け合わせたものを全トピックについて足し合わせたもので近似される。

行列分解の近似精度は、近似前後の行列のユークリッド距離を用いて次の損失関数で表現できる。一般的な非負行列因子分解では、損失関数を の非負条件下で最小化する事によって求められる。

chooらの半教師あり非負行列因子分解(Choo et al. 2013)では、 の他に追加の入力として、 に関しての参照行列 と、それぞれの参照の重みを示す対角行列 を受け取り、 の他に最適化する変数として、 の行ごとのスケーリングを行うための対角行列 を追加し、損失関数に に関するペナルティ項を追加し、次のように変更する。

ペナルティ項の追加により、 の各列は、 で与えられた重みに応じて、 の各列にできるだけ近づきながら の近似をする。この損失関数を、 の非負条件、 の対角条件下で最小化することで を求める。 に適切な値を設定することで、 の各列をそれぞれ監督できるため、各トピックの単語分布、各文書のトピック分布をそれぞれ個別に監督できる。これらに加えて、トピック数に対応する、 の列数、 の行数を調整することにより、トピックの分割、削除、見本文書からの新規作成が可能である。

半教師あり非負行列因子分解を用いた人間参加型トピックモデリングは、最初に通常の非負行列因子分解を用いて、初期のトピックモデルを作成し、 から得られる単語とトピックの関係、トピックと文書の関係を可視化することで、トピックの情報をユーザーに伝える。ユーザーがトピックの修正したい点を発見し、何らかの操作を行うと、ユーザーの操作に合わせて、 に適切な値を設定し、損失関数の最小化による、 の最適化を行い、再び から得られるトピックの情報を可視化し、と繰り返すことによって行われる。

## 提案手法

人間参加型トピックモデリングを学問分野の抽出に用いるには問題が二点あった。そのため、半教師あり非負行列因子分解を用いた人間参加型トピックモデリングをそのまま大規模文献集から学問分野の抽出に用いるには二つの問題がある。

一つは、大規模化に関する問題である。損失関数の最小化による、 の最適化を繰り返し行う。この最適化の計算量はであるため、大量のトピックを含むような大規模文献集を扱う場合、文書数、トピック数が増加するため、計算量が増加する。加えて、から得られる、各文書のトピック構成を可視化する事が困難になる。

二つは、ユーザーの認知的負荷と作業量が増大する問題である。ユーザーはから得られたトピックの情報を可視化から、トピックの修正したい点の発見を繰り返す。大量のトピックを含むような大規模文献集を扱う場合、ユーザーは可視化から大量のトピックの特徴を繰り返し把握する必要がある。

これらの問題は、扱うトピック数、文書数が大きいために起こっている。そのため、提案手法では、一度に扱うトピック数、文書数を削減することで、これらの問題に対処する。

扱うトピック数に関して、一度に扱うトピックを少数に限定する。全てのトピックを一度に求めることはせず、次のようにして段階的に求める。まずは、全体を大雑把に分割するような少数のトピックを求める。それぞれ分割されたトピックから、ユーザーは分析したいトピックを選択し、もう少し詳細に分割するような少数のトピックを求め、元のトピックをさらに分割する。この階層的な分割を繰り返すことで、一度に扱うトピックを少数に限定したままで、もともと求めるべきであった全てのトピックに対応するような詳細なトピックを取得する。

扱う文書数に関して、人間参加型トピックモデリングを、比較的小さなサンプル文書集合に対して行う。提案手法では、大雑把に分割するような、少数のトピックのみを繰り返し求める。そのようなトピックを発見するには、全体の文書ではなく、大規模化に関する問題が起きない程度の大きさ(本研究では1,000件とした)の無作為にサンプリングした文書を扱えば十分であると仮定する。大規模な文書集合全体のトピックを直接求めるのは困難なため、以下のようにして近似的に求める。

以下では、サンプル文書に対するトピックと、全体に対するトピックを区別するため、サンプル文書に関する記号には上線をつけて区別する。全体の単語・文書行列()の代わりに、サンプル文書集合の単語・文書行列()を入力として受け取り、人間参加型トピックモデリングを通してを求める。全体のトピックと単語の関係と、サンプル文書のトピックと単語の関係は同一であるという仮定の元、を全体の単語・トピック行列()と同一であるとして、全体のトピック・文書行列()のみを求める。繰り返し最適化されるのはであるため、ユーザーがトピックモデルの修正を行う度に、最適化の計算量は削減され、大規模化に関する問題は起きない。

以上によって、一度に扱う文書数、トピック数を削減し、人間参加型トピックモデリングを大規模文献集から学問分野の抽出に活用する。提案する手法をまとめると以下のステップのようになる。

