過去を参照するマイクロブログに関する分析

澄川　靖信† Adam Jatowt†† Marten Düring†††

† 東京理科大学理工学部情報科学科　〒 278–8510 千葉県野田市山崎 2641
†† 京都大学大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
††† University of Luxembourg, Institute for Contemporary History　Campus de Belval Maison des Sciences Humaines 11, Porte des Sciences, L-4366 Esch-sur-Alzette, Luxembourg
E-mail: †yas@cs.is.noda.tus.ac.jp, ††adam@dl.kuis.kyoto-u.ac.jp, †††marten.during@uni.lu

あらまし　本研究では，過去を参照するツイートを収集し，マイクロブログのユーザはどのような歴史を参照するのかを分析する．本稿では，データの収集方法として，1) 歴史に関連するハッシュタグによるツイートデータの検索，2) ハッシュタグの共起性に着目した新しい歴史的なハッシュタグ収集のためのブートストラップ法，の 2 種類の手法を用いる．これらのデータに対して，人々がいつ，どのような歴史を，何がきっかけで参照するのかを分析する．これらの分析結果は，あらゆる歴史について知識を増やし，理解を深めることを目的とする新たな情報推薦システムの実現といった時間指向の新たな情報検索の基礎となる効果が期待できる．

キーワード　ソーシャルメディア解析，集合的記憶解析，歴史，Twitter

1. はじめに

歴史に関する良い知識と理解は，現代社会の形成課程をより良く理解することや，様々な地域や時代で生じた事柄を類推して現代社会の諸問題について考察するための足場かけとなる効果があるので [1], [10]，社会的に重要な役割を果たしている．実際，歴史の授業は多くの国で小学校から開講されている基礎的な科目の 1 つであることや，歴史的類推を促すための学習デザインに関する研究 [23] が行われていることから，歴史の重要性は広く認識されている．

近年，情報科学の手法を用いて歴史を解析するヒストインフォマティクス（HistInformatics）の研究として，ニュース [2], [6] や Wikipedia [8], [17] を用いた研究が行われている．このようなメディアと同様に，マイクロブログも歴史に関する情報を共有するためにも用いられるが，マイクロブログを用いた研究の多くは，アメリカ大統領選挙のようなリアルタイムに生じているイベントへの社会的な関心を分析する傾向がある [22]．ヒストインフォマティクスに関係するものとして，第一次世界大戦に焦点を当てたものや [5]，国家間の伝統文化の比較といった研究はわずかに行われているが，大規模解析に基づいた分析は行われていない．

本研究では，次のリサーチクエスチョンに基づいて過去を参照するツイートを分析する．

- 人々はどのようにマイクロブログ上で歴史を参照するのか
- どの期間の歴史がよく参照されるのか
- 何がきっかけでマイクロブログ上で歴史を参照するのか
- 集合的記憶は Twitter 上でどのように表現されているのか

本稿では，2016 年 3 月から 2017 年 1 月の間に，130 個の歴史に関するハッシュタグと，それらのハッシュタグを含む 772,137 件のツイートを収集した．これらのデータに対して，時間，URL，固有表現，ハッシュタグに着目して分析する．また，歴史的なハッシュタグを分類するために，各ハッシュタグの意味を専門家が考慮して行った分類結果と，ハッシュタグと一緒に使用される固有表現に基づいたクラスタリングによるグループ分けの結果を示す．以上の分析によって，歴史的なイベントや固有表現に関する情報を構造化でき，次のような新たな情報検索や学習環境の実現に向けた研究の基礎となることが期待できる．

（1） 現代社会がどのように形成されたのか，という疑問に対する解を見出すための新たな知見．

（2） 歴史的な知識の共有を促進するための推薦システムの実現．このような推薦システムによって，日頃から現在や未来に関する膨大な量の情報に触れ，自分自身の国や世界の歴史に関する知識が不足している人々を対象とした，歴史への興味を促進するためのプロジェクト (注1)(注2)(注3)(注4) を更に発展させられることが期待できる．

（3） 歴史的な話題をつぶやいたり，会話したりする自動コンテンツ作成の実現．

本稿の貢献をまとめると，次の通りである．

（1） 過去を参照するマイクロブログ上で大規模解析を行い，一般的に，ユーザがどのような過去を参照するのかを分析する．

（2） マイクロブログでどのように集合的記憶が維持・形成されているのかをより良く理解するための新たな知見を明らかにする．

（3） 歴史的なハッシュタグに関する新しい分類法を提案する．

(注1)：https://twitter.com/RealTimeWWII
(注2)：https://twitter.com/civilwarwp
(注3)：https://twitter.com/1948War
(注4)：https://twitter.com/samuelpepys

2. 関 連 研 究

Halbwachs によって一般化された集合的記憶（社会的記憶）の概念は，社会的グループ内で共有されている過去を反映しているもの，と定義されている [12], [13]．集合的記憶の対照となる概念として，Jacoby によって定義された，不快なイベントに関する記憶を強制的，あるいは無意識的に抑制する集合的記憶喪失がある [15]．個人的記憶と同様に [7]，時間の経過とともに集合的記憶も薄れ，突発的なイベントや記念日といったきっかけによって思い出されることが知られている [2], [17], [18]．このような研究は，記憶や忘却のメカニズムを理解するためだけでなく，現代社会の中で過去の役割を説明するためにも重要である．

