Gunosyにおける言語処理応用

関 喜史 * [1,2]

[1] 株式会社 Gunosy
[2] 東京大学

1 はじめに

グノシーは株式会社 Gunosy が提供するスマートフォンアプリを中心に展開する情報サービスである．もともと 2011 年に Gunosy というサービスを当時データマイニングを研究していた学生 3 人がリリースし，それを 2012 年に株式会社 Gunosy として法人化し，2014 年春に「グノシー」としてリニューアルした．2016 年 2 月現在で 1300 万 DL を記録する国内最大級のニュースアプリケーションである．当サービスは複数のウェブメディアから提供されたコンテンツを取捨選択しサービスの利用者に提供しており，キュレーションメディアとも呼ばれる．Gunosy ではコンテンツの取捨選択を編集者の意思決定ではなく，自動的にもしくは半自動的に行っている．ウェブにおけるコンテンツは大半がテキスト情報であり，取捨選択の自動化のために言語処理技術が用いられている．本稿ではグノシーとはどのようなサービスであり，どのように変化してきたのかを述べる．言語処理技術を用いたシステムを提案するような論文は数多くあるが，そのシステムが人々に利用され拡大していく中でどのように変化してきたのかについてはあまり議論されておらず，本稿はそのようなサービスのケーススタディとして有益であると考えている．

グノシーのシステムはサービス開始時から 2014 年 3 月までの情報推薦を基盤にしたシステムと，それ以降のコンテンツ評価を基盤にしたシステムの 2 つに分けられる．本稿では初期のものを第 1 期, 現在のものを第 2 期として扱うこととし，まず 2 章において第 1 期のシステムとその当時の状況やサービスの成長に伴う変化について述べる．次に 3 章において第 2 期のシステムと，その中でどのような課題があり，どのように解決してきたかについて述べる．

2 第 1 期のシステムについて

2.1 システムの概要

グノシーは 2011 年の 9 月にサービスを開始した．第 1 期のシステムは Facebook, Twitter を連携してユーザ登録を行うとユーザの投稿内容からユーザの興味関心を分析し，その興味関心に適したウェブコンテンツを提供するサービスを提供するものであった．現在は「マイニュース」という名称でサービス内の 1 コンテンツとして提供されている．当初は 1 日に 25 件の URL のリストをユーザごとに生成しており，ユーザは指定した時間に配信されるメール，もしくはウェブサイトにログインすることで生成されたコンテンツリストをみることができた．メール, ウェブページ共にサムネイル画像, タイトル, スニペットが記載されており，クリックすることでリンク元のページへ遷移することができる．その際のクリック行動はユーザの興味関心情報として蓄積され翌日以降のコンテンツリストの生成に活用される．

システムとして見たとき第 1 期のグノシーは以下の要素で構築されている．

- URL をウェブ上から収集する
- 収集した URL を特徴量に変換する
- SNS の投稿情報からユーザの特徴量を構築する
- ユーザの特徴量と URL の特徴量の距離を求め，URL のリストを生成する

このようにオーソドックスな内容ベースフィルタリングのシステムとなっている．次節ではこのシステムがサービスを運営していく上でどのように変化していったのかを述べる．

*yoshifumi.seki@gunosy.com

2.2 ケーススタディ

2011 年のリリース後ソーシャルメディアでの投稿や，ブログ記事での紹介などを通じてユーザ数を伸ばした．このころ IPA 未踏プロジェクトの支援をうけており，コンテンツリストの多様性を高めることによりユーザの継続率を向上させることに成功している．このような実績を背景に 2012 年 11 月に法人化を行った．当時 2012 年末で登録ユーザ数は 4 万人に達している．その後 2013 年 1 月に iOS アプリを, 2013 年 4 月に Android アプリをリリースした．それまでメールが利用ユーザの大半を占めていたがアプリのリリース翌日からスマートフォンアプリケーションからの利用がユーザの半数以上を占めるようになり，以降サービスはスマートフォン中心にシフトしていくことになった．

このころからサービスの登録ユーザがアーリーアダプタと呼ばれる層から変化してきていた．その変化からいくつかの問題が発生しそれに応じてサービスを変化させてきた．ここでは特に言語処理システムとして特徴的な 2 点の問題を紹介する．

1 点目が SNS の活用度合いが低いユーザの登録割合が増えてきたことである．SNS 経由の登録やアカウント連携を促す段階でアプリケーションをダウンロードしたが登録しないユーザが増加していることや，SNS の投稿数が少ないユーザの継続率が低いことが明らかになった．そのため登録の際に SNS を必須にせず，ユーザに読みたい記事を選んでもらうアンケートを実施することによって興味関心の情報を獲得する登録のフローを開発し，メールアドレスでの登録のフローを追加した．その後アンケート機能の改善に取り組み，その登録フローから SNS 経由並の継続率を得てまもなくメールアドレス登録による登録が中心になり，SNS 経由の登録のフローはアップデートのうちに廃止された．パーソナライズを行うような言語処理技術を用いたシステムでは，ユーザの情報を獲得したり入力させたりするが，その方法はサービスの成長に合わせて変えていかなくてはならない．