1. 全体の文献集を一つのクラスタとして選択する。
2. 選択された文書クラスタから無作為にサンプル文書を1000件選択し、その単語・文書行列を とする。 に零行列を設定し、に対して、以下のステップで半教師あり非負行列因子分解による人間参加型トピックモデリングを行う。
3. 損失関数を最小化するような を求める
4. から得られるトピックの情報を可視化する。
5. ユーザーが修正すべき点を発見したら、それに合わせて操作する。修正すべき点がなければステップ3に進む
6. ユーザーの操作に合わせて に適切な値を設定し、ステップi に戻る。
7. ステップ2で得られた を選択中のクラスタ全体のものとして扱い、とする
8. 選択中の文書クラスタ全体の単語・文書行列 に対し、 を固定し、損失関数 を最小化する を求める。
9. の各列から、各文書のトピック分布がわかるので、もっとも関わりの強いトピックのクラスタにその文書が属するとして選択中の文書クラスタを分割(サブクラスタリング)する。
10. ユーザーはさらに分析したいトピックの文書クラスタを選択し、ステップ2に戻る。

提案手法では、各階層で、ユーザーの意図に沿ったトピックが取得でき、さらに、ユーザーが分析を進めるトピックを選択するため、ユーザーの興味に沿って分析を進めることができ、全体で把握するトピック数に関しての作業量を削減できると期待できる。また、人間参加型トピックモデリングをサンプル文書から少数のトピックに限定して行うことで、人間参加型トピックモデリングを大規模文献集から学問分野の抽出に活用する際の二つの問題点に対応した。

# システム

## 設計上の考慮点

提案手法において、ユーザーの行うことは、トピックから修正すべき点を発見すること、トピックの修正すべき点をシステムに伝えること、さらに分析したいとトピッククラスタを発見し選択することの三つである。そのため、システムでこれらをサポートするために必要なことは以下の三点である。

**DC1**. トピックの修正すべき点を容易に発見できる。

**DC2**. トピックの修正すべき点を容易にシステムに伝えられる。

**DC3**. 分析したいトピックのクラスタを発見できる。

それぞれについて、詳細に触れる。

##### fig4.1DC1

トピック修正すべき点として、以下のものを想定する。

* トピック中に複数の概念が混在している。(概念の混在)
* 同じ概念を指すトピックがある。(重複トピック)
* 意味が見出せないトピックがある。(不要トピック)
* トピックの指す概念で重要な単語がトピック中に出てこない。(重要単語の欠落)
* トピックの指す概念とは関係ない単語がトピック中に出現している。(不要単語の混入)
* 欲しいトピック、あるべきトピックがない。(トピックの欠落)

これらの修正すべき点を発見できるような機能が必要である。

##### DC2

DC1で挙げたトピックの修正すべき点に対して、半教師あり非負行列因子分解を用いた人間参加型トピックモデリングで可能なトピックモデルの修正方法を以下のように対応させることができる。

Table 2 Points to fix and how to fix

|  |  |
| --- | --- |
| トピックモデルの修正方法 | トピックの修正する点 |
| トピックの分割 | 概念の混在 |
| トピックの削除 | 重複トピック、　不要トピック |
| 単語の分布の監督 | 概念の混在、重要単語の欠落、　不要単語の混入 |
| 見本文書から新規トピックの作成 | トピックの欠落 |

これらの対応関係があるので、それぞれの修正方法を容易に呼び出せるようにすれば良い。

##### DC3

分析したいトピックのクラスタを発見には二つのパターンが考えられる。一つ目は、ユーザーがその時点で存在しているトピックのクラスタ群から分析したいトピックのクラスタを決めている場合、二つ目は、ユーザーがその時点で存在しているトピックのクラスタ群から注目に値するクラスタを探している場合である。

一つ目に対しては、トピックの階層構造を利用し、ファイルマネージャのような、階層構造に沿ったナビゲーションによって、発見が容易になると期待できる。一方、二つ目に対処するためには、各クラスタに対応するトピックの意味をみる必要がある。そのため、階層構造に沿ったナビゲーションと、各クラスタに対応するトピックの意味の可視化が必要になる。

## システム概要

DC1, DC2, DC3を満たすように設計し、加えて、効率的な分析をサポートするために、操作の取り消し機能、モデルの保存、読み込み機能をつけた。

Fig.2 User Interface

システムはFig.2 のように、三つのViewからなる。TreeView で分析をするクラスタを選択すると、そのクラスタからサンプル文書が取得され、WordView, DocView にから得られるサンプル文書に関するトピックの情報が可視化される。ユーザーは、トピックの修正すべき点の発見、トピックの修正、可視化を繰り返し、修正が十分と判断したら、選択中のクラスタを現在のトピックを用いてサブクラスタリングし、TreeView が更新される。

