DEIM Forum 2021 F25-3

ニュースアーカイブ探索のための記事間の関係抽出とその可視化

松本　直彰†　　湯本　高行††　　山本　岳洋††　　大島　裕明†,††

† 兵庫県立大学 応用情報科学研究科　〒 650–0047 兵庫県神戸市中央区港島南町 7-1-28
†† 兵庫県立大学 社会情報科学部　〒 651–2197 神戸市西区学園西町 8-2-1
E-mail: †{aa20y510,ohshima}@ai.u-hyogo.ac.jp, ††{yumoto,t.yamamoto}@sis.u-hyogo.ac.jp

あらまし　気になる話題を調べるためにニュースアーカイブが利用される．気になる話題のみをまとめるために検索を行う際，一度検索して読んだ記事を繰り返し検索したり，簡単に記事間を直接移動する手段がなかったり，記事同士の関係性を把握しづらいなどの非効率な検索を強いられる．本研究では，ユーザが自分の気に入った記事の関係性を確認しながら検索を行うことができるようにする．そのために，記事の閲覧やブックマークで更新されるグラフである探索者ビューを提案する．はじめに，ニュースアーカイブに存在するすべての記事に対してノードが記事，エッジが記事間のフォロー関係のアーカイブ記事グラフを作成する．そしてそれを参考に，ユーザの記事ブックマークや記事閲覧などの検索行動によって更新される探索者ビューを作成する．また，ブックマークした二つの記事が複数記事を間に持つ間接的なフォロー関係ならば，複数記事の中から 2 つの記事のどちらにも類似する記事を見落としている記事として推薦する．

キーワード　情報検索，デジタル図書館，自然言語処理，可視化，情報要約

1　はじめに

過去に起きた事件や出来事を調べる際にニュースアーカイブを利用することがある．大きな災害からの復興などの長期間の報道や，消費税増税などの大きく取り上げられた報道であれば，記事数が多く報道内容が多岐にわたる．その場合，報道の中でも自分が気になる話題を集中的に調べて，内容をまとめることが必要になってくる．

気になった話題のまとめを作成する場合，行う行動は大きく分けて 3 つある．まずはじめに，情報を集めるために複数回検索を行いながら多くの記事を読む．次に，ある程度情報が集まったら取捨選択を行い，自分の気に入った記事だけを残す．そして，気に入った記事の情報を一旦整理する．調べた情報を整理すると，まだ足りていない情報や，調べていく途中で気になった情報が出てくる．これらの情報について再び検索を行い，情報を取捨選択し，整理する．このように，検索，取捨選択，整理を何度も繰り返すことで，自分が気になった話題のまとめを作成することができる．

しかし，まとめを作成するにあたって非効率な検索を行わなければならない．例えば，同じ内容を重複して検索することや，記事間の直接的な移動ができないことが問題としてあげられる．

ある話題を調べているときに他の話題に興味が移ることがある．新しい話題へと途中で検索対象を変更する場合，現状の話題のまとめが中途半端な状態となる．これでは，新しい話題の検索がひと段落ついて元の話題について再び検索を始めようとした時に，自分がどこまで調べていたのかが分からなくなる．また，集めた記事同士の関係性はどうだったのかについても分からなくなってしまう．その結果，初めから記事を検索しなおしたり，すでに読んだ記事を再度読み直すことになる．

同じ内容について重複して検索する問題は，2 つの原因が挙げられる．1 つ目は，ブックマークしている記事同士の関係性をユーザが把握できていないことである．2 つ目は，ブックマークしている記事と閲覧中の記事との関係性をユーザが把握できていないことである．これは，検索時にブックマークした記事や閲覧している記事同士の関係を視覚的に分かりやすくユーザに見せることで，解決する問題だと考える．

また，記事アーカイブの検索には Web ブラウザ経由で検索システムを用いることが多くある．この時，一度閲覧した記事が再び気になって対象となる記事まで移動を行うことがある．その場合，ブラウザバックを複数回行って記事を移動するか，クエリを用いた検索によって移動する．つまり，記事移動するためにユーザは複数回の操作をする必要がある．

本研究では効率的な検索が行えるように，ユーザが自分の気に入った記事の関係性を確認しながら検索を行うことができるようになることを目的とする．そのために，記事の閲覧やブックマークで更新されるグラフである探索者ビューを見ながら記事検索を行える仕組みを提案する．

記事同士の関係をわかりやすく表現する方法として，記事間の関係をリンク構造として表現する研究は盛んにおこなわれている．しかし，それらの多くが，ユーザに提示している情報量が多すぎたり，ユーザが必要だとしない情報まで提示している．そのため，アーカイブ全体の記事同士の関係の大部分がユーザにとって不必要な情報であり，一見して必要な情報がどこにあるのかわからない．よって，あらかじめ作成しておいたアーカイブ記事グラフを参考に，ユーザの行動によって更新されるグラフである探索者ビューを作成する．図 1 のように，ユーザごとに必要とする情報だけ提示できると考えた．

探索者ビューを作成するために，ニュースアーカイブに存在するすべての記事に対してアーカイブ記事グラフを作成する．

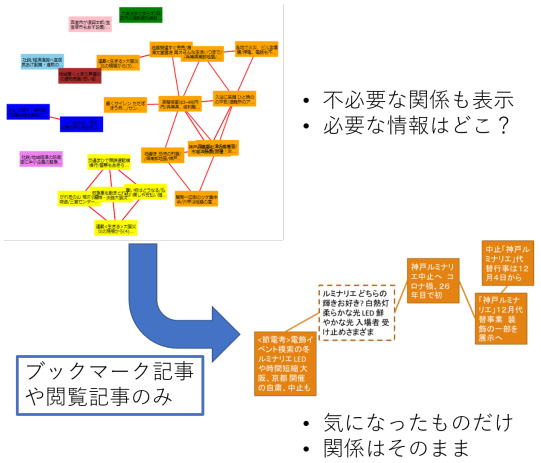


図 1　アーカイブ記事グラフから探索者ビューへのイメージ

ニュース記事では，新しい記事が追加情報を報じるために過去の記事の内容に関して言及することがある．この関係を本研究ではフォロー関係と呼び，新しい記事が過去の記事をフォローしていると表現する．アーカイブ記事グラフはノードが記事でエッジが記事のフォロー関係となる有向グラフである．

