オンライン小説におけるキーワードの時系列傾向分析

浦川, 隆寛
九州大学工学部電気情報工学科 | 九州大学情報基盤研究開発センター

伊東, 栄典
九州大学工学部電気情報工学科 | 九州大学情報基盤研究開発センター

出版情報：情報処理学会研究報告, pp.B-3-2-, 2013-06-07. 情報処理学会
バージョン：

オンライン小説におけるキーワードの時系列傾向分析

浦川隆寛[†1]　伊東栄典[†2]

Web 上の利用者投稿型メディアであるオンライン小説では，現在流行分野の小説を書かれることが多い．本研究では，作者が小説へ付与するキーワードの時系列分析を行い，ジャンルの流行り廃りを解析する．キーワードにはゆらぎがあるため，機械的に算出した関連語の傾向分析も行うことにした．本論文では，作成した時系列分析ツールの構成を述べる．また，傾向分析に用いたデータを説明し，最後にいくつかの興味深い分析結果を示す．

A study of keywords frequency trend analysis of online novels

TAKAHIRO URAKAWA[†1]　EISUKE ITO[†2]

Online novel sharing service, which is a user-generated media on Web, becomes popular and a large number of novels are being uploaded. Because authors like to write current popular genre, then current popular genre words may frequently appear. In this research, the authors apply the time series analysis to the keywords and genre words given to a novel by the author, and analyze the changes trend. Since keywords are not controlled, there was fluctuation in keyword. Then, the trend analysis system not only show the trend of query words but also show the trend of related terms automatically calculated using similarity. This paper describes the trend analysis system, the used data, and some interesting trend analysis results.

1.　はじめに

膨大なコンテンツを持つ投稿型コンテンツサービスで，利用者が求めるコンテンツを探すためには検索や推薦システムが重要である．利用者投稿型コンテンツサービスでは，YouTube, ニコニコ動画, YouKu などの動画が人気である．近年ではオンライン小説も徐々に人気になっている．日本の「小説家になろう」[1]や中国の「起点」[2]ではコンテンツ数や利用者数が増加している．これらのオンライン小説サイトに投稿された小説が，普及しつつある．

我々は，動画のランキング[4]及びカテゴリ分け手法[5]の研究だけでなく，学術論文の推薦[6]についても研究してきた．また，文書群の多面分析手法についても研究してきた[7]．これらの成果を用いて現在，日本のオンライン小説サイトである「小説家になろう」に投稿された小説を対象に，検索手法を検討している．このサイトに投稿される小説数は近年急増しており，2012 年 9 月現在，13 万件以上の小説が投稿されている．ほとんどの作者はアマチュアであるためか，多くの小説の品質は低い．しかし稀に高品質のコンテンツも投稿されている．

「小説家になろう」では，投稿者が自身の小説に対してキーワードを付与することができる．１つの小説に対して最大 15 個まで付けられるそれによって，カテゴリや作風，特徴的な概念などの情報を事前に読者に伝えることが可能になっている．また, このサイトの検索エンジンはキーワード検索にも対応している．その結果，作者がより多くの読者の目につくように，その時期流行の単語を積極的にキーワードに組み込むことが考えられる．

本研究ではキーワードの出現頻度から小説の流行について調べることを目的とする．そのため, 期間毎のキーワード出現頻度をグラフで視覚化するツールを作成する．検索語が明確でない場合のために，検索語の類似語も分析対象とし, これにより検索の幅を広げ, 利用者が詳しくないジャンルに対しても流行を把握できるようにした．

2.　基礎分析

「小説家になろう」はヒナプロジェクト社が提供するオンライン小説の投稿・閲覧サービス[1]である．小説数等の統計とメタデータについて述べる．

2.1　小説・読者・著者の数

表 1 に 2012 年 5,7,9 月の小説数, 読者数, 著者数を示す．著者は，１つ以上の小説を投稿した利用者である．小説をお気に入り登録している読者数をマイページ数としている．マイページについては，後で述べる．なお，2012 年 7 月から 2012 年 9 月にかけて, 小説数および著者数が減少しているのは，二次創作の小説が削除されたためである．

表 1：小説・読者・著者数の推移

	2012年1月	2012年5月	2012年7月	2012年9月
小説	134,763	159,090	168,396	148,278
読者	195,716	240,730	258,478	272,512
著者	46,938	53,396	56,214	44,585

図 1 に 2004 年 4 月から 2012 年 8 月までの期間における，各月の新規投稿小説数を示す．表 1 に示す現在の総小説数は約 13 万件で膨大ではない．しかし，この数年間の投稿小説は指数関数的に増大しているため，蓄積小説数は近い将来，膨大な数なると予想できる．

†1 九州大学工学部
Faculty of Engineering, Kyushu University
†2 九州大学情報基盤研究開発センター
Research Institute for IT, Kyushu University