集合的記憶に関する研究は，アーカイブ選択を意味する [18]．多くの研究では，一人一人のアカウントデータや政治的・文化的な組織ごとの活動記録といった小規模なデータを用いる．大規模なデータセットに対して情報科学の手法を適用した研究はあまり多くはない．Cook らは，20 世紀に発行された新聞記事を用いて，時間の経過と名声の失墜の関係を調査した [6]．Au Yeung らは 90 年分の英語版の新聞記事をデータセットとして，どのように過去が記憶され，どのように忘れられていくのかを調査した [2]．Jatowt らは人々が参照する過去と未来を可視化する枠組みを実現した [16]．Wikipedia を使用して，忘れられていたイベントが思い出されるきっかけを調査した研究も行われている [8], [17]．

3. データ収集と統計情報

本節で過去を参照するツイートの収集方法と，解析に用いるデータセットの統計情報を示す．

3.1 ハッシュタグに基づくツイート収集

歴史的なイベントや固有表現に強く関連するツイートを集めるために，本研究ではハッシュタグに基づくクローリングを行う．まず最初に，歴史に関連するハッシュタグを研究している専門家によって選別されたハッシュタグ（`#history`, `#HistoryTeacher`, `#WmnHist` など）を収集した(注5)．また，過去を参照する際に広く使用されている 7 個のハッシュタグ（`#throwbackthursday`，`#historyrepeating`，`#historicalevent`，`#thisdayinhistory`，`#otd`，`#onthisday`，`#timetravel`）も収集した．これらのハッシュタグを検索ワードとして，Twitter 社が提供している公式 API(注6) を用いてツイートを収集した．以降，歴史に関係するツイートを集めるために使用するハッシュタグを**シードハッシュタグ**と呼ぶ．

上記の方法によるツイートの収集を，2016 年 3 月 8 日から 2017 年 1 月 5 日まで実施した．またツイートの収集と同時に，シードハッシュタグの数を増やすために，シードハッシュタグと共によく使用され，未だシードハッシュタグとして扱われていない歴史的なハッシュタグをブートストラップ法によって検索し，専門家による検査の後にシードハッシュタグとして追加した．最終的に，本稿で用いるデータセットは，130 個のシードハッシュタグと，772,137 件のツイートデータを含む．シードハッシュタグの一部を，手動による分類結果として表 6 に示す．

3.2 データ処理

時間に着目した解析を行うために，**タイムスタンプ**と**時間参照**の 2 つの時間表現を定義する．本稿では，タイムスタンプをツイートが投稿された日時，時間参照をツイート中で言及された時間表現，とそれぞれ定義する．さらに，時間表現には**明示的**なものと**暗示的**なものの 2 種類がある [4]．明示的な時間表現とは,「1945 年 8 月 7 日」や「2009 年」のような時間軸上のある点や期間の絶対的な時間を表す．一方，暗示的な時間表現とは,「昨日」や「2 年前」といったある時点からの相対的な時間を表す．本稿では，あらかじめ，タイムスタンプを用いて暗示的な時間表現を明示的なものへと変換されていると仮定する．

Twitter には，リツイートと引用ツイートの 2 種類の特別なツイートがある．いずれのツイートも他のツイートを共有することを目的とし，リツイートは元のツイートと全く同じものを自身のタイムラインに再ポストするものであり，引用ツイートは新しい情報を付与してポストするものである．以降の解析では，どのような過去の情報が生成されるのかに着目して解析するために，時間参照について分析する 4.1 節のみでリツイートも使用し，その他の解析ではツイートのみを用いる．

時間表現の抽出は，ツイート処理に特化したオプションを提供している Heideltime [19] を用いた．Heideltime は時間表現を標準化した後，その結果を言語資源アノテーションの国際標準の 1 つである TimeML の TIMEX3 タグを用いて出力する．本稿で用いたデータセットの中には，6/11/16 や 3/19/88 といった簡略された時間表現を用いているものが存在している．Heideltime はこのような時間表現に対して”00”を年情報の先頭に付与する．本稿では，この修正が行われた年情報の下 2 桁が，17 以下なら”00”を”20”に，さもなければ”19”に変更した．

以上の処理を行い，次節以降の解析に用いるデータセットの統計情報を表 1 に示す．また，図 1 に，本データセットに含まれるツイートの使用言語の割合を示す．約 83%のツイートは英語で記述されていることがわかる．

表 1　データセットの統計情報.

総ハッシュタグ数	110,553
歴史的なハッシュタグ数	130
総ツイートデータ数	772,137
ツイート数	276,322
リツイート数	495,815
タイムスタンプの期間	2016/3/8 – 2017/1/5
時間参照の期間	8156 BC – 2029
時間参照を含むツイート数	88,309
URL 数	176,598
URL を含むツイート数	170,237
総ユーザ数	94,988

4. 一般的な解析

4.1 時 間 解 析

図 2 に，本稿で収集したツイートとリツイートのそれぞれに含まれる時間参照の分布を示す．2 つの分布の両方で，第一次

(注5)：http://blog.historians.org/2013/08/history-hashtags-exploring-a-visu\
\al-network-of-twitterstorians/

(注6)：https://dev.twitter.com/rest/public

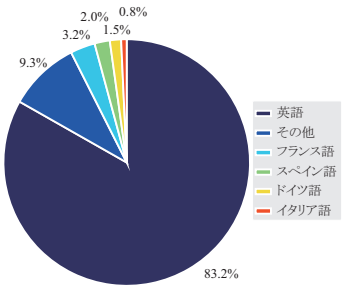

図 1　収集したツイートの使用言語上位 5 件.