そして，ユーザの検索行動によって更新される探索者ビューを作成する．図 2 に「神戸ルミナリエ」について検索を行っている際の探索者ビューのイメージを示す．図 3 にユーザインタフェースのイメージを示す．探索者ビューはノードが記事でエッジが記事のフォロー関係となる有向グラフである．探索者ビューはブックマークした記事や閲覧中の記事などのフォロー関係を示す．

記事のブックマーク時に，すでにブックマークされている記事集合に新しくブックマークした記事を追加し，アーカイブ記事グラフを参考に探索者ビューを作成し提示する．これにより，ブックマーク記事同士の関係性を記事から離れても把握することができる．また，記事閲覧時に，閲覧中の記事とブックマーク記事のフォロー関係を確認し一時的に探索者ビューに追加する．これにより，閲覧中の記事がブックマークした記事集合において，どの立ち位置なのかについて確認しながら読むことができる．そして，探索者ビューの記事同士のフォロー関係の間に，アーカイブ記事グラフでは複数の記事が存在する場合，フォロー関係となる 2 つの記事のどちらにも類似している記事を複数の記事から推薦する．これにより，ユーザが気に入るはずだが見落としている記事を推薦することができる．

2　関 連 研 究

ユーザに記事同士の関係を提示する研究は多くされている．Thomas らは，ある記事中で過去の何かに言及しているならば，その何かに関する記事があるはずだと定義した [1]．そして，時系列を考慮した同じテーマ性のある記事同士がリンクしたネットワークの構築を提案した．Shahaf らは，二つのニュース記事が与えられると，それらを結びつける首尾一貫した記事の連鎖を見つける方法を提案した [10]．Liu らは，記事から現実世界の出来事であるイベントを自動的に抽出し，それを時系列を考慮してつなげることで記事同士をリンクさせることを提案した [5]．Hu らは，記事に 2 つ以上の都市が記載されている場合，都市と起こった出来事を抽出することで，都市という観点から記事ネットワークを構築した [2]．Koutrika らは，ユーザが学習などの目的で文書を見つける場合，ユーザが求める知識の一般的な内容の文書から専門的な内容の文書をツリー状に整理して提案した [3]．Xi らは，トピックが時系列に沿ってどのように進化し発展していくのかをユーザが理解するために，ディリクレ処理に基づく新しいトピックモデルを開発した [11]．

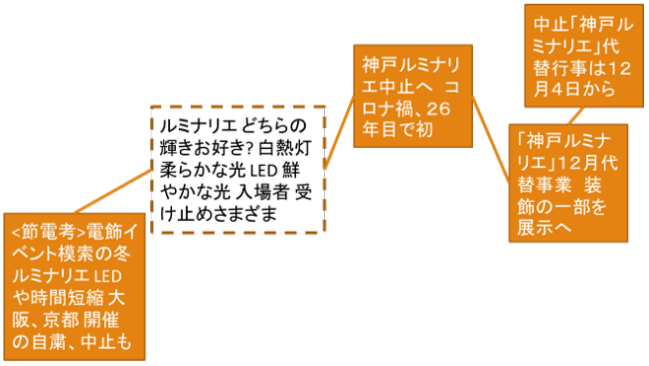


図 2　探索者ビューのイメージ

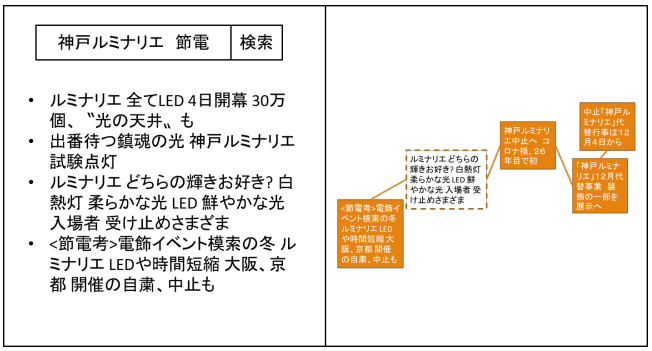


図 3　ユーザインタフェースのイメージ

これらの関連研究は，いずれもが大規模な記事データベースの記事を用いて記事同士の関係をグラフ化している．しかし，ユーザにとって必要なのは興味を持ったトピックのグラフの一部分だけであり，グラフの大部分は不必要となる．従って，アーカイブ記事全体のグラフを作成するために関連研究を参考にし，そこからユーザが必要とする情報だけを提示することを考えた．

3　アーカイブ記事グラフの作成

はじめに，ニュースアーカイブに存在するすべての記事に対してアーカイブ記事グラフを作成する．このグラフではノードが記事，エッジが記事のフォロー関係を表す．グラフの作成では Thomas らの方法 [1] を参考にする．

3.1　作 成 手 順

本研究ではニュース記事の関係性の一つとして，記事のフォロー関係を定義する． 3.2 節では，フォロー関係の具体例につ

東日本大震災で県、神戸市
公営住宅 300 戸提供廃棄物処理でも職員派遣 (2011/03/16)
... 広域緊急援助隊として 186 人を派遣した県警 ...

東北・関東大地震 県警、神戸市消防など
援助隊員を被災地派遣ボランティアらも支援 (2011/03/12)
兵庫県警は同日、広域緊急援助隊員として 166 人を岩手県に派遣。...