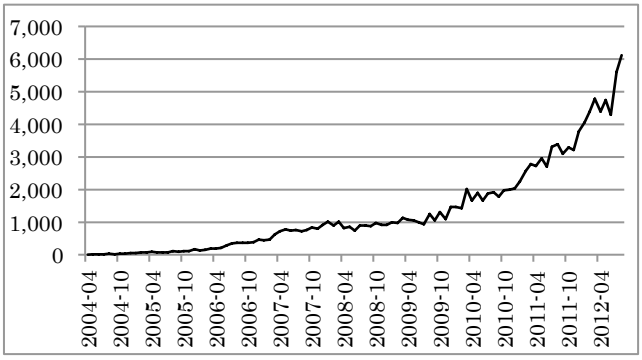

図 1　各月の新規投稿小説数

2.2　メタデータ

「小説家になろう」では，小説は単数あるいは複数のセクションから構成される.セクションが 1 つしかない場合，短編小説となり，それ以外は連載小説となる．連載小説の場合，著者は「完結済み」を設定できる．小説投稿の際，タイトル，著者名，ジャンル，あらすじ，キーワード，をメタデータとして一緒に投稿する．ジャンルはサイト運営側で指定された 15 個の単語から選ぶ.あらすじとキーワードは，最大文字数以下で自由に作者が付与できる．

投稿された小説は，利用登録なしに誰でも読むことができる．サイトに利用者登録をすると，小説に対してスコア付けと，コメント送付といったフィードバック機能を利用することが出来る．また，登録利用者は，小説のお気に入り登録（ブックマーク），お気に入り作者の登録ができる．登録した情報は，マイページと名付けられたページで閲覧できる．マイページを使うと，お気に入り登録した連載小説の更新情報や，お気に入り登録した作者の新作案内通知などの便利な機能を使うことが可能になる．

メタデータの例を図 2 に示す.この例は,小説 ID「n3556o」の小説である．例に示したように，メタデータには，タイトル，著者，あらすじ，ジャンル，キーワード欄がある．また，初回投稿日，最終更新日もある．更に，読者が与える評価値と，お気に入り登録数（ブックマーク数）の情報もある．

小説情報

作者：天酒之瓢

小説タイトル	Knight's & Magic
あらすじ	メカヲタ社会人が異世界に転生。 その世界に存在する巨大な魔導兵器の乗り手となるべく、彼は情熱と怨念と執念で全力疾走を開始する……。
キーワード	R15 残酷な描写あり 関西人 ロボット 魔法 ファンタジー 異世界 転生 学園
掲載日	2010年 10月16日 23時27分
最終投稿日	2013年 02月06日 00時09分

Nコード	N3556O	開示設定	開示されています	お気に入り登録	19,978件
ジャンル	ファンタジー	感想	1,388件	レビュー	4件
種別	連載：全58部	感想受付	受け付ける(ユーザのみ)	レビュー受付	受け付ける
年齢制限	なし	文字数	556,667文字	ポイント評価受付	受け付ける
総合評価	69,140pt	文章評価	14,417pt	ストーリー評価	14,767pt

図 2　小説メタデータの例（小説 ID：n3556o）

Knight's & Magic

作者：天酒之瓢

メカヲタ社会人が異世界に転生。 その世界に存在する巨大な魔導兵器の乗り手となるべく、彼は情熱と怨念と執念で全力疾走を開始する……。

第１章　転生、そして学園生活編

#1　別れと出会い	2010年 10月 16日 （改）
#2　魔法を使おう	2010年 10月 19日 （改）
#3　旅には道連れ	2010年 10月 20日 （改）
#4　発想の転換	2010年 10月 23日 （改）
#5　図書館にて	2010年 10月 27日 （改）
#6　入学式にて	2010年 10月 29日 （改）
#7　その武器の名は	2010年 10月 29日 （改）
#8　授業をうけよう	2010年 11月 03日 （改）
#9　決闘の時間	2010年 11月 06日 （改）
#10　決闘の決着	2010年 11月 07日 （改）

第２章　魔獣襲来編

#11　陸皇襲来	2010年 12月 14日 （改）
#12　見学しよう	2010年 12月 22日 （改）

図 3　小説表紙ページの例（小説 ID: n3556o）

2.3　小説数・単語数・共起単語対

時系列傾向分析で用いたデータについて述べる．分析には 2012 年 9 月時点の小説メタデータを用いた．収集したメタデータの小説数や単語数を表 2 に示す.

表 2　2012 年 9 月の小説数および単語数

記号	要素数	説明
D	148,278	全小説メタデータ集合
W	90,052	全単語集合
P	1,022,788	共起する単語対の集合

D は全小説メタデータの集合で，この要素数は投稿された全小説数に等しい．W はメタデータ群のキーワード欄に出現した一意な単語の集合である．P はキーワード欄で共起した単語対の集合である．

3.　出現頻度による傾向分析

小説群の流行分析を２つの方法で行う．１つ目は，小説で使われる単語の出現頻度を期間毎にプロットし，それにより時系列での単語出現頻度の増減傾向を調べる方法である．二つ目は，ある入力される検索語の傾向分析を行うだけでなく，検索語と類似する単語の傾向も合わせて調べる方法である．