世界大戦と第二次世界大戦に関する年と本稿のデータセットのタイムスタンプがほぼ網羅する年（2016 年）の 4 つのピークがある．また，2 つの分布の両方で現在に近づくにつれて時間参照の数が増加している．この結果は，より最近に起きた事柄の方が思い出しやすいことを表す．特にリツイートの分布は 1700 年頃から指数関数的に増加している．この曲線は，Au Yeung らによる新聞記事を用いた分析結果 [2] と同様に，本質的には忘却曲線と同じである．

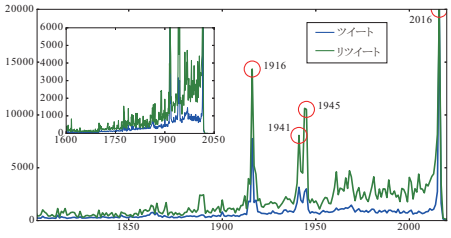

図 2　ツイート中の時間参照の分布.

表 2　ツイートで言及される固有表現とハッシュタグの上位 5 件.

ランク	固有表現	ハッシュタグ
1	United States	history
2	World War II	otd
3	United Kingdom	onthisday
4	Germany	wwii
5	France	throwbackthursday

表 3　リツイートで言及される固有表現とハッシュタグの上位 5 件.

ランク	固有表現	ハッシュタグ
1	United States	otd
2	United Kingdom	onthisday
3	France	throwbackthursday
4	World War II	history
5	Germany	ww1

表 2 と表 3 にツイートとリツイートそれぞれにおいて言及された固有表現とハッシュタグの上位 5 件を示す．なお，本研究では AIDA [14] を用いて固有表現を抽出した．固有表現の結果は，いずれのデータセットにおいても同じ 4 つの国と第二次世界大戦が含まれている．ハッシュタグも 5 件中 4 件は同じであり，残りの 1 件も 20 世紀の世界大戦に関わるものであった．これらの結果から，過去を振り返るとき，どのような国や地域が関係しているのかを明記していることが多いと考えられる．

次に，ツイートにおいて，1916, 1941, 1945, 2016 年のそれぞれの時間参照と一緒に言及される上位 5 件の固有表現とハッシュタグを，図 3 と図 4 に示す．図 3 では，国や地域が 7 件，人名（Adolf Hitler）が 1 件，イベント（World War II）が 1 件，新聞社（The Sun News-Pictorial）が 1 件，含まれている．なお，1916 年は第一次世界大戦の期間中の年であるが，第二次世界大戦について多く言及されている．一方，ハッシュタグに着目すると（図 4），1916 年と第一次世界大戦に関するハッシュタグが共に言及されている回数が多い．これらの結果から，1916 年では 2 つの世界大戦について比較していることが考えられる．

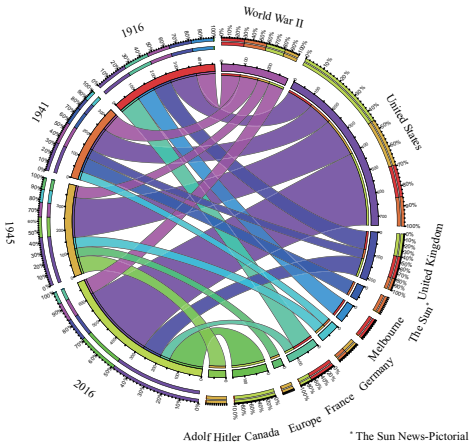

図 3　1916, 1941, 1945, 2016 年の各年を参照するツイートで言及された上位 5 件ずつの固有表現．各ストライプが 1 つの固有表現を表し，ストライプの太さが出現回数を表す.

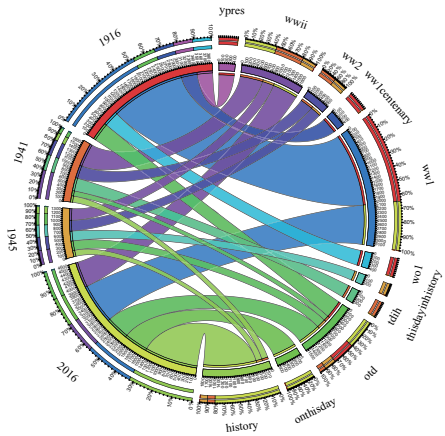

図 4　1916, 1941, 1945, 2016 年の各年を参照するツイートで言及された上位 5 件ずつのハッシュタグ．各ストライプが 1 つのハッシュタグを表し，ストライプの太さが出現回数を表す.

4.2　URL 解析

写真や動画，物，歴史的な文章などがきっかけとなって過去を思い出す事がしばしばある．実際，現代ではそのようなデータを提供する Web サービスが多く存在する．また，表 1 で示したように，半数以上のツイートには Web サービスへのリンクが含まれている．そこで，過去を参照するツイート中に含まれるリンクについて分析する．

表 4 に，本稿で使用するデータセットに含まれている Web サービスの上位 10 件を示す．前述したように，画像（Insta-

表 4　ツイート中で言及される URL の上位 10 件.

URL	出現数
www.instagram.com	42,762
twitter.com	37,683
www.youtube.com	13,026
www.facebook.com	11,763
www.history.com	11,335
www.amazon.com	6,417
paper.li	2,781
www.ebay.com	2,444
linkis.com	1,936
vine.co	1,499

gram），動画（YouTube, vine）は高い頻度で言及されていることがわかる．同様に，画像や動画を含むあらゆる歴史的なコンテンツを提供するサイト（www.history.com）も頻繁に言及されている．ここで，どのような種類の画像がリンクされているのかを分析したところ，個人的な歴史と一般的な歴史の 2 種類に分類できることが明らかになった．前者は，Twitter や Facebook で自分自身の幼少期の写真をポストする際に使用されるハッシュタグ`#throwbackthursday` と共に言及される傾向がある．一方，後者はあらゆる種類の過去の写真を意味する．