図 4　東日本大震災の緊急援助隊派遣についての記事

いて触れた後，記事の構成要素にどのようなものが存在するか，フォロー関係となる記事同士では記事の構成要素がどのような関係にあるかを述べる． 3.3 節では，実際の記事の構成要素をどのように抽出したかを述べる．アーカイブ記事グラフは，ノードが記事，エッジがフォロー関係となる有効グラフとして作成する． 3.4 節では，記事間の関連性を算出するための関連度を定義しフォロー関係となる記事を発見する．

3.2　記事構成要素と関係

記事同士の関係性として，ある記事中で過去の何かに言及しているならば，その何かに言及している記事があるはずである．例えば，図 4 は東日本大震災直後の緊急援助隊派遣について書かれた記事である．3 月 16 日の記事には兵庫県警が派遣した広域緊急援助隊について言及があり，3 月 12 日の記事はその広域緊急援助隊が結成され派遣された内容になっている．このようにニュース記事には，過去の記事の内容に対して新しい記事が新しい情報を追加するために言及することがある．本研究ではこのような関係を，新しい記事が過去の記事をフォローしている．フォロー関係であると呼ぶことにする．

3.2.1 では記事の構成要素について詳しく述べ， 3.2.2 では記事の構成要素から抽出された情報を用いて記事同士の関係について述べる．

3.2.1　記事構成要素

本論文では個々の記事は次の 4 つの構成要素を持っているとする．

- タイトル
- タイムスタンプ
- アブストラクト
- 段落

タイトルは記事のタイトル，タイムスタンプは記事の発行日である．アブストラクトは，記事本文の内容を簡単にまとめたものである．段落は，記事本文を構成する個々の段落である．

また，アブストラクトと段落は，次の 2 つを情報として持っている．

- 内容情報
- 時間情報

内容情報は段落を構成する文に含まれる名詞，動詞，形容詞の集合である．時間情報は段落を構成する文に含まれる曜日，日時，年などの時間である．

表 1 に，ある記事の 1 段落目，表 2 にその段落に含まれる内容情報，時間情報の例を示す．

表 1　「善意の物資ありがとう/お米など何とかメド/欲しいのは本，工具/備蓄基地/スペース確保課題に/阪神・淡路大震災/(1995/01/27)」の第 1 段落

大地震から十日余りがたち，兵庫県災害対策本部に全国から寄せられる救援物資と，被災者の要望の間にミスマッチが生じ始めている．乾パン，米，毛布などはぼ充足したとみられる品が，今もなお続々と到着し，備蓄基地は窮屈になる一方．担当者らは「せっかくの善意だが...」と少々，困惑気味だ．

表 2　表 1 の記事の第 1 段落の内容情報，時間情報

内容情報	地震，十，たつ，兵庫，災害，対策，本部，全国，寄せる，救援，物資，被災，要望，間，ミスマッチ，生ずる，始める，乾パン，米，毛布，ぼ，充足，する，みる，品，到着，する，窮屈だ，一方，担当，善意，困惑
時間情報	大地震から十日余りがたち

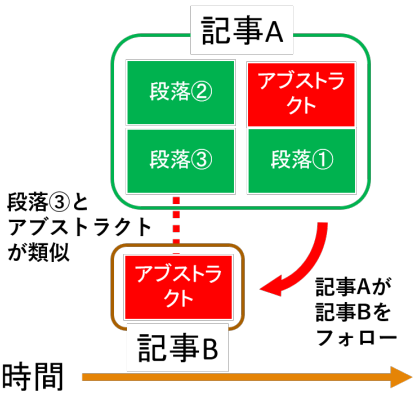


図 5　同じ時間情報を持ち類似するアブストラクトと段落

3.2.2　記事同士の関係

フォロー関係となる記事同士の記事構成要素は，次の 2 つの条件を満たすものとする．1 つ目は，フォローされる記事のアブストラクトの時間情報とフォローする記事の段落の時間情報が一致することである．2 つ目は，フォローされる記事のアブストラクトの内容とフォローする記事の段落の内容が類似することである．

図 5 は，記事同士がリンクしている様子を表したものである．ある出来事を報道した記事 B と，その出来事についての続報となる記事 A があるとする．記事 A は記事 B をフォローしている関係にある．この時，記事 A は出来事について詳しく説明している記事の本文中で，既に報道された記事 B の内容について触れている．図 5 では段落 3 が記事 B について言及した段落とした．同じ出来事を指す記述であれば，時間情報は一致し，内容情報は類似する．従って，フォロー関係と前記の 2 つの条件は必要十分の関係であると考えた．

3.3　記事構成要素の抽出

フォロー関係で構成されるアーカイブ記事グラフの作成のために，記事ごとの段落とアブストラクトからそれぞれの情報を抽出する．

表 3　時間情報の種類

直接的な時間情報	間接的な時間情報
1995 年 1 月 17 日	先週の火曜日
平成 20 年	先週末
3 年前	一昨日

表 4　間接的な時間情報の語例

間接的な時間情報
一昨日，昨日，先週， 先月，去年，前年， 日曜日，月曜日，火曜日…

記事データセットは，神戸新聞から提供されたの 1995 年 1 月から 2019 年 8 月までの 1,361,665 件の記事を使用する．

3.3.1　内容情報の抽出

内容情報の抽出は，段落やアブストラクトを構成する文章から，名詞，動詞，形容詞を抽出して内容情報とする．文章の形態素解析には，Juman++ [6] [12] を用いる．また，形態素解析の結果，品詞細分類が数詞の語をストップワードとする．

