セレンディピティを生み出す情報プラットフォームの開発

1. 背景

Twitter の公開情報を利用して、ユーザの興味を反映させることによって、読むのに 80 分かかる 1000 件のツイートを数分で読めるようにするためのツールを開発した。これは、情報の受容に知的好奇心が満たされた時に感じる、セレンディピティの醸成に最適化された新たな情報プラットフォームの開発を目的とするものである。Twitter などのミニブログはユーザが情報を生成、共有して拡散するための主要な情報プラットフォームとなっている。ユーザは、自らの知的好奇心を満たす目的で利用しているが、特定の人間をフレンドとして登録する現行の共有手法は、興味深い人間が常に一面的な興味に基づく情報を発信するわけではない点また、興味に基づく登録関係に補正しようと思っても、感情的な人的つながりに依存するために簡単につながりを更新することが出来ない点で、持続的なユーザの知的好奇心の最大化という達成目標に対して最適な仕組みとはいえない。本プロジェクトは、オンライン上に存在する大量の情報を利用可能な形に位置づけ利益を享受し管理するシステムである、推薦システムを利用している。推薦システムは、興味の意識が検索クエリの形になるほどには顕在化していないが興味を持つかもしれない場合や、既存の興味あるものと関連性があるために興味を持つに違いない場合の情報の発見に対して有用である。Twitter というツールには、フォローを増やして情報量が多くなればなるほど、便利になればなるほど、読むのに時間がかかるようになり、その結果として、重要な情報を見落とす、つまり、セレンディピティが失われる問題点がある。そこで、本プロジェクトはそれを補う形でのツールの開発に取り組んだものである。

2. 目的

Twitter などのミニブログはユーザが情報を生成、共有して拡散するための主要な情報プラットフォームとなっている。それには、一方的にフォローを増やして情報収集できるという Twitter の良い側面が作用しているためである。しかし、反面、Twitter の便利なところであるフォローによって情報量が多くなればなるほど、便利な情報がストリームに流れれば流れるほど、読むのに時間がかかるようになり、結果として、重要な情報を見落とす事態が発生している点で問題があった。そこで、本プロジェクトでは、読むのに 80 分かかる 1000 件のツイートを数分で読めるようにするためのツールを開発した。

3. 開発の内容

分類と順序付けにより、ツイート数量を圧縮する。上記から、問題はツイートの更新件数が多すぎるために、重要なツイートが埋もれてしまうことにあることがわかる。ここでいう、重要とは、客観的なものというよりは主観的なものであるべきであると考えた。なぜなら、Twitter 自体がもともと、ユーザが恣意的に選択したフォローの組み合わせである以上、そこからさらに選別する基準はユーザの趣味や嗜好にしたがった重要度に基づくべきであると考えるからである。そのため、ツイート数量を圧縮するための手法として、順序付けによる閲覧機会を傾斜させて配分させ、そのときの傾斜の仕方はユーザの Twitter 上での発言内容に基づいて形成されたユーザプロファイリングを基準に採用した。それによって、ユーザの趣味嗜好から主観的に重要と思われるツイートを優先させることが可能となる。また、埋もれてしまわないための方策の一つとして、ツイートを大きな話題を基準にグルーピングすることで、可読性の向上とユーザの主観的な判断を可能とする、という 2 つを実現できると考えた。そのために、あらかじめ用意された、9 つの分野にツイートを自動分類して、ユーザが自発的に選択しやすいインターフェイスの提供を可能としている。

4. 従来の技術（または機能）との相違

分類は、ベイズ分類により実装している。1分野に平均40弱のwikipedeia関連記事を学習データとして使用している。対象となる、それぞれの分野に属する学習データを使って、その分野に属する確率が最も高い分野を推定することが可能となる。今回ベイズ分類を選択したのには、実装が容易なことが利点として存在した。もっとも、当初、交差検定という精度評価手法で50%程度であり、実運用のレベルとしては、問題があった。加えて、ツイートは短いのも多く、十分な判定が出来ないという問題もあった。そこで、今回、コンプリメントナイーブベイズを実装した。これは、分野に属さない学習データを使って、属さない確率がもっとも低い分野を割り当てる手法である。これには、分類対象に分野のばらつきがある場合に、ばらつき量を減らせるという利点がある。そして、この他にも、頻出単語に引きづられないよう、出現回数の対数をとるなど、いくつか、ヒューリスティクスな手法も採用している。また、学習データの単語群を相互情報量の高い単語に絞る方法は試してみたが、精度に変化が余りないわりに、処理時間がかかるので今回は採用しなかった。その結果、コンプリメントナイーブベイズで精度は80%程度に向上。さらに、URLがある場合には、URLの文章も取得することで、ツイートの短さを補完している。また、「おなかすいた」「〇〇なう」などの特定の分野への指向性が低い場合は除去される。これは、特定の単語やフレーズをあらかじめ、登録するのではなく、ベイズ分類によっても分野への強い志向がなかった場合に、意味性の薄いツイートと推定することでなりたっている。

順位付けには、ユーザ発言属性により主観的な評価を基準においている。具体的には、ユーザの過去のツイート内容をデータとして使用している。当初、単純に、ユーザ発言中のそれぞれの単語の頻出数を数えて、ソートしていた。ただ、その結果として、「RT」「人」「目」「ブログ」など、一般的な単語が上位に来やすくなってしまい、これは、本来、意味のある重要なツイートを見逃さないようにしてピックアップするという趣旨からははずれるものであった。そこで、キーワード毎にTF-IDF値を計算することで、この問題にアプローチしている。

TF * log(totalUserCount/DF) TF = ユーザ発言中の頻出数
IDF = 全体の中での特徴の低さ DF = ベイズで用いた学習データと他ユーザのツイート中での頻出数

これにより、一般的な単語が上位に来にくくなった。

5. 期待される効果

ツイッターの普及により、自らのタイムラインの中から、重要な情報を見逃さないことに対する需要は増しているように思われる。本成果は、それに対するひとつの答えとして、価値がある。本成果を実現するために利用された、コンプリメント

ナイーブベイズとユーザプロファイリング情報を140文字以内の短文で構成されたツイートに適用することを可能とする、基礎的な技術を提供できたものと考える。また、これを基礎として、さらなら精度の高い技術が提供されるものと考える。

6. 普及（または活用）の見通し

本成果はiOSアプリケーションとして開発され、提供される。一般に、iOSアプリケーションはApple社の提供するApp Storeにてダウンロード可能であり、クリエータの個人的な経験則として、10000ダウンロードは比較的容易にされる。ツイッターが普及していることを考えると、本アプリの需要は高く、その10000ユーザがクチコミ等で広がれば、20000ユーザほどの利用者を早期に獲得できると思われる。

7. クリエータ名（所属）

吉牟田陽平