以下では、それぞれのViewの詳細を説明し、モデルの修正をするための操作方法を説明する。

### TreeView

トピックのクラスタ群を可視化し、DC3を満たすように設計されている。

画面上でクラスタをクリックすると、その子となるクラスタが、そのクラスタに対応したトピックに割り当てられた色で、図のように強調表示される。トピックごとに割り当てられた色は、WordView, DocView でも共通して使用されるため、それぞれの指すトピックの情報を容易にの組み合わせることができる。また、クラスタをダブルクリックすることで、子クラスタの展開、折り畳みが行われる。これらによって、ファイルマネージャのような、階層構造に沿ったナビゲーションが可能である。

また、各クラスタの配置は、クラスタの対応するトピックの単語の分布ベクトルをt-SNE (Maaten and Hinton 2008)を用いて二次元平面上に射影することによって決定されている。これによって、クラスタの配置から、そのクラスタの対応するトピックの類似度をみることができる。加えて、各クラスタは、対応するトピックに最も結びつきの強い単語で注釈されている。ユーザーが他の意味合いを発見した場合、その注釈を書き換えることができる。これによって、各クラスタの対応するトピックの具体的な意味も概観することができる。

### WordView

から得られるサンプル文書に対するトピックの単語分布を可視化し、DC1を満たすように設計されている。

各トピックに特徴的な単語は、そのトピックに属する文書でよく使われる単語であるため、その分布が大きいと考えられる。そのため、WordView では、適当な閾値を設け、各トピックごとに、閾値以上の分布を持つ単語のみをそのトピックに割り当てられた色で表示する。これらの単語から、ユーザーはトピックの指す概念を推測でき、トピックの修正すべき点の発見に役立てられる。

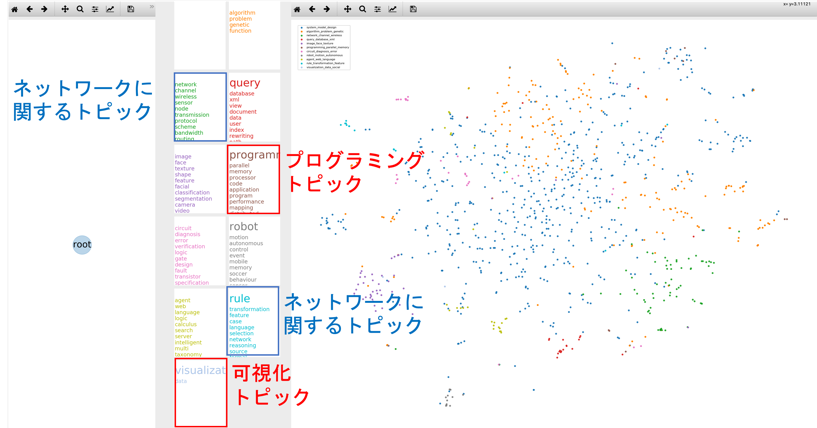
### Fig5.1DocView

から得られるサンプル文書のトピック分布を可視化し、DC1を満たすように設計されている。

各サンプル文書がどのトピックに属するかをTreeView, WordView に対応した色の頂点で可視化する。各頂点の配置は、サンプル文書の単語・文書行列をt-SNE (Maaten and Hinton 2008)を用いて二次元平面上に射影することによって決定されている。そのため、ユーザーはそれぞれの配置から、サンプル文書同士の類似度を概観することができる。各頂点の意味を確認するため、ユーザーは、マウスホバーによってその文書のタイトルを確認することができる。これらを組み合わせることで、トピックがサンプル文書中でどのように広がっているかを概観でき、トピックに属する文書をみることで、より具体的にトピック指す概念を見ることができる。

### Fig5.2モデルの修正をするための操作

単語の分布の監督は、WordView上で表示されている単語を上下にドラッグすることで、トピック中のその単語の重みを大小するように操作できる。トピックの削除はWordViewで、トピックを選択し、キーボードで-を押す事で選択中のトピックを削除することで行われる。トピックの分割は、WordViewでトピックを選択し、キーボードで+を押す事で、そのトピックのコピーが作成され、それぞれのトピックが混在していた別々の概念を指すように、単語の分布の監督を行うことによって実行される。見本文書から新規トピックの作成は、DocView上で見本となる文書を選択し、キーボードで+を押す事で選択中の文書を見本文書とした新規トピックが作成される。