3.3.2　時間情報の抽出

時間情報の抽出は，文書の中から曜日，日時，年月などの時間表現を抽出し，1 日単位の数値として規格化する．

このような時間情報には表 3 に示すように 2 つの種類があると考える．まず 1 つ目が，時間表現に数字が用いられる直接的な時間情報とする．そして 2 つ目が，時間表現に数字を用いない間接的な時間情報とする．それぞれの時間情報の抽出には異なる手法を用いる．

直接的な時間情報の抽出には数量表現・時間表現の規格化を行うツールである normalizeNumexp [7] を用いて，間接的な時間情報の抽出には独自にルールベースで時間表現の規格化を行った．

間接的時間情報は，数値情報を含まないため normalizeNumexp で時間情報を抽出することができない．そのため，間接的な時間情報はルールベースで抽出する．ニュース記事では読者に確実に情報を伝えるため，婉曲な間接的な時間表現はあまり使われない．よって，使用されている間接的な時間情報の表現は表記ゆれしない簡潔なものだと考える．これらは簡潔ゆえに種類が少なく，ルールベースで時間表現の規格化ができる．表 4 に示した語を用いて図 6 に示すような処理を行った．

まずは記事中から文節単位で間接的な表現を抽出する．図では 2021 年の 1 月 17 日に発行された記事に「先週の日曜日」という間接的な時間情報が文節単位で存在するとする．次に文節単位の間接的な時間情報を単語単位に分解する．「先週の日曜日」であれば「先週」と「日曜日」に分解する．最後に記事発行日を基準とした単語単位の時間情報が重なる部分を時間情報として数値化する．図では「先週」にあたる 1 月 4 日から 1 月 10 日までの期間と，「日曜日」にあたる 1 月 10 日から，2021 年 1 月 10 日を間接的な時間表現で抽出した時間情報とした．

3.3.3　アブストラクトの抽出

段落と比較する記事の要素であるアブストラクトの抽出を行う．アブストラクトは記事の概要を簡単に説明した文章である．

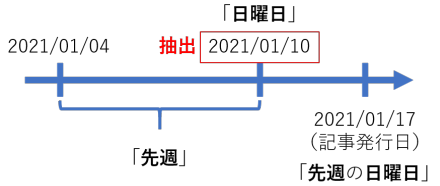


図 6　「先週の日曜日」の時間情報抽出例

また，記事本文の前半には報道を要約したリードと呼ばれる文章が存在する．したがって，記事におけるリードの部分をアブストラクトとして用いる．

3.4　記事間のフォロー関係の発見

すべての記事の要素が抜き出せたら，それを用いて記事間のフォロー関係の発見を行う． 3.2.2 節では記事間のフォロー関係に方向性があることを示した．アーカイブ記事集合を A とし，アーカイブ記事グラフ G は式 1 のように定義する．これはノードが記事，エッジが記事間のフォロー関係となる有向グラフである．

$$G := (A, E)$$
$$E \subset A \times A \tag{1}$$

このフォロー関係を発見するために，記事間の関連度を求める．関連度は，フォロー関係となる記事同士の候補を見つけ，それぞれの候補のアブストラクトと段落の類似度を求め，その最大を関連度とする．記事 $a \in A$ と $a' \in A$ の関連度がしきい値を超える場合，a から a' に向かってエッジが張られる．

まずは，フォロー関係となる記事の候補を見つける．アーカイブ記事集合の中で同じ時間情報を持つアブストラクトの記事と段落の記事を，それぞれフォローされる記事とフォローする記事の候補とする．

次に，フォロー関係となる記事の候補の関連度算出のために，アブストラクトと段落の類似度算出を行う．本研究では Word Mover's Distance（WMD）を用いて計算したアブストラクトと段落の文章間距離を用いて類似度を計算した [4] [8] [9].

WMD は文章間の距離を単語の分散表現を用いて計算する方法である．2 つの文章のすべての単語において，一方の文章の単語からもう一方の文章への単語までの最短距離を計算する．

一般に，WMD では語の距離計算にユークリッド距離を採用している．式 2 は，N 次元でベクトル化された語 $\mathbf{w_1}, \mathbf{w_2}$ のユークリッド距離を示している．しかし，これでは WMD の最大値を設定することができない．本研究では語のベクトルの長さを 1 に正規化したベクトルを用いてユークリッド距離を算出する．また，文章 c の特徴ベクトルを $\mathbf{D_c}$ と定義する．$\mathbf{D_c}$ の単語 w の次元に対応する値は，文章 c 中で単語 w が出現した回数を文章 c の総単語数で割ったものである．これらを用いて WMD を式 3 のように定義する．c は文章，$\mathbf{T}$ は文章間の変換行列，k_c は文章の内容情報である．

比較する 2 つの文章に含まれる語数が異なる場合，語同士を

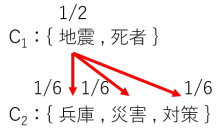


図 7 「地震」が c_2 の各語に等しい重みで変換されている様子

1 対 1 対応することができない．任意の語数の文章同士の比較を行うために，一方の文章の語から他方の文章の語へとどのような重みで変換したかを表したものを変換行列と呼び $\mathbf{T}$ で表す．例えば，$c_1 =$ { 地震, 死者 } と $c_2 =$ { 兵庫, 災害, 対策 } という語集合を 2 つの文章を比較する場合を考える．2 つの文章の特徴ベクトルは式 4 と式 5 のようになる．2 つの文章の語が他方に含まれる語にすべて等しい重みで変換できる場合，その変換行列は式 6 のように $\mathbf{T_{c_{12}}}$ で表せる．図 7 は c_1 の「地震」が c_2 の各語に等しい重みで変換されている様子を表したものである．この重みを様々な値に変化させ，式 3 で表すように距離が最も小さくなるような変換行列を $\mathbf{T}$ とする．