また，オンラインニュースを提供するサイト（paper.li）もよく言及されている．これは, [2] が明らかにしたように，現在進行中のイベントとの比較や考察の題材として過去のイベントが思い出されることを示唆する．

最後に，ユーザはショッピングサイト（amazon）も歴史的なハッシュタグと共に言及する傾向があることが明らかになった (注7)．これらのツイートを調べたところ，歴史的なイベントや人物などをテーマとした映画について言及していることがわかった．

4.3　固有表現解析

本節で，どのような人やイベントがよく思い出されるのかを解析する．

4.3.1　頻度に基づいた解析

まず，各固有表現の出現回数の上位 30 件を図 5 に示す．表 2 の結果と同様に，国や都市といった場所が多く含まれている．具体的には，場所に関する固有表現が 21 件，人名が 3 件，イベントが 2 件，グループが 3 件，自然言語が 1 件ある．

次に，固有表現の型ごとの解析結果を示す．本稿では固有表現の型を DBpedia [3] から取得し，その結果を**人**，民族や組織を表す**グループ**，国や地域などの**場所**，**イベント**，**その他**の 5 種類に分類する．本稿で使用するデータセットに含まれる固有表現の型の割合を図 6 に示す．人とグループがデータセットの約 73%を構成していることから，過去を参照するとき，あらゆる人やグループについて，国や地域と一緒に言及していると考えられる．

次に，人，グループ，イベントのカテゴリごとの出現数が多い上位 10 件を表 5 に示す．人とイベントに関しては，歴史的に広く知られている固有表現（Adolf Hitler, Abraham Lincoln, World War II など）が頻繁に言及されていることがわかる．また，本稿のデータセットはアメリカ大統領選が行われた 2016 年のツイートを多く含むので，4 名のアメリカ大統領（Abraham Lincoln, Donald Trump, Bill Clinton, Barack Obama）が多く言及されている．その他の人は歌手（Sharon Corr，Marvin Gaye），第二次世界大戦に関する本を執筆する著者（Jerrard Tickell, David Irving）とプロダーツプレイヤー（Alan Evans）であった．グループは，民族（Jews），国家的な組織，オンラインサービスを提供する会社（Facebook, BBC）が多く言及されている．

(注7)：`https://twitter.com/JA\_Redmond/status/727121139370188 80\\1/photo/1`

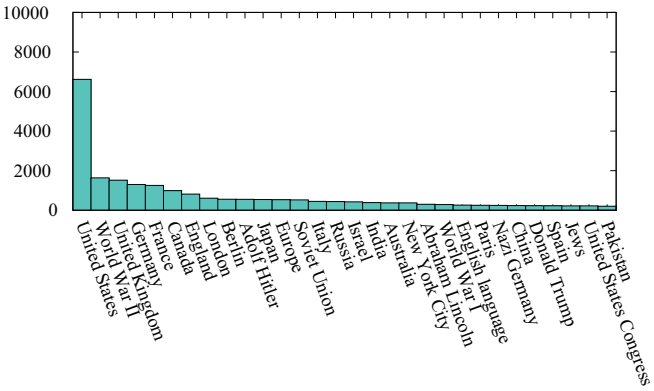

図 5　ツイート中の出現数上位 30 件の固有表現

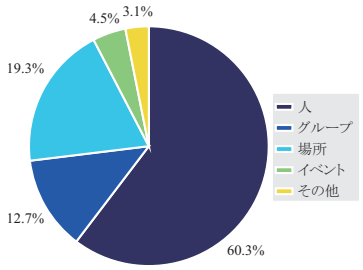

図 6　固有表現の型の割合.

表 5　ツイートに含まれる人，グループ，イベントの上位 10 件.

人	グループ	イベント
Adolf Hitler	Jews	World War II
Abraham Lincoln	United States Army	Vietnam War
Sharon Corr	United States Congress	Omaha Beach
Alan Evans	United States Navy	Battle of Verdun
Donald Trump	Facebook	Korean War
Jerrard Tickell	BBC	Battles of Saratoga
Bill Clinton	Royal Navy	American Revolutionary War
Marvin Gaye	NASA	Cinco de Mayo
Barack Obama	Federal Bureau of Investigation	Battle of Gettysburg
David Irving	United States Marine Corps	Siege of Yorktown

5.　ハッシュタグに基づく解析

5.1　ハッシュタグの使用頻度

本節で，各ハッシュタグの使用頻度に基づいて分析する．図 7 にデータセット全体で使用頻度が高い上位 30 件を示す．次に，図 8 に，各ハッシュタグを使用しているアカウント数が多い上位 30 件を示す．これらのデータから，`#history`,

#throwbackthursday，#wwii, #onthisday が広く使用されていることがわかる．また，自分自身の過去の写真をポストする際に使用される#throwbackthursday と#tbt は，図 7 よりも図 8 の方が上位に位置している．一方，世界大戦に関係するハッシュタグ（#wwii, #wwi, #ww2, #ww1）を使用するユーザ数は減少している．これらの結果から，多くの人にとって個人的な経験は思い出しやすいが，後者のような一般的な過去はその影響が比較的小さいと考えられる．