システムは、モデルの操作を行うために、ユーザーの各操作に合わせて、 に自動的に適切な値を設定する。ユーザーが操作する際には、これらの行列に直接値を設定することはない。ユーザーは、これらの行列を意識することなく、WordView とDocView を用いて、容易にモデルの修正を行える。

# 実際の解析

## データセット

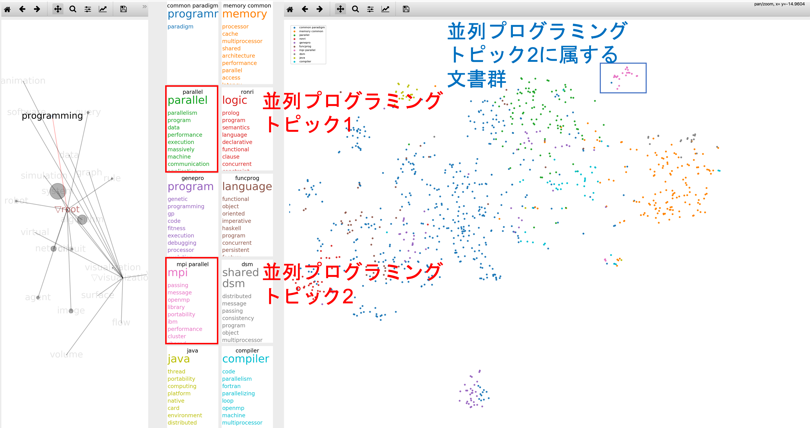
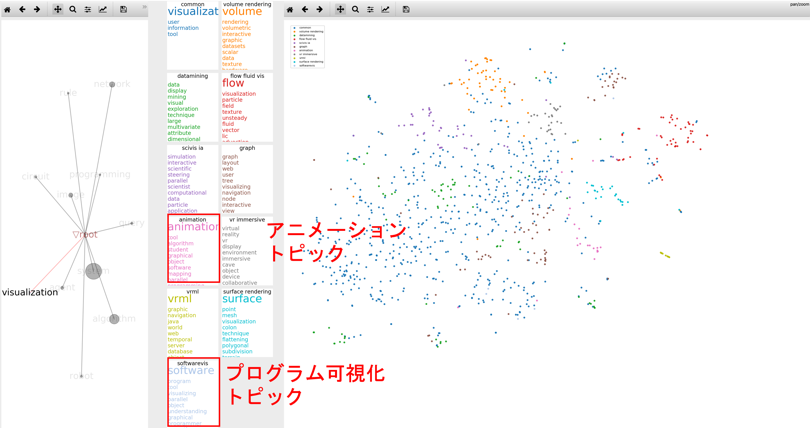
実データの分析を通して考察を行うため、主に計算機科学関連の文献からなるCiteSeer Metadataを分析する。これは1,355,242件の文献のタイトルと要旨とその他メタデータを持つ。これらの文献集には、英語以外の文書、要旨が短すぎるもの、重複した文書等のノイズとなる文書が多数含まれているため、自然言語解析モジュールpolyglot (Chen and Skiena 2014)による言語判定で英語とならない文書、要旨が100単語以下の文書、タイトルが重複している文書は除去した結果、686,324件の文献が残った。これらについて、タイトルと要旨をつなげたものを一つの文書として扱い、作成したシステムによる分析を行なった。

## 適用事例

Fig.3 Initial state of a topic obtained from a cluster of whole references. User cannot find a topic of visualization. Create a new topic based on a document of visualization.

Fig.4 The visualization topic has been created. Dissatisfied with the word distribution of visualization and the programming topics, User supervises the word distribution.

Fig.5 User finished the correction of the topic. It can’t be determined which topic cluster a document about OS is included in. There are two network related topics

Fig.6 Initial state of a topic obtained from a cluster of the visualization topic. Since there was a mixture of concepts, User divided the topic.