$$d(\mathbf{w_1}, \mathbf{w_2}) = \sqrt{\sum_{n=1}^{N} (w_{1_n} - w_{2_n})^2} \tag{2}$$

$$wmd(c_1, c_2) = \min_{\mathbf{T} \geqq 0} \sum_{i,j=1}^{n} \mathbf{T}_{i,j} d(k_{c_1 i}, k_{c_2 j}) \tag{3}$$

$$(0 \leqq wmd \leqq 2)$$

$$\mathbf{D_{c_1}} = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{array}{ccccc} \text{地震} & \text{死者} & \text{兵庫} & \text{災害} & \text{対策} \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{4}$$

$$\mathbf{D_{c_2}} = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{array}{ccccc} \text{地震} & \text{死者} & \text{兵庫} & \text{災害} & \text{対策} \end{array} \\ \begin{pmatrix} 0 & 0 & 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array} \tag{5}$$

$$\mathbf{T_{c_{12}}} = \begin{array}{c} \\ \text{地震} \\ \text{死者} \\ \text{兵庫} \\ \text{災害} \\ \text{対策} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{地震} & \text{死者} & \text{兵庫} & \text{災害} & \text{対策} \end{array} \\ \begin{pmatrix} 0 & 0 & 1/6 & 1/6 & 1/6 \\ 0 & 0 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 0 & 0 & 0 \\ 1/6 & 1/6 & 0 & 0 & 0 \\ 1/6 & 1/6 & 0 & 0 & 0 \end{pmatrix} \end{array} \tag{6}$$

式 3 を用いて，フォロー関係となる記事の候補のアブストラクトと段落の類似度を式 7 のように定義する．

$$sim(c_1, c_2) = 1 - \frac{wmd(c_1, c_2)}{2} \tag{7}$$

$$(0 \leqq sim \leqq 1)$$

フォロー関係となる記事の候補のアブストラクトと段落の類似度を求め，その中でも最大の類似度をフォロー関係となる候補記事同士の関連度とする．式 8 にフォロー候補記事間の関連度の定義を示す．a はフォローする記事，a' はフォローされる記事，$Par(a)$ は a の段落集合，$abst(a')$ は a' のアブストラクトを表す．フォロー関係を求める関連度計算には p_i と $abst(a')$ が同じ時間情報をもつものを使用した．時間関係を考慮する関連度計算を rel と表す．関連度にはしきい値を設け，しきい値以上となるフォロー候補記事がフォロー関係であるとする．

$$rel(a, a') := \max_{p_i \in Par(a)} sim(p_i, abst(a')) \tag{8}$$

4 探索者ビューの作成

調べるトピックが異なればブックマークする記事や閲覧する記事が異なる．そのため，ユーザが求める記事同士の関係はユーザごとに異なる．本研究では，ユーザによって表示されるグラフが異なりその後の検索行動によって随時グラフがアップデートされる，探索者ビューを提案する．これは第 3 章で作成したアーカイブ全体のアーカイブ記事グラフをもとに，ユーザのブックマーク記事集合や閲覧記事から，ユーザが必要とする記事同士の関係のみを見せるグラフである．

探索者ビューは，記事をブックマークしたり，記事を閲覧したりすることによってグラフを更新する．記事ブックマーク時に起こる探索者ビューの更新を 4.1 節で，記事閲覧時に起こる探索者ビューの更新を 4.2 節で述べる．そして，記事検索時に見落とし記事についての推薦を探索者ビュー上で行う．記事検索時に起こる探索者ビューの更新を 4.3 節で述べる．ユーザからは自らの検索行動でグラフが作成され更新されていくように見せる．

4.1 記事ブックマーク時の処理

ユーザは検索中に気に入った記事をブックマークする．従来の検索システムでは記事情報のみが保存され，後で見返した場合に記事同士の関係が分からず，再び記事を読んで関係性を把握する手間が発生する．そこで，記事ブックマーク時は，ブックマーク記事集合に含まれる記事がどのような関係であるかについて可視化する．これで，気に入った記事同士がどのような関係を持っているのかを常に確認しながら検索を行うことができる．

ブックマーク記事集合を A_{bm} とすると，探索者ビュー SV は式 9 のように定義できる．これは G と同じくノードが記事，エッジが記事間のフォロー関係となる有向グラフとなり，図 2 のようなグラフを作成する．A_{bm} は A の部分集合であり，検索開始時において A_{bm} は空集合である．記事 $a_i \in A_{bm}$ と $a_j \in A_{bm}$ の関連度がしきい値を超える場合，a_i から a_j に向かってエッジが張られる．また，a_i と a_j の関連度がしきい値を超えていない場合でも，図 8 のように G において記事 $a_k \in A$ を介して a_i と a_j がつながっているならばエッジを張る．このため E_{bm} は E に含まれないエッジを含む場合がある．ブックマークへの記事追加や削除が行われるたびに探索者ビューが更新される．