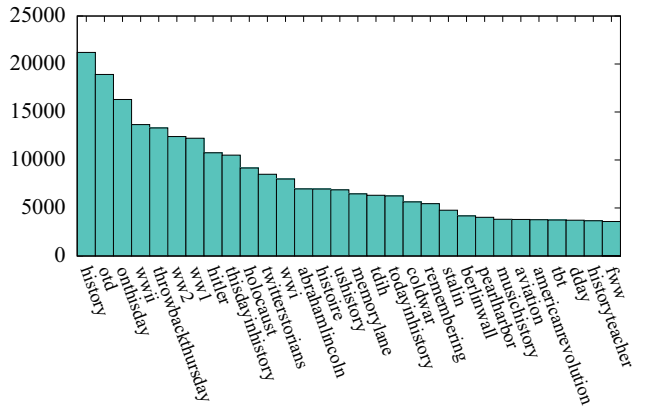

図 7　ツイートに含まれるハッシュタグの上位 30 件．x 軸はツイート数を意味する．

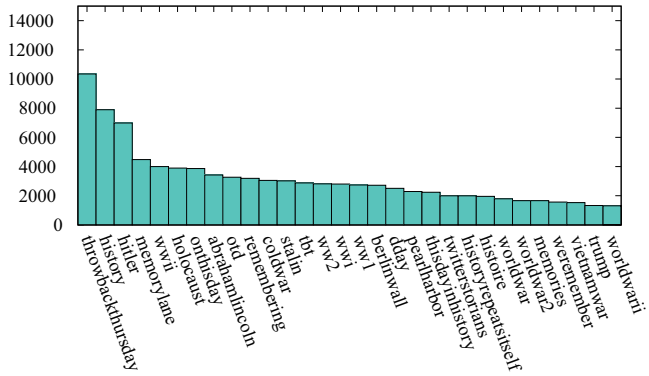

図 8　ユーザに使われているハッシュタグの上位 30 件．x 軸はそのハッシュタグを含むツイートを生成したことがあるユーザ数を意味する．

5.2　カテゴリに基づいた解析

本節で，歴史に関係するハッシュタグのカテゴリを歴史家による分析結果に基づいて定義し，カテゴリごとの分析を行う．本稿では，ブートストラップ法で収集した 130 個の歴史に関するハッシュタグを以下の 6 つのカテゴリに分類する．

1. **一般的な歴史**　：#history や #historicalcontext といった，一般的な歴史を参照するツイートで使用されているハッシュタグを含む．
2. **一国史・地域史**　：#canadianhistory や #ushistory のような，ある特定の国や地域に関する歴史について言及しているツイートで使用されているハッシュタグを含む．
3. **テーマ史**　：美術史やスポーツ史といった，歴史の特定の側面に関連するハッシュタグを含む．
4. **コメモレーション**　：このカテゴリは，#onthisday, #otd, #todayweremember, #4yearsago のような，ある特定の日や期間に生じた事柄への記憶を表すハッシュタグを含む．
5. **歴史的なイベント**　：過去の特定のイベントに関係するハッシュタグを含む．例えば，#wwI，#SevenYearsWar などが含まれる．
6. **歴史的な固有表現**　：人や組織や物などの固有表現（ただし，国や地域，土地は除く）に関するハッシュタグを含む．

これらのカテゴリに属するハッシュタグの一部を表 6 に示す．また，各カテゴリのハッシュタグを含むツイートの割合を図 9 に示す．1 番多く出現するテーマ史が約 30%含まれており，次いでコメモレーション，一般的な歴史が共に約 21%程度出現する．

表 6　手動によるグループ分けしたカテゴリと，各カテゴリに属するハッシュタグの例

カテゴリ	ハッシュタグ
一般的な歴史	history, historyfacts, oldpicture, historyteacher, historicalcontext
一国史・地域史	canadianhistory, ushistory, nazigermany, ottoman, ancientchina
テーマ史	arthistory, sporthistory, womenshistory, musichistory, historyscience
コメモレーション	onthisday, otd, otdh, ThisDayIn, thisdayinhistory
歴史的なイベント	gulfwar, ColdWar, ww2, ww1, americanrevolution,
歴史的な固有表現	stalin, hitler, abrahamlincoln, rudolfhess, napoleon

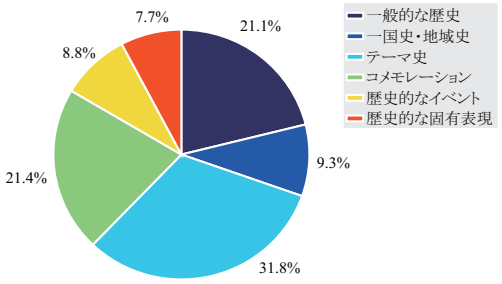

図 9　カテゴリの割合

5.2.1　カテゴリ間の類似度

次に，カテゴリ間の類似度を分析する．ここで，本稿では，1 つのツイートで一緒に使用されるハッシュタグは類似し，その使用頻度が高いほどそれらのハッシュタグの類似度も高いと仮定する．カテゴリの類似度も同様に，カテゴリに属するハッシュタグが一緒に使用される回数で評価する．カテゴリ間の類似度を次式のジャッカード係数によって定義する．

$$CatSim(A,B) = \frac{|\,T_A \cap T_B\,|}{|\,T_A \cup T_B\,|} \tag{1}$$

ここで，$|\cdot|$ は集合の要素数を表し，T_A, T_B はそれぞれカテゴリ A と B に含まれるハッシュタグを含むツイートの集合を表す．

図 10 にカテゴリ間の類似度を示す．一般的な歴史は全体的

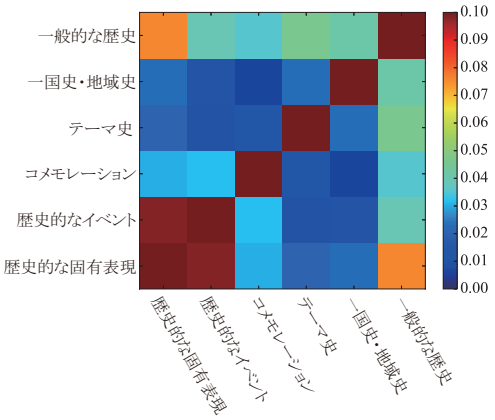

図 10　カテゴリの共起性

に他のカテゴリと共に用いられる傾向がある．一方，**歴史的な固有表現**と**歴史的なイベント**の類似度が最も高い．これは，過去のイベントとどのような人が関係していたのかを述べていることが多いことを示す．同様に，**歴史的な固有表現**と**テーマ史**の共起性が高いことから，特定の歴史の側面について述べるときも関連する人についても言及していると考えられる．

