ＳＯＭとキーグラフによる、コンテンツアナリシス

伊藤貴一
慶應義塾大学大学院政策・メディア研究科
〒252-0816 神奈川県藤沢市遠藤 5322 熊坂研究室
kiichi@sfc.keio.ac.jp

Abstract:大量のデータの構造と変化を可視化し、コンテンツ分析に役立つツールを開発している。自己組織化マップとキーグラフを組み合わせることによって、大量のデータでも詳細なコンテンツ分析が行なえることが確かめられた。

1.　研究目的

・　コンテンツアナリシスツールの開発

インターネットが一般化し、多くの人々が電子掲示板、BLOG、SNS への自分の意見を気軽に書き込める時代になった。クオリティの高い作品が登場すると書き込みは加速し、口コミの連鎖を起こし、ヒット作になる。そして、ヒット作品を分析することは、時代を読むことに繋がる。この大量の書き込みの細部を踏まえながらどのように分析するのか？大量の書き込みの構造と変化をどのように計量的に捉えるか？そして、そういった分析の敷居を下げることはできないか？これが課題になっている。ゆえに著者らは、そういった課題に答えるコンテンツアナリシスツールの開発を行なっている[1]。
ここでは 2006 年にヒットしたアニメ『涼宮ハルヒの憂鬱』に関するインターネット掲示板『２ちゃんねる』における書き込みのデータを、著者の開発したツール（自己組織化マップ（SOM）とキーグラフ）を使った分析を紹介する。

・　涼宮ハルヒの憂鬱

『涼宮ハルヒの憂鬱』は 2006 年 4 月から 7 月まで全 14 回放送の深夜アニメである。独立系 UTF 局のみの放送で、放送時の平均視聴率２％程度であったにも関わらず、アニメ放映開始後、原作のライトノベルは半年で 150 万部を売り、主題歌や挿入歌集はチャート上位に続々進出など様々な記録を打ち立てた。この大ヒットには口コミが強く作用したといわれており、現に『２ちゃんねる』のアニメ板の『涼宮ハルヒの憂鬱』に関するスレッドは書き込み量が非常に多く、番組終了の 7 月中旬までの 3 ヶ月間で、約５６０スレッドを消費している。図１はスレッドの消費を週ごとに表したものである。この中で何が語られ、どう変化していったのか、これを計量であらわすことができないか、というのが本研究の課題である。

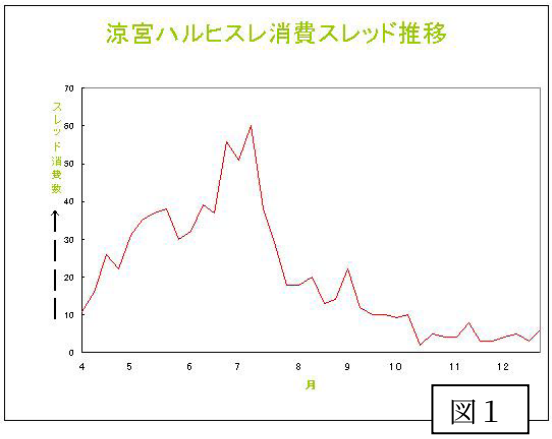


図１

2.　データの取得・加工

２ちゃんねるの過去ログを読める、「にくちゃんねる」（http://makimo.to/2ch/）からデータを取得。基本的には、一つの書き込みあたり、一つの塊とみなす。２ちゃんねる専用のブラウザでは書き込みに「>>100」などと書くと、100 番目のレス（書き込み）を参照することを意味する機能がある。ただの一つの書き込みでは関係性は取りづらいと思われるので、参照先も含めた形で一つの塊とした。しかし、多重参照は認めない。また、慣例として、スレッド開始から 10 番目くらいまでは、定型文章が書き込まれるため、これらを含まないようにした。このようなデータに対し、茶筅を使い形態素にした。そして、表記ゆれや類義語をまとめるなどの処理をした。