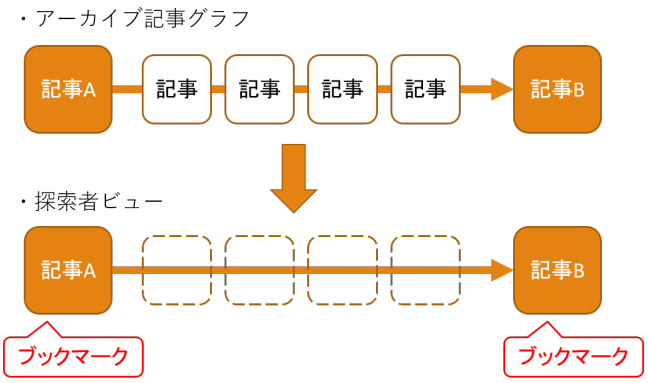


図 8　探索者ビューではブックマーク記事のみ表示

$$SV := (A_{bm}, E_{bm})$$
$$E_{bm} \subset A_{bm} \times A_{bm} \tag{9}$$

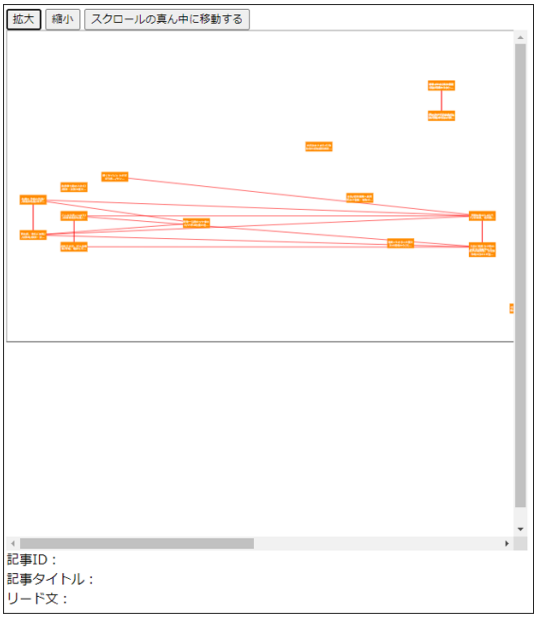
図 9 はブックマーク記事集合を用いて探索者ビューを作成した例である．まず，探索者ビューは拡大縮小が自由に行える．ユーザは図 9 の上部のように記事のつながりの概観や，図 9 の下部のように詳細な記事のつながりを自由に見ることができる．次に，探索者ビューの左から右にかけて記事が新しくなるように，ノードを時間関係を考慮して配置する．ユーザは記事同士の時間関係を容易に把握することができる．また，ノードはドラッグ操作で任意の場所に移動させることができる．ノードを離すとノードは時系列を考慮した位置に自動で移動する．ブックマークした記事が増えるとノードやエッジが増えて，エッジの上にノードが重なることがある．見にくくなった探索者ビューであってもノードを移動することで，関係性を見やすくすることができる．そして，ノードをクリックすると記事を一意に決める記事 ID，記事のタイトル，記事のリード文を表示する．ユーザは一度ブックマークした記事をもう一度確認したい場合，その概要について容易に確認することができる．

4.2　記事閲覧時の処理

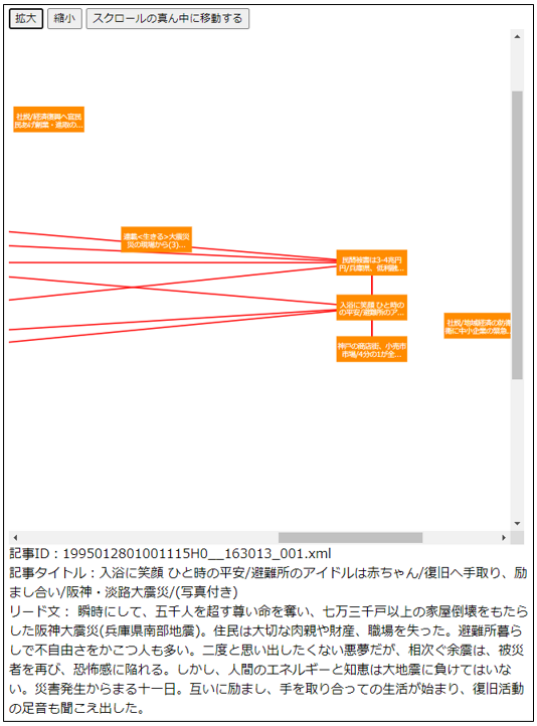
記事閲覧時，従来の検索システムではユーザは記事を読み終わってから他の記事との関係性を考えなくてはならない．そこで，閲覧記事が探索者ビューにおいてどの記事と関係があるかを，記事を読む前に確認することで記事の理解がしやすいと考える．そのために，図 10 のように閲覧中の記事を一時的に探索者ビューに追加することで閲覧記事の理解を助ける．図 10 の赤のノードが一時的に探索者ビューに追加された閲覧中の記事である．

閲覧中の記事を a_v とすると，A_{bm} と E_{bm} は一時的に更新される．A_{bm} の記事集合には a_v が一時的に追加される．E_{bm} は A_{bm} が更新されたことによってエッジが新たに追加される．4.1 で示した E_{bm} と同じく，a と a_v もしくは a_v と a において，関連度がしきい値を超える場合か G において $a_k \in A$ を介して繋がっている場合，新たにリンクを張る．

ブックマークせずに記事を移動すれば，A_{bm} と E_{bm} から一時的に追加していた a_v とそのエッジを削除する．その場合，探索者ビューから閲覧していた記事のノードとエッジを削除して，記事閲覧前の探索者ビューの状態に戻す．



縮小して全体像をとらえる



拡大して詳細につながりを確認

図 9　実際の探索者ビュー

4.3　見落とし記事の推薦

大規模な記事アーカイブで検索する場合，記事閲覧候補が多すぎてユーザが気に入るはずの記事を見落とすことがある．探索者ビューではブックマークした記事や閲覧中の記事のみを表示させるため，図 8 のような記事アーカイブグラフでは経路に含まれるが探索者ビューでは表示しない間接的なフォロー関係が発生する．間接的なフォロー関係の途中に含まれる記事集合