3.1　時系列における単語の出現頻度

時系列に対するキーワードの検証方法について説明する．投稿日時を期間毎に区切り．小説の初回投稿日が区間 t にある小説群を対象に，キーワード欄の単語 w の出現頻度 tf(w, t)として，tf(w, t)を時間 t に対してプロットする．この tf(w, t)の時系列変化から流行の傾向変化を見る．ある期間に tf(w, t)が増加・減少すれば，その期間が単語 w を含む小説が流行していた期間，あるいは流行が廃れてきた期間であると判断できる．

3.1.1 対象レコード

小説のメタデータが持つレコードのうち，単語の出現回数の数え上げ対象になりうるものには，キーワード欄の他

に，題名，あらすじ，ジャンルがある．本研究では，キーワード欄だけを数え上げの対象にした．「小説家になろう」では，作者が自分の作品にキーワードを与えるため，キーワード欄の単語は，作者が自分の作品の特徴を表すと信じている単語であろう．そのため，流行を調べる傾向分析には適している．また，キーワード欄の単語数は少ないため処理が単純になり，最初の傾向分析としても適している．

題名・あらすじ・ジャンルおよびキーワード欄を対象に，単語の出現回数を調べる方法は将来検討する．題名やあらすじの中でも，頻繁に出現する単語は，その各小説の特徴を表す単語であると想定できるため，頻出単語を重く扱う手法も適切であろう．本論文のキーワード欄だけを見る手法を比較したい．

3.1.2 対象期間

「小説家になろう」では，連載小説と短編小説の２つを投稿できる．短編小説とは，節（セクション）が１つだけの小説である．投稿されている小説のほとんどは連載小説である．長い期間連載されている小説も多い．例えば図 2 の小説（小説 ID: n3556o）は３年以上連載されており，2013 年 1 月現在でも連載が継続している．実際，完結（連載終了）と指定された小説は少なく，多くの連載小説は途中で放置されている．

ある小説メタデータにおける単語の出現頻度を，どの期間に割り当てるかについて，３つの方法を考えた．小説の初回投稿日のみ,初回投稿日と最終投稿日,「章」の設置日，の３つである.初回投稿日と最終投稿日は自明であるため，「章」の設置日を説明する．図 3 に，図２と同じ小説の表紙ページを示す．図 3 に有るように，作者は小説に「章」を設定できる．長期連載小説の「章」は，紙印刷される本における単行本発行日と考えられるため,「章」の設置日を重要視しても良い.

本論文では，第１の手法である「初回投稿日」だけを扱う．初回日が最も連載開始時の流行を反映していると考えられるため,流行を調べる時系列傾向分析には良い.将来，他の期間を考慮した傾向分析を行い，本論文の手法と比較したい.

3.2 類似語の傾向分析

入力される検索語だけでなく，検索語と類似する単語も傾向を分析する.

3.2.1 類似語も対象とする理由

本論部の時系列傾向分析でも，また一般の情報検索でも，検索語与えることから始まる．一般的な事がらであれば，検索語の指定に問題は少ない．例えば，天気予報が知りたい場合,「天気」を検索語に選び，検索対象を絞るために天気を調べたい地名を検索語とする．電車はバスの時刻表を調べたい場合,「時刻表」を検索語にするであろう.

専門的な分野の場合，検索語の指定が難しくなる．それでも，自分が詳しい分野については検索語を選定できる．例えば，情報分野に詳しい人がスマートフォンの性能を調べたい場合，機種名や OS 名，アプリの名称，通信規格名などを入力するであろう．しかし，情報分野を専門としない人には，適切な検索語の選定は困難である.

オンライン小説でも，自分の興味がある分野であれば，検索語を比較的適切に選定可能であろう．しかし，図 2 に示したように,オンライ小説のキーワード数は膨大であり，また多数の作者が自由に小説を書き，かつ自由にキーワードを与えるため，小説を特定する単語は広大な多様性を持つことになる．新語も登録されるため，ますます多様性が大きくなる.

あるジャンルに関連する検索語が与えられたとき，その分野について詳しい利用者は分野の他の単語を知っていて,それらを用いた検索ができる．一方，その分野に詳しくない利用者は当然他の関連語を知らないため，検索を深めたり広げたりする手段に乏しくなる．あらかじめ単語間の類似度を計算しておけば，利用者入力した検索語に対して関連語を返すことができる．検索者が詳細に調べたい可能性がある単語を導きだせば，検索を進める助けになる.

3.2.2 類似度計算手法

単語間の類似度計算手法については様々な方法がある．本論文ではコサイン類似度と Jaccard 係数の二つを用いた．コサイン類似度は，文書と単語の関係を文書ベクトルモデルで表現した場合における，単語ベクトルを用いることで表現できる．Jaccard 係数は集合間の演算であるため，単語の文書頻度と，共起する２つの単語対の文書頻度で計算できる.

コサイン類似度と Jaccard 係数を計算する場合，高頻出語と，低頻出語が問題になる．例えば単語 x と y の頻度がどちらも 1 で，かつ x と y が一つの文書にだけ出現した場合，x と y の類似度はコサイン類似度でも Jaccard 係数でも最高値の 1 になる．この問題を避けるため，閾値以下の出現頻度の語は足切りすることにした.