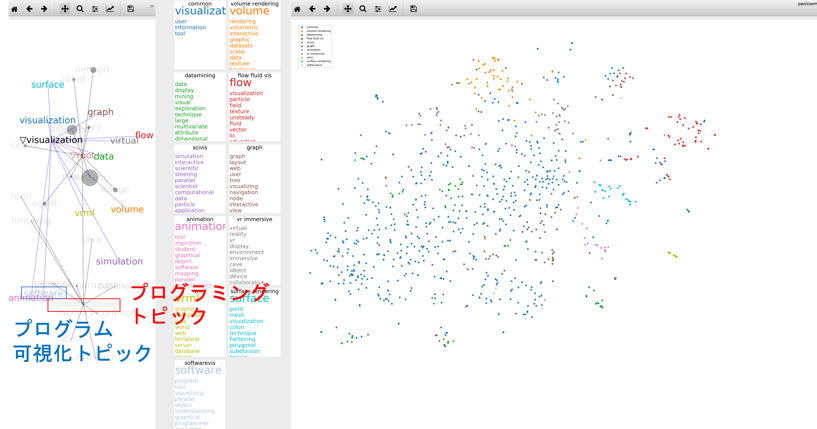
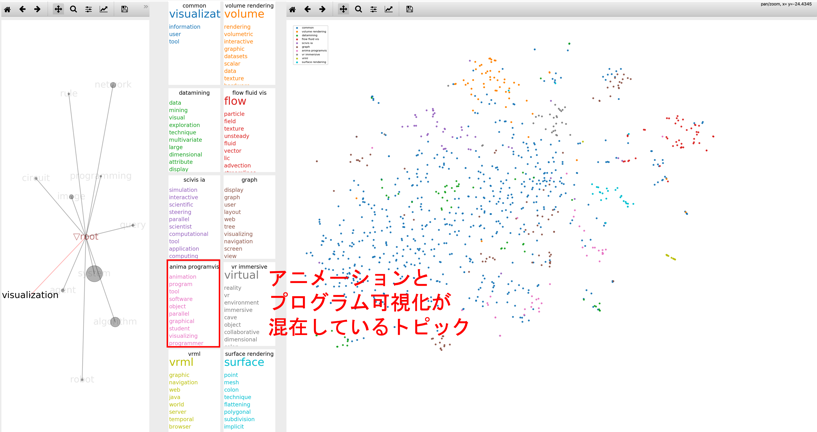
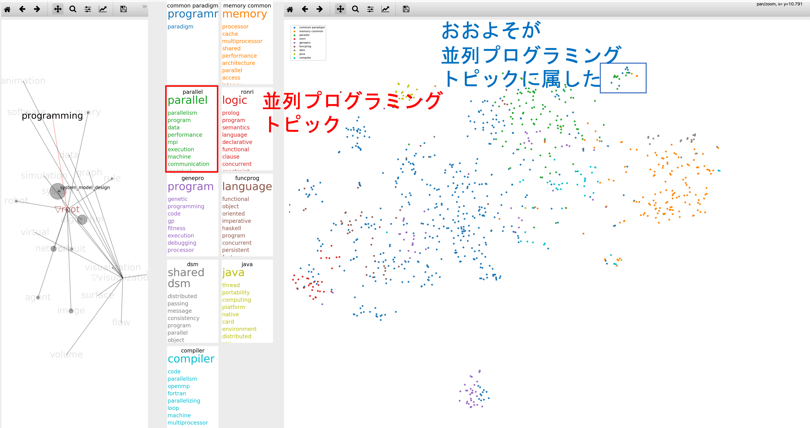
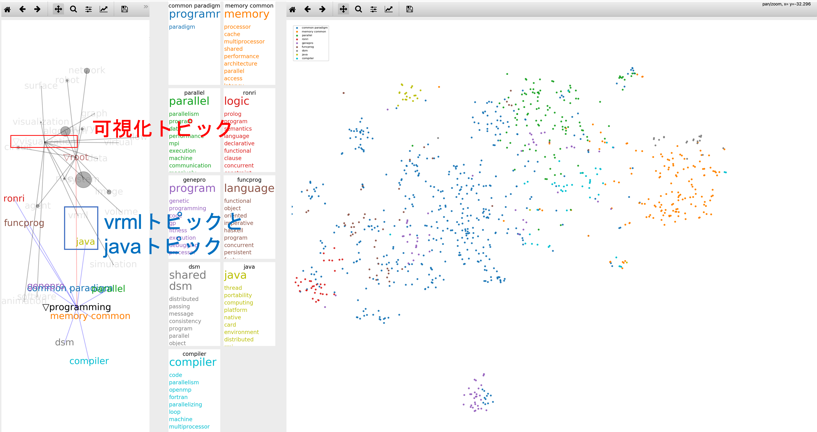
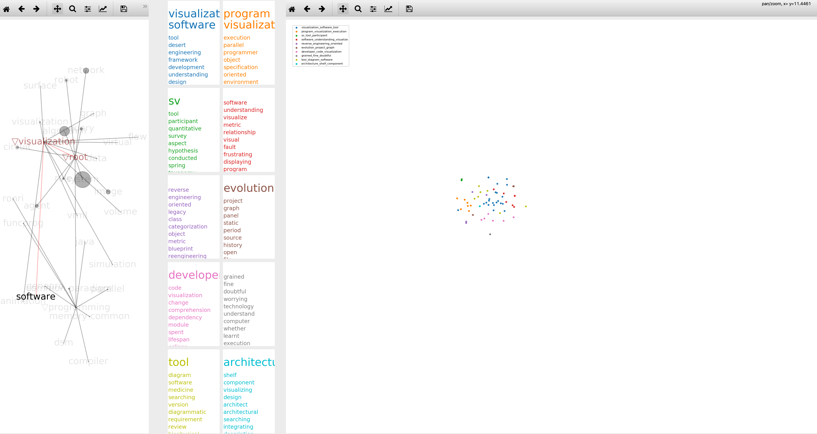
 Fig.8 Initial state of a topic obtained from a cluster of the programming topic. Since duplicate topics were seen, delete topics.