5.2.2　カテゴリの時間参照分布

次に，時間参照に基づいてカテゴリを分析する．まず，各カテゴリのハッシュタグと時間参照が一緒に言及されているツイートの割合と，時間参照の標準偏差を表 7 に示す．**コメモレーション**は`#onthisday` のような過去の特定の日や期間を参照するハッシュタグを含むので，半数以上のツイートが時間参照を含む．一方，過去のイベントや人物といった要素について言及する傾向がある**一般的な歴史**，**一国史・地域史**，**テーマ史**は約 20–30%のツイートが時間参照を含む．標準偏差に関しては，あらゆる年代の事柄を参照する**コメモレーション**と**歴史的なイベント**は 1000 を超え，第二次世界大戦のような特定の時期に活躍した人物を含む**歴史的な固有表現**は 160 程度である．

次に，同じカテゴリに属するハッシュタグが同一の時期を参照しているのかどうかをコサイン類似度に従って分析する．まず，各ハッシュタグを時間参照を基にベクトル化する．図 2 が示すように，多くのツイートは 1750 年から 2040 年を参照しているので，この期間の出現回数を値とするベクトルを作成する．同一カテゴリ内のハッシュタグのペアワイズ類似度の平均を表 7 に示す．**一般的な歴史**と**歴史的な固有表現**のコサイン類似度は比較的値が高い．すなわち，これらのカテゴリには同じ期間の過去を参照するハッシュタグを多く含むことがわかる．一方，**テーマ史**と**歴史的なイベント**の値は低いので，テーマやイベントのハッシュタグは互いに異なる期間の過去を参照することがわかる．特に**歴史的なイベント**は，標準偏差の値は高いがコサイン類似度の値は低いことから，ハッシュタグごとに様々な年代の過去を参照していると考えられる．

最後に，各カテゴリの時間参照の分布を図 11 に示す．**歴史的なイベント**は 1900 年以降は全体的に各年への参照を多く含むが，特に 2 つの世界大戦に関係する年と 2016 年の 3 箇所でピークがある．**コメモレーション**もこれらの世界大戦の年でのピークが見受けられるが，他のカテゴリよりも各年の参照数が多く，ナポレオンやアメリカ独立戦争に関係する年（18 世紀後半）や南北戦争（1861–1865 年）の期間を多く参照している．**一国史・地域史**と**歴史的な固有表現**は 2016 年でピークが存在する．この結果は，これらのカテゴリでは，現代のイベントや固有表現が過去を思い出すきっかけとなっていることを示す．しかし，**テーマ史**は 1916 年の参照数が比較的多いことから，現代との関連性は見受けられない．

表 7　カテゴリに対するコサイン類似度の平均値

カテゴリ	時間参照を含む割合	標準偏差	コサイン類似度
一般的な歴史	20.1%	581.862	0.824
一国史・地域史	20.5%	258.862	0.661
テーマ史	32.9%	335.115	0.559
コメモレーション	55.6%	1000.777	0.607
歴史的なイベント	18.5%	1161.994	0.464
歴史的な固有表現	9.2%	159.832	0.830

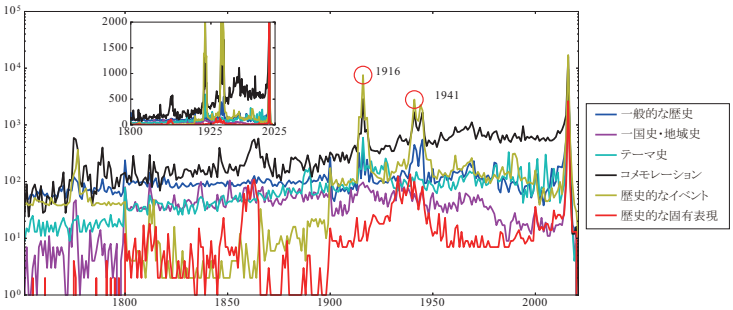

図 11　カテゴリの時間参照分布

5.2.3　エントロピー

次に，エントロピーに関する分析を行う．まず，固有表現のエントロピーとツイート数の相関関係を図 12 に示す．多くのハッシュタグにおいて，ツイート数が増加すると言及される固有表現の種類も増加している．特に，`#otd`，`#onthisday`，`#tdih`，`#history` といったあらゆる過去を参照しうるハッシュタグは，多くのツイートが存在し，あらゆる種類の固有表現と共に使用されている．なお，**コメモレーション**に属する多くのハッシュタグは，多くの種類の固有表現と一緒に言及されている傾向がある．