もう 1 つの大きな問題が読まれるコンテンツの変化である．アーリーアダプタが新規ユーザの大半を占めていた頃は，ビジネス・Web サービス・デバイス・デザインなどが読まれる記事の中心となっており，当初からシステムの評価をユーザからの反応の良し悪し（一人当たりのクリック数や継続率など）で行っていたため，アーリーアダプタの興味関心を抽出しやすく，彼らが好むコンテンツをより配信しやすくなっていた．しかしユーザ層の拡大によって読まれるコンテンツにも変化が生まれてきた．具体的に読まれるようになったコンテンツとしては，恋愛や健康に関するコラムやエンタメ・スポーツ情報などが挙げられる．登録のフローにアンケートを導入したことでそのような興味関心をもっているユーザが増えていること，そしてそのようなユーザに対して適切なコンテンツを提供できておらず，継続率が低くなっていることが明らかとなった．この課題は「興味関心を抽出できていない」「配信可能なコンテンツが無い」という 2 つの問題から生まれている．まずは配信可能なコンテンツを増やすという対策を行った．この際多くのニュースサービスをもつポータルサイトを調査し，それらのポータルサイトで掲載されるコンテンツを充足することを評価指標として，それを高めるようなコンテンツを追加していった．これはユーザ層がポータルサイトにアクセスするような人々に近づいているという仮説からである．

配信可能なコンテンツが増えてもそれが配信されないと解決にはならない．コンテンツを配信するためには,「コンテンツから特徴量が適切に抽出されること」「ユーザからそのようなコンテンツにつく特徴量が得られること」の 2 点が満たされなくてはならない．そのためにまず収集したコンテンツに適切な特徴量が与えられるようにしなくてはならない．コンテンツへの特徴量の付与には教師あり分類を用いていたが，それを改善するためにはよりよい教師データが必要である．サイト内から抽出したカテゴリ情報や, クラウドソーシングを用いて付けたラベルを教師データとすることで, コンテンツに特徴量を与えられるようにした．

そしてユーザに特徴量を与えることができるようにするために，登録導線で選んだカテゴリに対して一定期間ルールで対象カテゴリに関連する記事を配信するようにした．その結果継続したユーザの特徴量はクリックのフィードバックによりよい特徴量を得たユーザが現れた．それらのユーザの特徴量を同じようなカテゴリを選んだ新規のユーザに与えることで，初期登録時にユーザの特徴量が適切な状態になっているようにした結果，そういったカテゴリを選んだユーザの継続率が改善した．

またすでに登録していたユーザについてはそのようなコンテンツが過去配信されていなかったことや，ユーザ特徴量を得づらくなっていたことから，そのようなカテゴリでサービス内で人気となった記事を一定確率で配信リストに追加することで，これまで得られていなかった興味関心データを既存ユーザについても付加できるよう試みた．

これらのケーススタディから言語処理のシステムを実社会に提供する上では，サービスの成長にどのよう

にシステムを追従させていくかが課題になるということが示唆される．とくに BtoC のシステムの場合，開発時に対象としているようなユーザに対してはモデルの改善などでシステムをよりよくしていくことが可能であるが，ユーザ層が変化していくとコンテンツを提供するモデルやユーザの情報を抽出するモデルではなく，そもそものユーザに与える体験やモデルを構築するためのデータソースの部分を見直していかないと，サービスの成長にシステムが追従できない．そしてその追従のためにはよりよいモデルではなく，アイデアであったり人手での改善が効果的な場合が往々にしてある．言語処理にかぎらず，統計的なモデルを用いたようなシステムではユーザの体験の変化は定性的に追いにくい．このような変化を適切に捉え，広い視野に立った改善を行っていくことがサービスを成長させ，よりよい体験をユーザに提供していく上では重要であると考える．

3　第 2 期のシステムについて

3.1　システムの概要

本節では第 2 期の概要について述べる．2014 年 3 月にグノシーは大規模なシステムアップデートを行った．アプリ名の「Gunosy」から「グノシー」への変化や，アルファベットの G を用いたアイコンから紙飛行機のアイコンに変わったのも同一時期である．背景としてはテレビ CM をはじめとする大規模プロモーションに向けて，より広い範囲のユーザが関心を持ってもらうプロダクトにする必要があったことがある．