3.　分析手法

シンプルな出現頻度の数え上げ、という手法を乗り越えるために、キーグラフ[2]を用いる。キーグラフは文章データの可視化手法の一つである。上位頻度の単語から、共起確率を使い単語をクラスタ化し、クラスタとクラスタを結びつける、キーワードを探索するアルゴリズムである。この性質により、仮に低頻度の単語であっても、重要な単語の抽出を期待できる。

しかしキーグラフは、巨大なデータには十分な効果を発揮できない性質がある。一つは、ネットワークの形状が single-scale ネットワークを想定したアルゴリズムである[3]ためだ。scale-free ネットワークのような、有名なものはどんどん有名になっていく（優先的選択）という性質のあるものに対しては、キーグラフは有効に機能しない。上位頻度のもの同士が強くクラスタ化してしまい、巨大なクラスタと小さなクラスタを結びつけるものや、共起確率が高くなる低頻度のもの同士でのブリッジしか見つけられなくなってしまうからである。[3]では、頻度の高い単語を削除していけば、single-scale ネットワークに近づき、キーグラフで分析可能だとしているが、助詞、助動詞などの不要語を除いた後では、そのような単語はむしろ重要語であることが多い。本対象のような 2 ちゃんねるのスレッドのデータでもこの問題は発生している。「ハルヒ」「長門」「みくる」といった主要登場人物は頻度が高く、語られる内容も多彩になるためネットワーク的には集中ハブになってしまうと傾向がある。同時に、頻度が高すぎるため、他の単語との共起関係は相対的に弱くなってしまうため、ノード数が増えていったとき、孤立してしまう傾向にある。

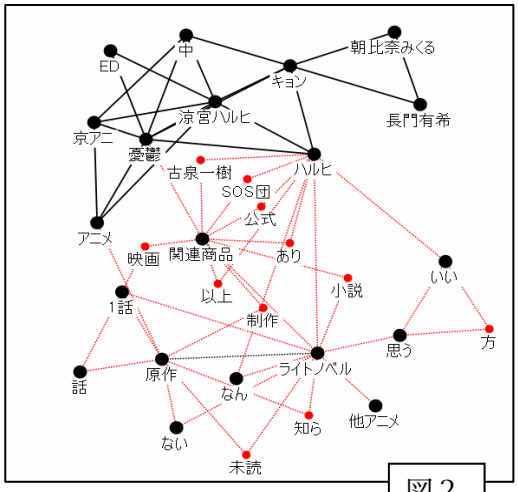


図２

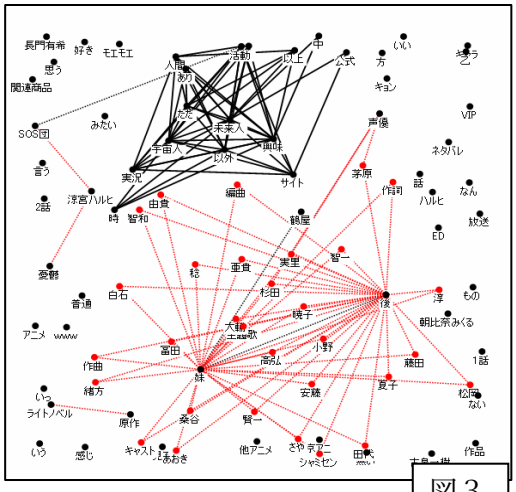


図３

図２と図３は涼宮ハルヒの憂鬱が開始して半月(4 月 16 日)までの２ちゃんねるのデータをキーグラフ化したものである。図１では「ハルヒ」「キョン」「みくる」といった主要登場人物たちによる巨大なクラスタと単独のノードとの関係のみを表している。一見きれいなグラフになっているようだが、約９メガバイトのファイルからの出力としてはシンプルすぎる。もっと様々な関係を見たいが、ノード数を増やしていくと図２のようになってしまい、解読不能に陥る。