次に高頻出語を考える．高頻出語は，多くの単語と同時に出現するため，共起頻度も多くなり，そのため類似度も高くなる．しかし高頻出語は一般的な語が多く，本論文で扱うオンライン小説では「ジャンル」の単語が頻出する．これらの高品質語は詳細な傾向を調査する際に邪魔となるため，影響を小さくしたい．そこで， IDF 値（Inversed document frequency）を掛け算することで，高頻出語の影響を低くすることに検討する．IDF 値の効果については後述する.

4. 時系列傾向分析ツールの構築

作成した時系列傾向分析ツールについて述べる．ツールは前処理部と検索部の２つから構成される．前処理部ではメタデータ収集，単語の出現頻度，共起頻度，類似度計算を行う．検索部では，入力された検索語に対して，類似度

に基づく関連語を導き，最後に検索語および関連語の時系列での出現傾向を示す．

4.1 前処理部

図 4 に前処理部の流れを示す．

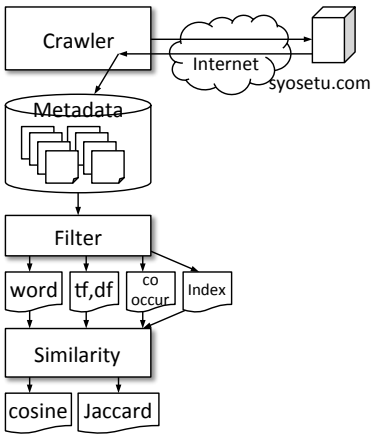

図 4 前処理部

4.1.1 メタデータ収集

「小説家になろう」のデータ提供用の WebAPI を用い，Ruby 言語を用いて小説メタデータ収集プログラムを作成した．「小説家になろう」サイトでは，小説も利用者にも ID 番号が割り当てられている．利用者 ID は連番の数値であるため，順番に数値を増やし，全利用者のメタデータを集めた．利用者 ID から，その利用者の投稿作品メタデータ群を返す WebAPI を用い，全小説のメタデータを入手した．なお，メタデータは図 2 で示した Web ページの HTML 形式ではなく，YAML 形式のテキストデータを入手した．

4.1.2 頻度分析

単語毎にその出現頻度を数え上げる作業には，ハッシュを利用した．ハッシュのキーに単語を割り振り，キーと同じ単語が現れれば，ハッシュ値に 1 加える．同様にして共起頻度もハッシュで数え上げた．

4.1.3 類似度計算

共起した単語対からさらに関連語を導く手段として類似度計算を実行した．前節で述べたように，共起する単語対について，コサイン類似度と Jaccard 係数を算出した．また，コサイン類似度と Jaccard 係数とに IDF 値を掛けたものも算出した．

4.2 検索部

検索部は Web CGI プログラムとして作成した．Web サーバには Apache を用い，CGI プログラムは Ruby 言語で作成した．図 5 に構成を示す．

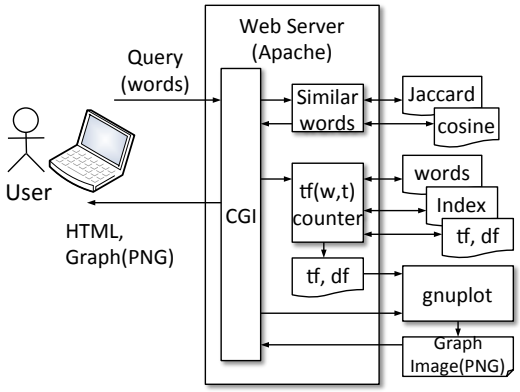

図 5 検索部

利用者は Web ブラウザから複数の検索語を入力する．検索語をプログラムは受け取り，検索語ごとに類似語と類似度を抽出する．上位の類似語を固定数（実験では 5 個とした）抽出する．入力された検索語と，上位の類似語について，各月ごとの出現頻度を抽出する．得られた出現頻度に従って時系列推移のグラフを gnuplot で作成する．最後に表とグラフを含む HTML として出力する．

試作した時系列傾向分析ツールの画面を図 6 に示す．左の欄に検索語の一覧が表示される．クリックすると，右側の欄に，検索語とその関連語がならび，それぞれの語の傾向が折れ線グラフで表示される．

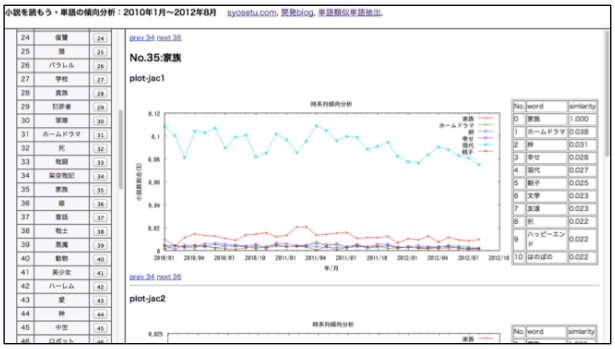

図 6 時系列傾向分析ツールの画面

5. 評価と考察

5.1 データセット

表 3 に評価実験に使用したデータセットを示す．

表 3 評価実験に使用したデータセット

記号	要素数	説明
D'	135,164	キーワード欄に単語のある小説のメタデータ集合
W'	6,484	出現頻度が 5 以上の単語集合
P'	36,804	共起出現頻度が 5 以上の単語ペア