具体的には「3 分で旬のニュースをまとめ読み」をキーワードに世の中で話題となっているニュースを閲覧できるアプリケーションを目指すこととなった．システムとしては以下の要素で構築されている．

- URL をウェブ上から収集する
- 収集した URL をカテゴリに分類する (教師あり分類)
- カテゴリ内で同一の話題を述べている URL をまとめる (教師なし分類)
- 話題の中でユーザに提示する URL を決定する
- 話題を評価し，並び替えて URL のリストを生成する

第 1 期のシステムであったユーザとのマッチングの要素が排除され，その分よいリストを生成するための仕組みが多く導入された．例えるなら新聞の各ページを自動で生成するようなシステムを目指したものである．

この中で我々がとくに重要だと考えている点は，同一の話題をまとめる教師なし分類と，話題の評価である．特に教師なし分類はユーザ体験を高めるためには幅広い話題をリスト内で提供すべきという仮説が背景にあるが，それだけではなく同一の話題に分類される URL の数を，その話題がどれだけ注目を集めているか変数として用いていることもありその話題の評価にも影響を与えている．

話題の評価については大きく 2 つの評価を行っている．1 つはその話題がどれだけ社会的に重要であるのかという点であり，もう 1 つはその話題がサービス内でどれだけ注目を集めているかという点である．後者については新しい記事には適用できないため，前者の指標をベースにしながら後者の指標も評価に加え，その評価が総合的に高い話題をリストの上部に表示するようにしている．

次節のケーススタディでは教師なし分類の事例を元にサービスを運営する中でどのようにシステムを改善してきたかを紹介する．

3.2　ケーススタディ

最初期の教師なし分類においては単語ベクトルによる文書間の距離を規定し，その距離が一定以上短いものを同一の話題として捉えて対象となるカテゴリ内すべての組み合わせを計算した上で話題を特定するようにしていた．

実際にサービスを運用していく中で，極端に多い URL を持つ話題が出ることが度々起こった．この原因は「今週の人気だった記事」というような多くの URL へのリンクを含むようなコンテンツが，様々な URL と薄く結びつき大きな話題を形成することであった．特にスポーツの試合結果などは他球場で行われた試合結果なども掲載することが多くあり，そういった URL が原因となって大きな話題が形成されてしまうことになった．大きな話題が形成されると本来リストに表示されるべき複数の話題が 1 つの話題になってしまうため，リストに表示されるコンテンツが少なくなってしまう．また話題に含まれる URL の数は外部での注目度として話題の評価にも用いられているため，不必要にその話題が高く評価されることになる．

このように 1 つの URL によって複数の話題が結び付けられることを防ぐために，すべての URL の関係性を評価することをやめ，新しい URL と現在特定されている話題との距離を用いて逐次的に処理することにした．

こうした教師なし分類の手法を一部のユーザに反映した結果，とくにスポーツのタブで 5〜10%程度のクリック数の改善がみられた．この結果から教師なし分類の改善がユーザの体験を大きく改善することを明らかにすることができた．

4　まとめ

言語処理を用いたシステムはこれまで数多く提案・報告されているが，実際にそれを運用し改善し成長させている例はあまり多くない．本稿では用いた技術の詳細より，システムを運用する中でどのような課題があり，それをどのように改善し，どのように成長させてきたのかを中心に紹介した．

我々はこのように様々な改善を行っているがその中で重要視していることはその改善がユーザの体験をどのように良くしていくかということである．それはより良いモデルを使うことであっても，人手によるものであっても同等である．特にモデルの場合は実装に時間がかかる上改善がなかなか見えにくい．そのためそのモデルを用いることによってサービスで提供されるコンテンツがどのように変わるのかの仮説を出した上で，それと同じ状況を人手で再現しユーザの行動が改善するのかなどの検討を行っている．

サービスを運用する中で興味深い課題に出会うことも多い．特に第 1 期から第 2 期の移行期のは情報推薦とはどうあるべきなのか，本当に推薦は，パーソナライズは必要なのかという課題に大きく悩まされ，パーソナライズをほぼ行わないような形で現在は運用されている．しかしこれはパーソナライズが不要ということではなく，パーソナライズされることを魅力に思う人が決して多くはないということであり，パーソナライズ以前にまずコンテンツ自体を正しく評価するような仕組みが重要であることだと我々は考えている．

現在はますます多くのユーザに利用していただくべく，マンガ等多種多様なコンテンツを扱うサービスへと進化を続けている．その中でよりサービスが成長しユーザの皆様に満足していただけるために言語処理技術の重要度はますます上がってきている．今後多くの技術を取り入れながらその結果をこのように学術研究の場でフィードバックしていくことで，サービスの改善につなげつつ言語処理技術の発展にも貢献していければ幸いである．