Fig.10 View subclusters of the programming topic. There is no topic that is directly close to visualization. However, it turns out that the java topic is close to the VRML topic, which is a subcluster of the visualization topic

Fig.7 User separated the mixed concept. User finished the correction of the topic.

Fig.9 User eliminates duplicate topics. User finished the correction of the topic.

Fig.11 View subclusters of the visualization topic. User is interested and selected because program visualization topic is close to the programming topic

Fig.12 Initial state of a topic obtained from a cluster of the program visualization topic.

ユーザーは可視化とプログラミングに興味があり、それらに関連する分野を抽出したいと考えているとする。

まず、文献集全体のクラスタを選択し、それらに対するトピックを見た。(Fig.3) 初期状態において、可視化に関するトピックは見つからないため、ユーザーは、可視化に関する文献を見本文書に新規トピックの作成を行い、可視化に関するトピックを得た。(Fig.4) 可視化トピック、プログラミングトピックに対して、不必要な単語が見つかったため、それらのトピックに対して、単語の分布の監督を行い、それらの単語を削除し、ここでトピックの修正を終了し、文献集全体のクラスタのサブクラスタリングを行なった。(Fig.5)

その後、可視化トピックのクラスタを選択し、それらに対するトピックを見た。(Fig.6) アニメーションとプログラム可視化が一つのトピック中に存在し、これらを分離したいと考え、トピックの分割を行い、ここでトピックの修正を終了し、可視化トピックのクラスタのサブクラスタリングを行なった。(Fig.7)

その後、プログラミングトピックのクラスタを選択し、それらに対するトピックを見た。(Fig.8) 並列プログラミングに関するトピックが二つ見つかったため、そのうちの一つを削除したいと考え、削除し、ここでトピックの修正を終了し、プログラミングトピックのクラスタのサブクラスタリングを行なった。(Fig.9)

ここで、ユーザーは注目に値するクラスタをTreeViewを用いて探した。まずは、プログラミングトピックのサブクラスタを見た。(Fig.10) 可視化トピックと近いものは見つからなかった。続いて、可視化トピックのサブクラスタを見た。(Fig.11) プログラム可視化トピックがプログラミングトピックと近いことがわかったので、このトピックに注目し、分析を進めた。(Fig.12)

このようにして、ユーザーは、トピック修正と、分析するクラスタの選択を通して、トピックを大量に含むような大規模文献集から、トピックの修正を行いながら、ユーザーの興味に沿って分析することができた。ユーザーがトピックモデルの操作を行った際、再びトピックを出力するまでの時間は1秒以内であった。

ここで、ユーザーが興味を向けなかった部分を見てみる。文献集全体のクラスタのトピックに対して、OSに関する文書がどのトピックのクラスタに属するか判別できない一方で、ネットワークに関連するトピックが二つ見られるといった、(Fig.5) トピックモデルが好ましくない様子が見られた。また、トピックツリーにおいて、可視化トピックのサブクラスタであるvrmlトピックと、プログラミングトピックのサブクラスタであるjavaトピックは近しいことがわかるが、(Fig.10) この二つのトピックが同時に強調されることはないため、トピックが増加していったとき、発見は困難になりうる。

# 考察

学問分野の抽出の際には、大規模な文献集を扱えること、大量のトピックを持つ文書を扱えること、人間によってトピックモデルを修正できること、の三点が必要である。本研究では、人間参加型トピックモデリングを学問分野の抽出に用いる際の大規模化に関する問題とユーザーの認知的負荷と作業量が増大する問題を解決することでこれらの三点を満たすことを目的としている。ここでは、トピックモデルの修正と、二つの問題に関して、適用事例から考察を行う。