表 3 の D'は，表 2 に示した 2012 年 9 月時点のデータから，キーワード欄に単語が無いメタデータを除外したものである．W'および P'については，出現頻度 5 以下の低頻度語を足切りしたデータセットになる．これのデータを用いて傾向分析を行う．

5.2 評価方法

提案する手法において特徴的な数値およびその変化が現れた場合，それは対象とする単語ないし単語組に関して一時的・集中的な人気がある，いわゆる流行している状態にあると判断する．今回は 100 個の検索語に対して傾向分析を行い，それら人手で定性的に評価した．

評価で用いた 100 個の検索語を表 4 に示す. 表 4 の語は，出現頻度が 51〜150 位に該当した単語である．高頻出語は類似語が多過ぎるため評価に用いにくい．低頻出語すぎる語も類似語が分かりにくい．特徴的かつある程度の出現頻度を持つ語を調べたいため，51〜150 位の単語を選定した．

表 4　評価実験に用いた検索語

1	ミステリ	26	パラレル	51	推理	76	初恋
2	転生	27	学校	52	記憶喪失	77	大自然
3	らぶらぶ	28	貴族	53	戦記	78	トリップ
4	日常	29	犯罪者	54	国家/民族	79	主婦
5	年の差	30	軍隊	55	純愛	80	恋愛？
6	三角関係	31	ホームドラマ	56	ドラゴン	81	失恋
7	オカルト	32	死	57	その他	82	性転換
8	妖怪	33	戦闘	58	魔術	83	学生
9	実話系	34	架空戦記	59	ノンフィクション	84	片思い
10	小学生	35	家族	60	ミステリー	85	吸血鬼
11	社会問題	36	姫	61	幽霊	86	人生
12	ガールズラブ	37	童話	62	殺人鬼	87	心
13	ギャグ	38	戦士	63	R15(15禁)	88	別れ
14	ロマンス	39	悪魔	64	ツンデレ	89	霊界/地獄/天国
15	夢	40	動物	65	昭和	90	近未来
16	勇者	41	美少女	66	らぶえっち	91	精霊
17	恋	42	ハーレム	67	宇宙	92	自殺
18	モンスター	43	愛	68	ライトノベル	93	王子
19	切ない	44	神	69	海	94	スポーツ
20	天使	45	中世	70	歴史	95	旅
21	霊	46	ロボット	71	騎士	96	魔法使い
22	チート	47	剣	72	最強	97	雨
23	オリジナル	48	主人公最強	73	SF	98	ゲーム
24	復讐	49	異世界トリップ	74	夏	99	ヤンデレ
25	猫	50	アクション	75	妖精	100	妹

5.2.1 類似度計算手法の比較

類似度の計算については，コサイン類似度と Jaccard 係数の二種類に対して，IDF を掛け合わせるか否かの計 4 つで計算した．例として「家族」の類似語を表 5 に示す．

表 5　「家族」の類似語（上位 10 個）

(1) Jaccard

No.	word	simlarity
0	家族	1
1	ホームドラマ	0.038
2	絆	0.031
3	幸せ	0.028
4	現代	0.027
5	親子	0.025
6	文学	0.023
7	友達	0.023
8	死	0.022
9	ハッピーエンド	0.022
10	ほのぼの	0.022

(2) Jaccard x idf

No.	word	simlarity
0	家族	1
1	ホームドラマ	0.294
2	絆	0.246
3	幸せ	0.221
4	親子	0.205
5	友達	0.185
6	現代	0.184
7	兄弟	0.178
8	死	0.173
9	心	0.171
10	感動	0.163

(3) Cosine

No.	word	simlarity
0	家族	1
1	ホームドラマ	0.085
2	絆	0.07
3	親子	0.064
4	現代	0.062
5	幸せ	0.061
6	ほのぼの	0.061
7	文学	0.053
8	兄弟	0.052
9	ハッピーエンド	0.051
10	父	0.051

(4) Cosine x idf

No.	word	simlarity
0	家族	1
1	ホームドラマ	0.656
2	絆	0.555
3	親子	0.524
4	幸せ	0.49
5	妻	0.463
6	父	0.455
7	兄弟	0.435
8	現代	0.426
9	ほのぼの	0.401
10	母親	0.394

表 7 を見ると，IDF を掛けない Jaccard 係数による類似度と，コサイン類似度では，「ハッピーエンド」という高頻出語が上位になる．IDF を掛けた値では，高頻出語が上位にならず，特徴的な語の類似度が高い. コサイン類似度に IDF を掛けたものでも，高頻出語が上位になる場合がある．Jaccard 係数に IDF を掛けたものが，出現頻度による差が比較的小さく安定しており，またより多くの特徴的な単語をとらえていると判断した．