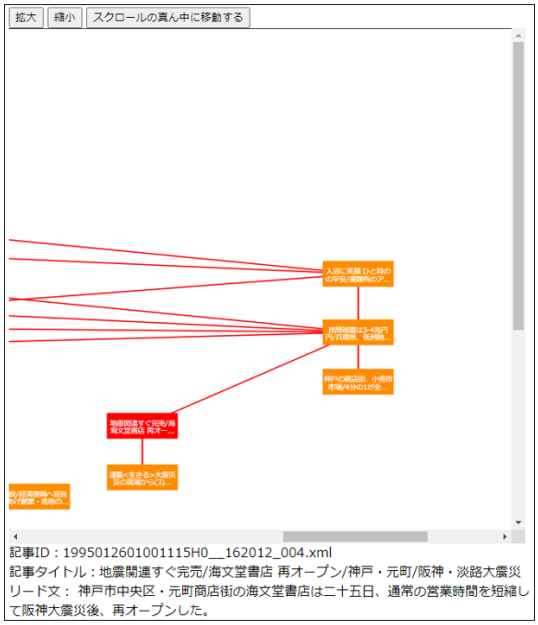


図 10　記事閲覧時のイメージ

は，ユーザが興味を持った記事同士に挟まれているので，ユーザが興味を持ちそうな記事の集合だと考えられる．そこで，記事検索時に間接的なフォロー関係に含まれる記事集合の中から両端の記事に類似するものを推薦することで，ユーザが興味を持ちそうな記事の見落としを防ぐ．図 11 の緑のノードのように推薦する記事を一時的に探索者ビューに追加することで見落とし記事の推薦を行う．

記事推薦を行うために，2 つの記事の間接的なフォロー関係に含まれる記事集合の記事の推薦度を定義する．間接的なフォロー関係である 2 つの記事を a_1, a_2 とし，集合に含まれる記事 a_c の推薦度 $reco$ を式 10 に示す．この推薦度の定義に含まれる関連度では記事のアブストラクトと段落の時間関係は考慮しない．時間関係は考慮しない関連度を rel^* と表す．集合のすべての記事に対して推薦度を求めて，推薦度が最大のものを推薦する．

推薦する記事は一時的に探索者ビューに加える． 4.2 節の一時的な更新と同じように，A_{bm} の記事集合には a_c が一時的に追加される．また，E_{bm} は A_{bm} が更新されたことによって $\{(a_1, a_c), (a_c, a_2)\}$ のエッジが新たに追加される．ユーザが推薦記事もしくはその他の記事の閲覧を行った場合，A_{bm} と E_{bm} から一時的に追加していた a_c とそのエッジを削除する．

$$reco(a_c|a_1, a_2) = \frac{rel^*(a_1, a_c) + rel^*(a_c, a_2)}{2} \qquad (10)$$

5　実　行　例

実際に検索を行った例を図 12 に探索者ビューを示しながら述べる．この実行例では，ユーザが調べるトピックとしてとして，阪神・淡路大震災発生時の商店街の様子について調べているものとする．ユーザは既に，気になった記事として 1 つ目の記事をブックマークしているものとする．1 つ目の記事は「地

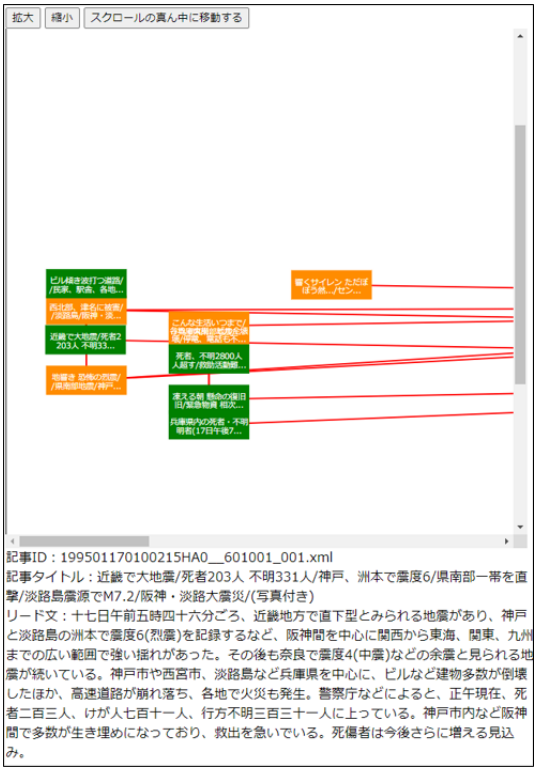


図 11　探索者ビューによる見落とし記事の推薦

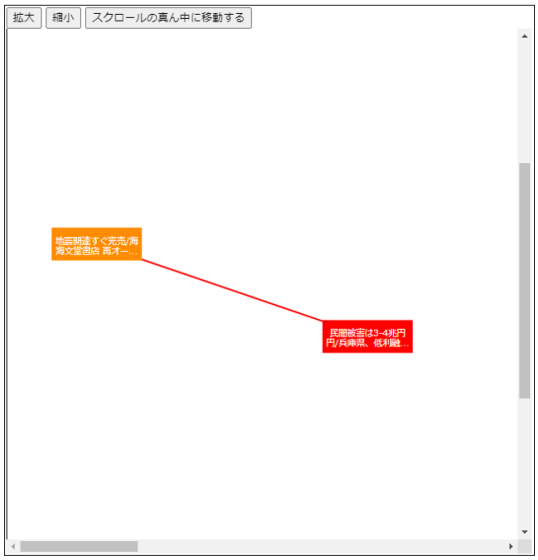
震関連すぐ完売/海文堂書店 再オープン/神戸・元町/阪神・淡路大震災」という記事であり，神戸市中央区の元町商店街にある海文堂書店が阪神・淡路大震災後，再オープンした内容の記事である．そして図 12 の上部で，2 つ目の記事を閲覧しているときの探索者ビューを示している．2 つ目の記事は「民間被害は 3-4 兆円/兵庫県，低利融資を検討/阪神・淡路大震災」という記事であり，兵庫県が阪神・淡路大震災で被災した県下の民間企業に対し新しい融資制度の検討を始めた内容の記事である．この 2 つの記事は，2 つ目の記事が 1 つ目の記事に対して間接的なフォロー関係にあり，探索者ビューでは 2 つの記事ノードにエッジが張られる．ユーザは 2 つ目の記事の本文を読む前に，閲覧中の記事がブックマークした記事とフォロー関係にあることを知り，関係性を意識しながら読むことができる．