まず、トピックモデルの修正に関して考察する。適用事例において、各クラスタから得られたトピックについて、問題点を発見し、そのトピックモデルの修正を行うことで、ユーザーの感覚に沿ったトピックを取得する様子が見られた。しかし、修正を行った際の、ユーザーがトピックにこのように変化して欲しいという期待と実際のトピックの変化がどの程度一致しているかの評価はできていない。この評価を行うことが必要となる。

ユーザーが興味を向けなかった部分から、適切な操作が行われないと、好ましくないトピックモデルが得られることがわかる。実際には、ユーザーが全てのトピックに対して、適切に操作することは困難であると予想される。そのため、ACM’s 2012 Computing Classification Scheme (CCS)等の既存のカテゴリ分類を利用し、ユーザーが興味のない部分、専門知識を持たない部分に対してサポートしていくことが必要になる。

続いて、大規模化に関する問題に関して考察する。全体の文書の代わりに、サンプル文書に対してのトピックモデリングを行うことで、この問題の解決を図った。ユーザーのトピックモデルの操作に対して、1秒以内に再びトピックを出力することができた。

適用事例において、可視化、プログラミングに関する複数の詳細なトピックが取得できたことから、サンプル文書から有益なトピックが得られることがわかった。しかし、全体の文書に対してトピックモデリングを行った場合、得られるトピックがどのように変化するかの評価を行う必要がある。

適用事例において、OSに関する文書がどのトピックに属しているか判別できない様子が見られた。サンプル文書から得られたトピックでは全体を適切に分割できないことを考慮にいれ、クラスタ全体の文書をクラスタリングする手法を見直す必要がある。

最後に、ユーザーの認知的負荷と作業量が増大する問題に関して考察する。少数のトピックの発見を階層的に繰り返すことでこの問題の解決を図った。適用事例において、可視化、プログラミングに関する複数の詳細なトピックが取得できたことから、一度に扱うトピック数は少数にも関わらず、詳細なトピックを取得できることがわかった。ユーザーの分析の興味が定まっている場合、興味のない部分については、大雑把なトピック以上は求めないことで、トピックを認識する作業量の削減ができた。

提案手法において、いくつかの仮定の不適切さが見られた。提案手法では、大雑把な分割を繰り返すことで階層構造の幅は狭くて十分と仮定していた。しかし、適用事例において、OSに関する文書がどのトピックに属しているか判別できない様子が見られたように、適切に大雑把に分割することは困難である。また、大雑把に分割するような概念は、特徴的な単語で表すことが困難であるかもしれない。ある程度詳細な概念から始めるために、階層構造の幅を広くする必要がある。一度に扱うトピック数を増やすと、ユーザーの負荷が増大するため、トピックの発見を幅方向へ繰り返すことが必要である。

提案手法では、トピックの木構造を仮定していた。しかし、複数の既存の分野を統合して扱う分野である境界領域では、その複数の分野の子として境界領域のトピックが出現することが考えられる。そうした場合、トピック群に関して、木構造ではなく、複数の親を持つ有向非巡回グラフのような構造を許すべきである。木構造に沿って発見したトピック群を、ユーザーによって、もしくは自動的に木構造を編集する機能が必要になる。

# 今後の課題

今後の課題についてまとめる。

提案手法に関する評価について、二つの評価が不足している。一つは、トピックモデルの操作の直感性、二つは、サンプリングにより得られるトピックがどのように変化するかである。これらの評価をしていくことは今後の課題である。

考察で発見された問題を解決するため、3つの機能を追加することを考えている。一つは、既存のカテゴリ分類によるサポート機能、二つは、サンプル文書から得られたどのトピックにも属さない文書のクラスタ機能、三つは、木構造の編集機能である。どのトピックにも属さない文書のクラスタ機能によって、全体の文書のクラスタリングの不適切さが改善されることを期待し、このクラスタからトピックを発見することで、トピックの発見を幅方向へ繰り返すことができる。

# まとめ

本研究では、学問分野の抽出に必要な、大規模な文献集を扱えること、大量のトピックを持つ文書を扱えること、人間によってトピックモデルを修正できること、の三点が必要であることを述べた。この三点を満たすため、人間参加型トピックモデリングを学問分野の抽出に用いる際の二つの問題、大規模化に関する問題とユーザーの認知的負荷と作業量が増大する問題を解決することを目的とした。少数のトピックの発見を階層的に繰り返すこと、トピックの発見をサンプル文書から行うことによってこれらに対処する手法を提案した。

提案手法を用いた分析システムを作成し、約70万件の文献集の分析を通して、トピックを大量に含むような大規模文献集から、トピックの修正を行いながら、トピックを取得することができること、ユーザーが分析したい分野が定まっている場合、ユーザーの興味に従って、分析を進めることができることがわかった。一方で、課題が複数発見された。これらの対応をした上で、得られたトピックを用いて、引用関係、発表年数等のメタデータを用いて、有益な情報を可視化することを今後の方針とする。

# 参　考　文　献

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. “Latent Dirichlet Allocation.” *Journal of Machine Learning Research* 3 (Jan): 993–1022.