このような問題を解決するために、自己組織化マップ[4]を使い、事前にデータを分割する。語られる主要ワードで大雑把に分類してから、キーグラフにかけることがこの目的である。そもそも『涼宮ハルヒの憂鬱』スレッドとなっているが、語られる内容は幅広いため、語られる内容ごとにキーグラフを作ろうという戦略である。

自己組織化マップへのインプット情報は、図４のように、高頻度の単語から順に分類に効力を発揮しそう 100 個程度の単語を選択し、それを特徴ベクトルにした。それぞれの文章ごとに、その単語が含まれていれば１、なければ０と判断し、最後は平均と分散による標準化を行った。平均は単語の出現確率になるため、標準化は低頻度の単語の影響力を増やすために行なう処置である。

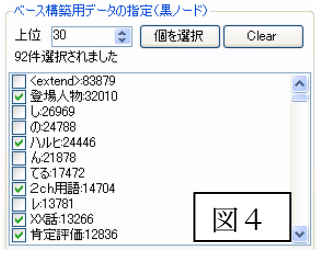


図４

このようなデータを用い、自己組織化を行った。そのあと、マップのベクトルを元に、K-means 法でクラスタリングを行った。

図 5 は、「京アニ」という属性での重みの分布を示す（色のついているところは値が高い）。このように、単語ごとに固まって配置されるようになる。その「京アニ」（当アニメのアニメーション制作会社「京都アニメーション」の略称）の属性が反応しているクラスタに所属している文章を見ると図 6 のようになる。

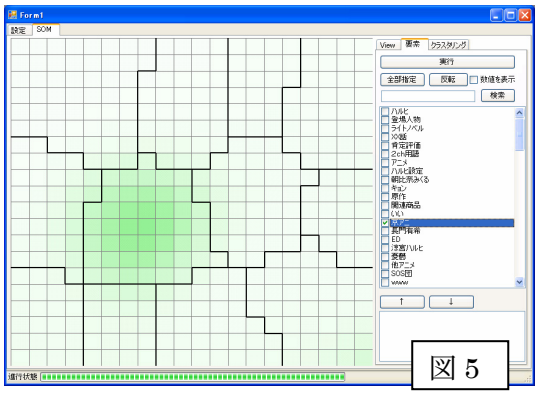
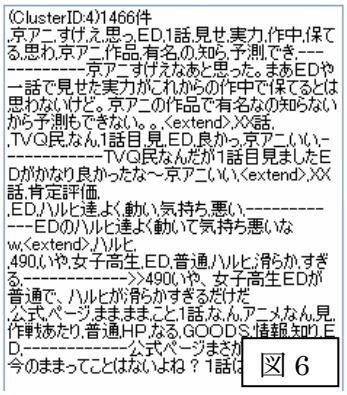


図 5


(ClusterID:4)1466件
,京アニ,すげえ,思っ,ED,1話,見せ,実力,作中,保てる,思わ,京アニ,作品,有名,の,知ら,予測,でき,----------京アニすげえなあと思った。まあEDや一話で見せた実力がこれからの作中で保てるとは思わないけど。京アニの作品で有名なの知らないから予測もできない。。,<extend>,XX話,,TVQ民,なん,1話目,見,ED,良かっ,京アニ,い,----------TVQ民なんだが1話目見ましたEDがかなり良かったな～京アニい,い,<extend>,XX話,肯定評価,,ED,ハルヒ達,よく,動い,気持ち,悪い,----------EDのハルヒ達よく動いて気持ち悪いなw,<extend>,ハルヒ,,490,いや,女子高生,ED,普通,ハルヒ,滑らか,すぎる,----------＞＞490,いや、女子高生EDが普通で、ハルヒが滑らかすぎるだけだ,公式,ページ,まま,こと1話,なん,アニメなん,見,作画,あたり,普通,HP,なる,GOODS,情報,知り,ED,----------公式ページまさか今のままってことはないよね？1話は
図 6