5.2.2 期間の制限

調査した期間について述べる．当初，「小説家になろう」のサービス開始時からの傾向を分析した．その結果，2009 年後半に，値が大きく変化するグラフが多く出現した．値が大きく変化するグラフの例を図 8 に示す．

原因を調べた結果，２つの事がわかった．2009 年後半に「小説家になろう」サイトの障害が発生しており，その結果，小説投稿数が減っていた．また，2009 年 11 月から，200 文字以下の小説は投稿できなくなる規制が導入された．これら２つの影響でキーワード小説投稿数が激減し，図 8 に示すよう 2009 年 11 月から単語の出現数傾向も激減している．なお，2010 年 1 月以降は安定している．

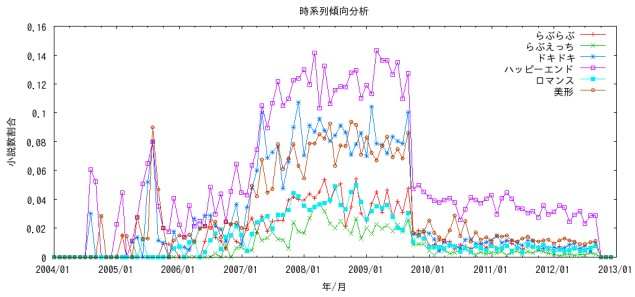

図 7　2009 年後半の急落傾向の例

5.3　傾向分析事例

以下に，調査した 100 単語のうち，興味深い傾向を示したものを列挙する．傾向が安定した 2010 年 1 月～2012 年 9 月の期間の傾向だけを，グラフとして出力している．

5.3.1 例 1 ：夢

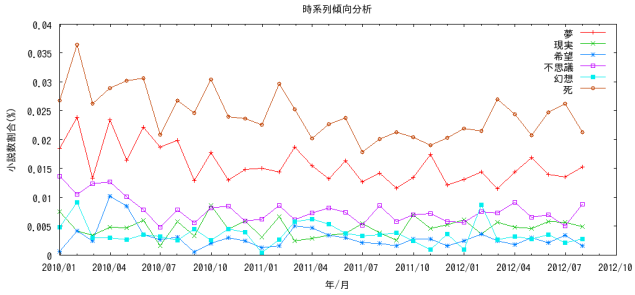

図 8　「夢」と類似語の傾向

多くの単語は月毎の変動が大きく全体傾向が判別しづらい．この単語では，類似語として挙げられた「死」とともに，全体的に減少傾向を示している．類似語として抽出した単語が似た増減傾向を示しているのは，その 2 つの関連度が大きい．

5.3.2 例 2 ：ハーレム

例 1 とは逆に，検索結果の複数の単語で増加傾向が見られる．このグラフ傾向は，「ハーレム」「チート」「主人公最強」の要素を合わせ持った小説が最近流行していると判断することができる．

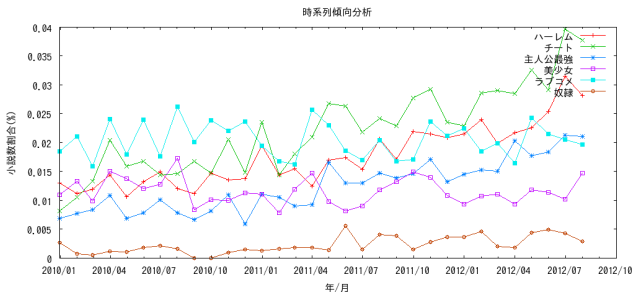

図 9　「ハーレム」と類似語の傾向

5.3.3 例 3 ：ツンデレ

「○○デレ」という単語が類似語として並んで表れている．赤線の「ツンデレ」は横ばいまたは下降気味であるのに対し，「ヤンデレ」が上昇傾向を示していることから，これらの流行が取って代わる形で訪れているのではないかという予測ができる．

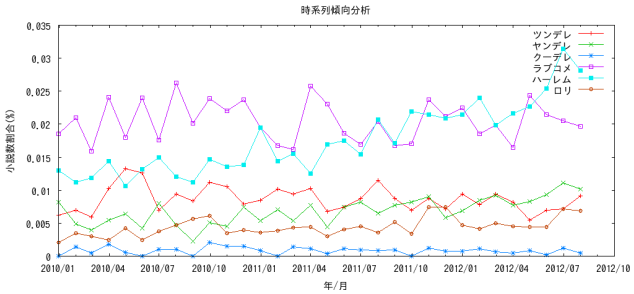

図 10　「ツンデレ」と類似語の傾向

5.3.4 例 4 ：海

類似語として表れた「夏」の推移が非常に特徴的で，グラフから分かるとおり夏の時期に大きく偏って出現している．他にも季節に関する要素は該当する時期に偏る傾向があり，No.97:雨の類似語として表れる「梅雨」は 6 月にピークを持っている．