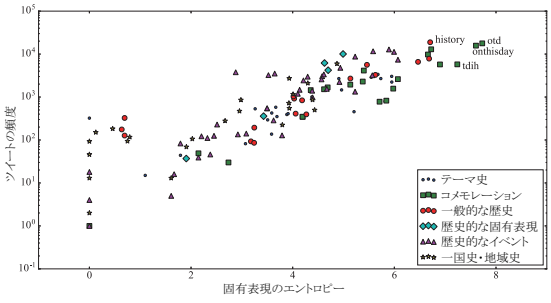

図 12　歴史に関するハッシュタグの固有表現に基づくエントロピー（x 軸）とそれらを含むツイート数（y 軸）の相関関係．

次に，時間参照の種類と固有表現の種類の相関関係の分析結果を図 13 に示す．先ほどと同様に，全体的に正の相関を確認で

き，特に，#otd，#onthisday，#tdih，#history，#thisday といったハッシュタグは様々な年を参照していることが明らかである．

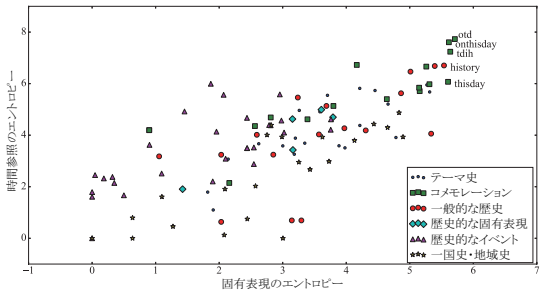

図 13　歴史に関するハッシュタグの固有表現に基づくエントロピー（x 軸）と年に基づくエントロピー（y 軸）の相関関係．

最後に，ユーザの種類と固有表現の種類の相関関係の分析結果を図 14 に示す．これまでの図と同様に，正の相関が確認できる．また，#otd，#onthisday，#tdih は多くのユーザに使用されていることが観察できるが，固有表現のエントロピーは減少している．一方，#throwbackthursday と #hitler は多くのユーザに，あらゆる固有表現と共に言及されていることがわかる．

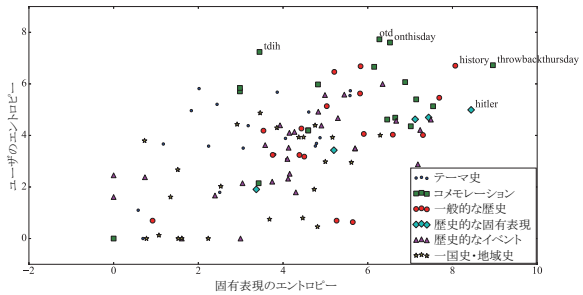

図 14　歴史に関するハッシュタグの固有表現に基づくエントロピー（x 軸）とユーザに基づくエントロピー（y 軸）の相関関係．

5.3　固有表現に基づくクラスタリング

ハッシュタグのグループを自動的に作成するために，歴史に関係するハッシュタグの階層型クラスタリングを適用する．本稿では，各ハッシュタグと同じツイートで出現する固有表現の回数を使用して特徴ベクトルを作成する．特徴ベクトル間の類似度の評価関数として次の式を用いる．

$$d(v_1, v_2) = 1 - \frac{(v_1 - \bar{v_1}) \cdot (v_2 - \bar{v_2})}{|| (v_1 - \bar{v_1}) ||_2 || (v_2 - \bar{v_2}) ||_2} \quad (2)$$

ここで，$\bar{v}$ はベクトル v の要素の平均，$x \cdot y$ はベクトル x, y の内積を表す．クラスタ間の類似度は次式で定義される群平均法を用いる．

$$d(U, V) = \sum_{i,j} \frac{d(U_i, V_j)}{| U | * | V |} \quad (3)$$

出現数上位 20 個のハッシュタグに対して階層型クラスタリングを適用した結果を図 15 に示す．

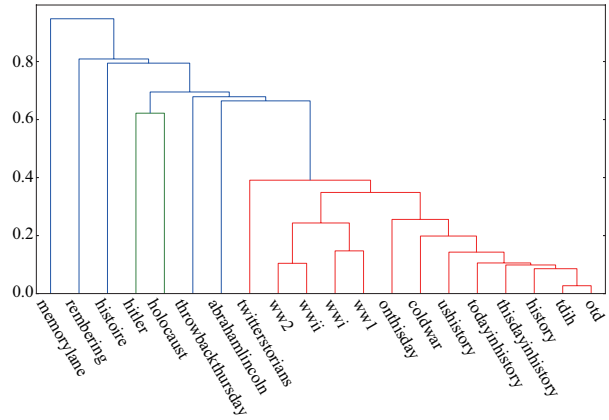

図 15　固有表現に基づいたハッシュタグクラスタリング

意味的に類似するハッシュタグ（例えば，ww2 - wwii, wwi - ww1, todayinhistory - thisdayinhistory - history - tdih - otd)）がクラスタを形成している．また，同じイベントに関連する固有表現のクラスタ（hitler - holocaust）も形成されている．このように，ハッシュタグと固有表現の組み合わせは，ハッシュタグの意味を解析するために有効であると考えられる．しかしながら，(hitler - holocaust）は第二次世界大戦のクラスタ（ww2 - wwii）と意味的な類似度が高いにも関わらず，同じクラスタを形成すると他のハッシュタグを含む．したがって，時間参照のような他の特徴を併用することによってクラスタリングの精度を向上させることが期待できる．

6.　議　　論

6.1　何がきっかけでマイクロブログ上で歴史を参照するのか

本稿の分析結果から過去を参照するきっかけについて議論する．まず，使用頻度が高いハッシュタグを示す結果（図 7 および図 8）において，世界大戦に関するハッシュタグや，#thisdayinhistory, #throwbackthurthday, #history, #otd, #onthisday といった特別な日や曜日に使用するハッシュタグが上位に含まれていることが観察できる．これは，日付が過去を振り返る際の強いきっかけになることが考えられる．