このクラスタに所属しているデータをキーグラフにかけると図 7 のようになる。

キーグラフは線で繋がったものを想像で補いながら読んでいけばいい。読んでいくと「よくわけのわからないアニメだが、京アニの作画がすばらしく（神）、特に ED（エンディングのスタッフロール）のアニメが動きすぎで、いい作品だと思う。駄目な（糞）アニメだといっている人もいるけどね」このような感じであろうか。

解説すると、ハルヒの第一話は、なんの人間関係や設定の説明もなく、登場人物たちが作った素人っぽい自主制作映画の上映がいきなり始まる、という構成をしていた。原作を知っている人にはニヤリとするものだったが、原作を知らない人たちには、まさしく、よくわけのわからないアニメだった。

しかし、作画のクオリティは高く、素人っぽさをうまく演出していた。エンディングでは、テレビ放送では珍しいフルアニメーション（動きを省略しないアニメーション）で製作されたダンスが評判となり、後に実際に踊る人まで現れた。注１）
このように、図２，３のキーグラフより明らかにディティールを拾い出すことができている。

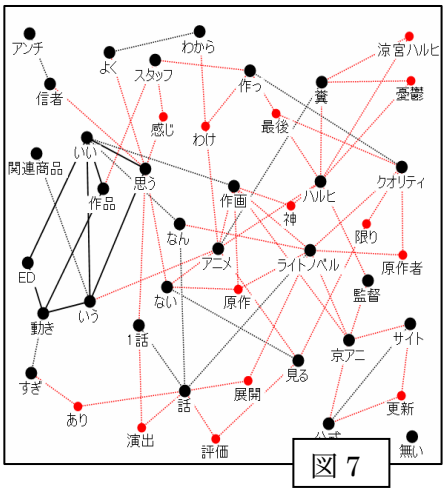


図７

4.　課題と今後の展望

このように個々のクラスタをそれぞれキーグラフにしていけば、部分的なディティールが得られるのである。しかし、この手法は、あくまで部分に過ぎず、全体的に統合したときどのようになるのか、という視点が欠けている。これが課題の一つである。キーグラフのコンセプトの延長上になるが、クラスタとクラスタを繋ぐものは何か?という分析も必要であろう。
自己組織化マップでの分類についての問題点は、上位頻度の単語を元に分類しているため、分類網に引っかからないものが多く出てしまう点である。（今回の事例では全体の 46.2%）これらはいったいどういう意味があるのか、という疑問がある。おそらくスレッドに流れる空気が関係するのだと思われる。これに関連し、どのような言葉を特徴ベクトルにしていけばいいのか、という課題もある。
また、『涼宮ハルヒの憂鬱』に関しては、今回示したのは、初回の分析に過ぎない。これが変化していくさまを時系列的に可視化することがこの研究を始めた動機である。キーグラフが変化していく様を捉える、紙芝居キーグラフという手法[5]がすでにあるため、これとの組み合わせを期待したい。

今回使ったツールは公開予定である。
キーグラフのツールはすでに、http://web.sfc.keio.ac.jp/~kiichi/kamisibaiwiki/　で公開している。

[1] 木幡敬史，渡邊紀文，岡本潤，小野田哲弥，伊藤貴一：コラボレーション!―SFC という「融合の現場」,2007，慶應義塾大学出版会，1 章 Case No.4「IS2」P66-80
[2] 大澤幸生：チャンス発見の情報技術―ポストデータマイニング時代の意志決定支援,2003，東京電機大学出版局
[3] 松尾豊：予兆発見とスモールワールド,2003，人工知能学会誌、Vol.18，No.3
[4] Teuvo Kohonen：Self-Organizing Maps ,2001,Springer-Verlag
[5] 大澤幸生：チャンス発見のデータ分析―モデル化+可視化+コミュニケーション→シナリオ創発,2006，東京電機大学出版局
注１）Youtube (http://www.youtube.com/)で「haruhi dance」と検索すると多くの踊っている動画を見ることができる。（2007 年 5 月）