そして 2 つ目の記事をブックマークすると記事検索画面では，図 12 の下部のように記事推薦が起きる．3 つ目の記事として「転居先足りず/被災者ほん走/阪神・淡路大震災」が推薦され，家をなくした被災者が役所に仮設住宅への転居を申し込んでいる内容となっている．1 つ目と 2 つ目の記事の，復興に向けての話題と公的機関の対処の話題が考慮されている記事が推薦されていると考える．

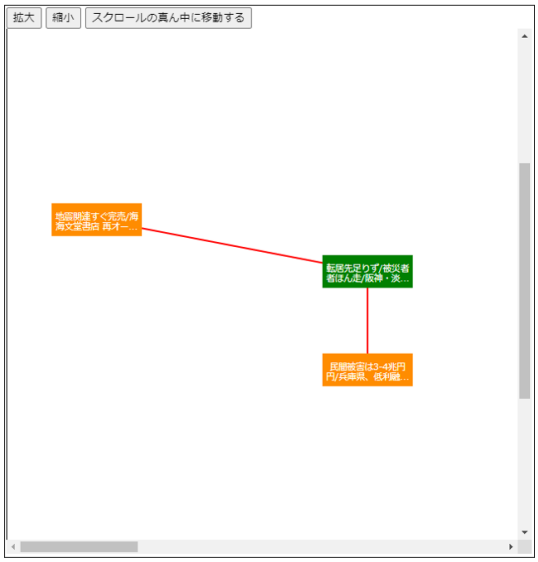
このように，推薦記事を参考にしながら記事の閲覧とブックマークを繰り返すことで，ユーザ独自の気に入った記事集合の関係性をグラフ化することができる．

6　ま　と　め

本研究では，ニュースアーカイブ探索のためのインタフェースとして記事の閲覧やブックマークで更新されるグラフである



記事閲覧時



記事検索時

図 12　実　行　例

探索者ビューを見ながら記事検索を行える仕組みを提案した．探索者ビューはアーカイブ記事全体をグラフとしたアーカイブ記事グラフから，必要とする記事の部分だけユーザに見せるグラフである．まずは，アーカイブ記事グラフを作成するため，記事間の関係性としてフォロー関係について述べ，ノードが記事でエッジが記事間のフォロー関係となるアーカイブ記事グラフの作成について述べた．次に，アーカイブ記事グラフ作成のために，記事の構成要素とその抽出を述べ，フォロー関係となる記事同士の構成要素の関係について述べた．そして，記事間のフォロー関係の発見のためにフォロー候補記事の関連度について定義した．探索者ビューはユーザの検索行動によってグラフが更新される．記事ブックマーク時には新規でブックマークされた記事と既存のブックマーク記事の関係性をグラフに追加する．記事閲覧時には閲覧中の記事と既存のブックマーク記事との関係性を一時的にグラフに追加する．記事検索時にはユーザが見落としていると考えられる記事を推薦し，既存のブックマーク記事との関係性を一時的にグラフに追加する．最後に簡単な実行例を示し，提案手法の有用性について述べた．

謝　　辞

本研究の一部は JSPS 科学研究費助成事業 JP18H03494, JP18H03243 による助成を受けたものです．ここに記して謝意を表します．加えて，データをご提供いただいた神戸新聞社，特にメディアビジネス局企画推進部の豊川聡氏，峯大二郎氏に深く感謝致します．

文　　献

[1] Thomas Bögel and Michael Gertz. Time will tell: Temporal linking of news stories. In *Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries*, pp. 195–204, 2015.

[2] Yingjie Hu, Xinyue Ye, and Shih-Lung Shaw. Extracting and analyzing semantic relatedness between cities using news articles. *International Journal of Geographical Information Science*, Vol. 31, No. 12, pp. 2427–2451, 2017.

[3] Georgia Koutrika, Lei Liu, and Steve Simske. Generating reading orders over document collections. In *Proceedings of IEEE 31st International Conference on Data Engineering*, pp. 507–518, 2015.

[4] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *Proceedings of International conference on machine learning*, pp. 957–966, 2015.

[5] Yaopeng Liu, Hao Peng, Jie Guo, Tao He, Xiong Li, Yangqiu Song, and Jianxin Li. Event detection and evolution based on knowledge base. In *Proceedings of Knowledge Base Construction, Reasoning and Mining 2018*, 2018.

[6] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 2292–2297, 2015.

[7] Katsuma Narisawa, Yotaro Watanabe, Junta Mizuno, Naoaki Okazaki, and Kentaro Inui. Is a 204 cm man tall or small? acquisition of numerical common sense from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 382–391, 2013.

[8] Ofir Pele and Michael Werman. A linear time histogram metric for improved sift matching. In *Proceedings of European conference on computer vision*, pp. 495–508, 2008.

[9] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *Proceedings of IEEE 12th International Conference on Computer Vision*, pp. 460–467, 2009.

[10] Dafna Shahaf and Carlos Guestrin. Connecting the dots between news articles. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–632, 2010.

[11] Yaoyi Xi, Gang Chen, Bicheng Li, and Yongwang Tang. Topic evolution analysis based on cluster topic model. *Journal of Advanced Computational Intelligence and Intelligent Informatic*, Vol. 20, No. 1, pp. 66–75, 2016.

[12] 森田一, 黒橋禎夫. RNN 言語モデルを用いた日本語形態素解析の実用化. 情報処理学会第 78 回全国大会講演論文集, Vol. 2016, No. 1, pp. 13–14, 2016.