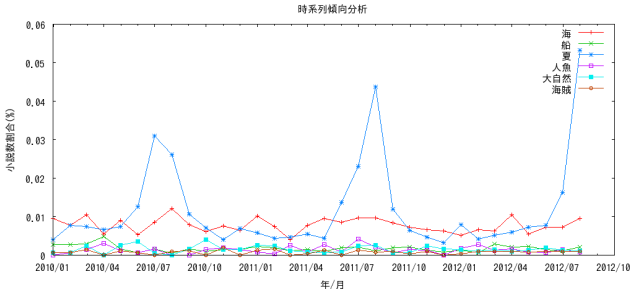

図 11　「海」と類似語の傾向

5.3.5 例 5 ：ゲーム

分かりやすい上昇傾向を示した単語のひとつ．類似語も含め急上昇を見せている．最近流行の単語だと推察できる．また，類似語には「RPG」「オンライン」など，ゲームのジャンルや形態を示すものが多く表れている．

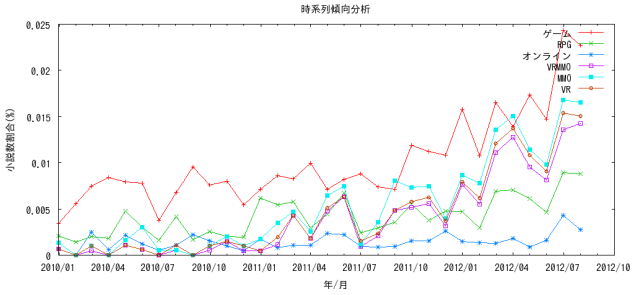

図 12　「ゲーム」と類似語の傾向

6. 関連研究

Yin, Hirokawa らは，日本の科学研究費申請書を対象に，研究分野の動向を時系列 分析する手法を提案している[9]. また，その手法および分析システムを試作し，e ラーニング分野における科学研究費申請動向の分析結果を評価している．彼らが文献で扱っている科学研究費申請書は，大学等に属する研究者が申請するものである．日本の研究者数は多いものの，研究テーマ・研究分野は多様であるため，一つの研究テーマに携わる研究者の数は，それほど多くない．そのため，時系列傾向分析では，キーワードが疎に出現する事になる．図 13 に Hirokawa らの検索システムを示す．

一方，本論文の対象であるオンライン小説は新しいサービスであるため，現在までの所，廃れたキーワードは見つからない． これは，人気ランキングの影響であると考えられる．読者は人気の高い小説を読もうとする．ファンタジー等の人気が高い分野の小説には多くの読者が付くため，ファンタジー小説への評価値は集まりやすく，その結果ランキングも上昇するという正のフィードバックが発生している．そのため，作者も人気が高い分野の小説を書き，かつ人気分野の単語をキーワード欄に付ける傾向がある．

Top 5 words of each year with respect to the score of words
T : the number of years when the word was in the top rank
L : the number of articles that contain the word in those years
G : the total number of articles that contain the word

```
   2005          2010       T word (L/G)
. . . . . . . 1 . . 1 questionnaire(1/18)
. . . . . . . 1 . . 1 tam(1/3)
. . . . . . . 1 . . 1 lms(1/7)
. . . . . . . 1 . . 1 heis(1/1)
. . . . . . . 1 . . 1 instructors(1/10)
. . . . . 1 + . . 1 puzzle(1/9)
. . . . . 2 + . . 1 smart(2/62)
. . . . . 2 . + . . 1 nsf(2/10)
. . . . . 2 . + . . 1 disciplinary(2/18)
. . . . . 2 . + . . 1 cyberization(2/14)
. . . . 4 . . + . . 1 national(4/133)
. . . . 2 . . + . . 1 greenston(2/22)
. . . . 2 . . + . . 1 professional(2/62)
. . . 9 . . 4 + . . 2 students(13/96)
. . . 4 2 . . + . . 2 teachers(6/28)
. . . 5 . . 1 + . . 2 classrooms(6/18)
. . . 3 . . . + . . 1 americans(3/14)
. . 1 . . . . + . . 1 ldt(1/1)
. . 2 . . . . + . . 1 pedagogical(2/11)
. . 1 . . 4 . + . . 2 nsdl(5/25)
. . 1 . . . . + . . 1 escience(1/2)
. 2 . . . . . + . . 1 hypermedia(2/39)
. 5 . . . . . + . . 1 learners(5/145)
. 3 . . . . . + . . 1 instructional(3/78)
. 3 . . . . . + . . 1 assembly(3/61)
. 14 6 13 9 5 10 + . . 6 education(57/165)
```

図 13 Milky way trend の例

期間毎の単語の出現数を数え上げてグラフとして出力するという手法は，Google でも取り入れられている．Google Trends [10]は入力された単語の検索要求数をグラフで示す機能があり，その単語を含むニュースと結びつけることでどのような出来事があったときに検索数が増加したかを知ることができる．また，世界中で利用されている Google に特徴的な機能として，期間のみならず，国・地域にも範囲の指定を入れて傾向を見ることが可能になっている．