次に，本稿のデータセットに含まれるツイートの半分以上が URL を含み，それらのリンク先を調べた結果（表 4）によると，歴史的な画像や動画，文章が頻繁に参照されていた．リンク先を分析したところ，過去のイベントを描く写真だけでなく，ユーザ自身の過去の写真や，歴史的な内容を題材とした映画を販売する Web サイトについて言及しているものが確認できた．すなわち，過去を振り返るきっかけとしてメディアも強い影響を与えることが考えられる．また，ハッシュタグのカテゴリごとの時間参照の分布の結果（図 11）によると，現代との比較のために過去を参照する事も観察できる．新聞記事を用いた研究 [2] や新しい歴史の学習環境を実現する研究 [23] でも議論されているように，現代社会が直面している問題について考察するために過去のイベントを参照することがマイクロブログでも確認できる．なお，図 3 の結果は，単に現代と過去のイベント

を比較するだけでなく，過去のイベントをきっかけに他の過去のイベントについても議論している様子を示唆する．

6.2 集合的記憶は Twitter 上でどのように表現されているのか

エントロピーに関する分析結果を示す 3 つの結果（図 12, 図 13, 図 14）から，特定の期間や話題について一部の人が思い出すのではなく，様々な人があらゆる年代について言及していることがわかる．また，図 15 が示すように，世界大戦に関するハッシュタグ（`ww2`, `wwi` など）やコメモレーションに関するハッシュタグ（`todayinhistory`, `otd` など）といった意味的に類似するハッシュタグが同じクラスタを形成できていることから，単に過去の事柄を説明するだけでなく，比較としても過去の記憶を表現していることがわかる．

7. まとめと今後の課題

本稿では，過去を参照するツイートを時間，URL，固有表現，ハッシュタグに着目して分析した．前述したように，本稿の分析結果は歴史学習に関する研究の基礎となり，特に歴史に特化した推薦システムの実現は，ユーザに歴史への興味を促進することが期待できる．多くの先進国において，近年，歴史への関心が薄れてきていることが報告されている．例えば，アメリカの全国テストの結果では，アメリカ史に関して十分な成績を収めた学生は 4 分の 1 にも満たず，また，国の歴史が基礎的な要素であることを認識できていないことも示されている [21]．同様な現象がオーストラリアでも確認されている [11]．

`#throwbackthursday` のような個人の歴史を振り返るツイートと，`#onthisday`，`#ww2`，`#ww1` といった歴史的な事実や記憶を表すハッシュタグを含むツイートの違いを，人気度に着目して分析し，歴史への興味を促すための知見を明らかにすることは，1 つの重要な今後の課題である．また，過去を思い出すきっかけを調査した研究はこれまでにも行われているが [9], [17], [20]，多くは個々人に着目した調査である．大勢の人が過去を思い出すきっかけを深く分析するために，どのようなツイートが再ポストされるのかを詳細に分析することも今後の重要な課題である．

文　献

[1] R. P. Abelson and A. Levi. Decision making and decision theory, handbook of social psychology. pages 231–309, 1985.

[2] C.-m. Au Yeung and A. Jatowt. Studying how the past is remembered: Towards computational history through large scale text mining. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1231–1240, New York, NY, USA, 2011. ACM.

[3] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[4] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):15, 2015.

[5] F. Clavert, B. Majerus, and N. Beaupre. #ww1. twitter, the centenary of the first world war and the historian.

[6] J. Cook, A. Das Sarma, A. Fabrikant, and A. Tomkins. Your two weeks of fame and your grandmother's. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 919–928, New York, NY, USA, 2012. ACM.

[7] H. Ebbinghaus. *Memory: A Contribution to Experimental Psychology.* 1913.

[8] M. Ferron and P. Massa. Collective memory building in wikipedia: The case of north african uprisings. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 114–123, New York, NY, USA, 2011. ACM.

[9] R. Garcia-Gavilanes, A. Mollgaard, M. Tsvetkova, and T. Yasseri. Memory remains: Understanding collective memory in the digital age.

[10] T. Gilovich. Seeing the past in the present: The effect of associations to familiar events on judgments and decisions. *Journal of Personality and Social Psychology*, 40(5):797, 1981.

[11] J. Gregory. Don't know much about history: Australian history in schools today. *The Critic*, 6, 2007.

[12] M. Halbwachs. *La Memoire Collective.* Les Presses universitaires de France, (in French), 1950.

[13] C. Hoerl and T. McCormack. *Time and Memory: Issues in Philosophy and Psychology.* 2001.

[14] J. Hoffart, S. Seufert, D. B. Nguyen, M. Theobald, and G. Weikum. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 545–554, New York, NY, USA, 2012. ACM.

[15] R. Jacoby. *Social Amnesia: A Critique of Contemporary Psychology.* 1997.

[16] A. Jatowt, E. Antoine, Y. Kawai, and T. Akiyama. Mapping temporal horizons: Analysis of collective future and past related attention in twitter. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 484–494, New York, NY, USA, 2015. ACM.

[17] N. Kanhabua, T. N. Nguyen, and C. Niederée. What triggers human remembering of events?: A large-scale analysis of catalysts for collective memory in wikipedia. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 341–350, Piscataway, NJ, USA, 2014. IEEE Press.

[18] N. Kanhabua, C. Niederee, and W. Siberski. Towards concise preservation by managed forgetting: Research issues and case study. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres)*, 2013.

[19] E. Kuzey, J. Strötgen, V. Setty, and G. Weikum. Temponym tagging: Temporal scopes for textual phrases. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 841–842, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.

[20] M. Proust and J.-Y. Tadie. *A la recherche du temps perdu.* Gallimard.

[21] J. Soboroff. If students fail history, does it matter?, 2011.

[22] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. 10:178–185.

[23] 池尻良平，澄川靖信. 真正な社会参画を促す世界史の授業開発 – その日のニュースと関連した歴史を検索できるシステムを用いて –. **全国社会科教育学会，社会科研究**, 84:37–48, July 2016.