Chen, Yanqing, and Steven Skiena. 2014. “Building Sentiment Lexicons for All Major Languages.” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, 383–89.

Choo, Jaegul, Changhyun Lee, Chandan K. Reddy, and Haesun Park. 2013. “UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization.” *Visualization and Computer Graphics, IEEE Transactions on*, TVCG, 19 (12): 1992–2001. <https://doi.org/10.1109/TVCG.2013.212>.

Council, National Research. 2014. Edited by Kaye Husbands Fealing, Alexandra S. Beatty, and Constance F. Citro. Washington, DC: The National Academies Press. <https://doi.org/10.17226/18741>.

Cui, W., S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. 2011. “TextFlow: Towards Better Understanding of Evolving Topics in Text.” *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2412–21. <https://doi.org/10.1109/TVCG.2011.239>.

Dou, Wenwen, Li Yu, Xiaoyu Wang, Zhiqiang Ma, and William Ribarsky. 2013. “HierarchicalTopics: Visually Exploring Large Text Collections Using Topic Hierarchies.” *Visualization and Computer Graphics, IEEE Transactions on*, TVCG, 19 (12): 2002–11. <https://doi.org/10.1109/TVCG.2013.162>.

Federico, P., F. Heimerl, S. Koch, and S. Miksch. 2017. “A Survey on Visual Approaches for Analyzing Scientific Literature and Patents.” *IEEE Transactions on Visualization and Computer Graphics* 23 (9): 2179–98. <https://doi.org/10.1109/TVCG.2016.2610422>.

Hoque, Enamul, and Giuseppe Carenini. 2015. “ConVisIT: Interactive Topic Modeling for Exploring Asynchronous Online Conversations.” In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 169–80. IUI ’15. New York, NY, USA: ACM. <https://doi.org/10.1145/2678025.2701370>.

Hu, Yuening, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. “Interactive Topic Modeling.” *Machine Learning* 95 (3). Springer: 423–69.

Lee, Hanseung, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. “iVisClustering: An Interactive Visual Document Clustering via Topic Modeling.” *Computer Graphics Forum*, CGF,. The Eurographics Association; Blackwell Publishing Ltd. <https://doi.org/10.1111/j.1467-8659.2012.03108.x>.

Lund, Jeffrey, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. “Tandem Anchoring: A Multiword Anchor Approach for Interactive Topic Modeling.” In, 896–905. <https://doi.org/10.18653/v1/P17-1083>.

Maaten, Laurens van der, and Geoffrey E. Hinton. 2008. “Visualizing Data Using T-Sne.” In.

Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. “Optimizing Semantic Coherence in Topic Models.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–72. EMNLP ’11. Stroudsburg, PA, USA: Association for Computational Linguistics. <http://dl.acm.org/citation.cfm?id=2145432.2145462>.

Nakazawa, Rina, Takayuki Itoh, and Takafumi Saito. 2015. “A visualization of research papers based on the topics and citation network.” In *Proceedings of the 19th Conference on Information Visualization*. IV. <https://doi.org/10.1109/iV.2015.58>.

Saeidi, A. M., J. Hage, R. Khadka, and S. Jansen. 2015. “ITMViz: Interactive Topic Modeling for Source Code Analysis.” In *2015 Ieee 23rd International Conference on Program Comprehension*, 295–98. <https://doi.org/10.1109/ICPC.2015.44>.

Smith, Alison, Varun Kumar, Jordan L. Boyd-Graber, Kevin D. Seppi, and Leah Findlater. 2018. “Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System.” In *IUI*.

Wei, Furu, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. “TIARA: A Visual Exploratory Text Analytic System.” In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 153–62. <https://doi.org/10.1145/1835804.1835827>.

1. https://ieeexplore.ieee.org/Xplore/home.jsp [↑](#footnote-ref-1)
2. https://dl.acm.org [↑](#footnote-ref-2)
3. https://dl.acm.org/ccs/ccs\_flat.cfm [↑](#footnote-ref-3)
4. https://www.ipsj.or.jp/kenkyukai/bunya2018.html [↑](#footnote-ref-4)