図 14 に Google trends の例を示す．図 8 の例では，前章で示した No.98 と同じ期間(2010 年 1 月から 2012 年 9 月)，同じ語（ゲーム・RPG・オンライン・VRMMO・MMO）の傾向を示した．本研究で行った関連語の検索・出力はこのようなシステムをさらに拡張させることの提案と言え，適切な関連語の選択が行えれば，1 つの単語から，大きなジャンルでの流行を見ることも可能になるだろう．

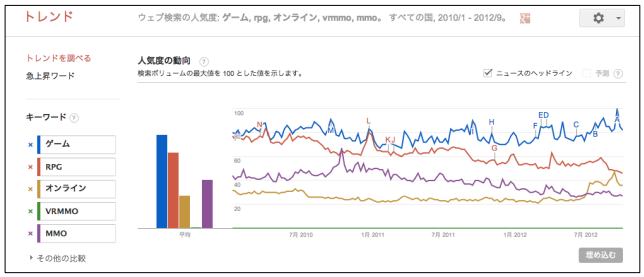

図 14 Google トレンド
（期間：2010 年 1 月～2012 年 9 月，単語：ゲーム・RPG・オンライン・VRMMO・MMO）

頻繁に上下する値を滑らかな変化として評価する手法として，時間毎に変化する値の平均を取っていくという方法（移動平均）がある．実際に社会でこれを利用しているものに株式市場があり，主に 13 週平均，26 週平均といった値が市場の動向を予測するために用いられる．本研究でも単語の出現頻度の推移を分かりやすくために採用することを検討したが，あらかじめグラフの傾向が予測された実験では無かったので，あえて移動平均を用いない状態での傾向分析を試みることにした．また，移動平均は過去の数値を計算に加える特性上最新の数値から遅れた推移を見せることになるので，短期的な流行を捉える際には考慮が必要である．

従来の協調フィルタリングによる推薦方法では，推薦の精度を最適化することに重点を置いているが，すでに知っているアイテムが多く推薦されるという問題がある．Hijikata らは，新しいものを推薦する尺度として novelty の概念を提案し，それらを実現する 3 つのアルゴリズムを提案・評価した[11]．Hijikata らの研究では，利用者にとって新規性があり興味のあるコンテンツの推薦を行うことに重点を置いている．ノベルティの概念と，傾向分析を組み合わせる事で，自分が詳しくない分野の小説を発見できやすくなる可能性がある．

7. おわりに

検索は，しばしば新しく知識を得るために行われるため，ある検索語を与えたときに，その検索語に詳しくない人でも新鮮で多様な知識を得られることが望ましい．本実験においては，時系列順の出現頻度による単語の流行の視覚化

および，与えられた検索語に対して類似度の高い言葉を計算し関連語として出力を試みた．これらのプログラムを用いて検索を試みた結果，いくつかの単語に関して，特徴的な出現頻度の傾向およびそれに連動したような傾向を示す関連語の出現を確認することが出来た．

本実験の今後の目標としては，移動平均などの手法によるグラフの傾向分析や，IDF 以外に全体での出現頻度の大小による差の大きさの改善方法の検討が挙げられる．その他にも，単語の選定や，試みたもの以外の類似度の計算法による結果の変化を検討したい．

謝辞　本研究は JSPS 科研費 2350099 の助成を受けたものである．

参考文献

1)　ヒナプロジェクト社: 小説家になろう. http://www.syosetu.com/ .
2)　ヒナプロジェクト社: 小説家になろう API, http://dev.syosetu.com/man/api/ .
3)　起点. http://www.qidian.com/.
4)　Murakami, N. and Ito, E.: Emotional video ranking based on user comments. Proc. of ACM iiWAS2011, pp. 499–502, ACM (2011).
5)　村上直至, 伊東栄典: 動画投稿サイトで付与された動画タグの階層化, 情処研報 Vol. 2010-MPS-81, No. 17, pp. 1-6 (2010).
6)　Baba, K., Ito, E., and Hirokawa S.,: Co-occurrence analysis of access log of institutional repository. Proc. of JCAICT2011, pp. 25-29 (2011).
7)　Eisuke Ito, Hirokawa, S. and Shimizu, K.: Introducing faceted views in diversity of online novels, Proc. of ICDIM2012 (Seventh International Conference on Digital Information Management), pp.145-148, IEEE, (2012).
8)　Eisuke Ito and Kazunori Shimizu: Frequency and link analysis of online novels toward social contents ranking, Proc. of SCA2012 (The 2nd International Conference on Social Computing and its Applications), pp. 531–536, IEEE (2012).
9)　Yin, C., Hirokawa, S., Yau, J. Y., Nakatoh, T, Hashimoto, K., and Tabata, Y.: Analyzing research trends with cross tabulation search engine, International Journal of Distance Education Technologies, Vol.11, No.1 (2013).
10)　Google 社: Google Trend, http://www.google.com/trends/ .
11)　Hijikata, Y., Shimizu, T. and Nishida, S.: Discovery-oriented collaborative filtering for improving user satisfaction, Proc. of IUI2009, pp. 67-76. ACM (2009).