5S-01

Twitterデータを利用したLinked Open Data作成手法の検討

麻田 修平† 平川 豊‡ 大関 和夫‡
†芝浦工業大学大学院理工学研究科 ‡芝浦工業大学工学部

1. はじめに

近年、Linked Open Data(LOD)が注目を集めている。これは、Web サイト上で公開されているデータの参照や再利用などの利便性を向上させる取り組みである。しかし、人手でデータの LOD 化をするには、公開するデータの収集や、RDF 化(共通なデータモデルへの変換)に労力が伴う。

そこで本研究では、Web API を通して利用可能な Twitter[1]の投稿データに着目し、自動的にオンラインゲームなどのデジタルコンテンツに関する情報を収集、及びテキスト内のキーワードの関連性から RDF 化することで、LOD として利用するためのシステムを提案する。

2. Twitter API

Twitter の投稿データは、公開されている API を利用することで取得することができ、API を利用して取得したデータは引用・転載可能なデータとして利用することができる。また、投稿された本文だけでなく、投稿者の名前や、プロフィール情報、フォロー数やフォロワー数などの情報も同時に取得することができる。以下で、Twitter から投稿データを取得するための二つの API を紹介する。

2.1. キーワード検索 API

キーワードを用いて Twitter の投稿データを検索することができる。しかし、以下のような制限がある。

- 検索ができるのは、15 分の間に 180 回までである
- 検索で取得できる投稿データは一度に最大で 100 件までである
- 検索で取得できる投稿データは、一週間前のものまでである。

2.2. ストリーミング API

指定した投稿者の投稿データをリアルタイムに監視し、取得することができる。しかし、監視できる投稿者の最大数は 5000 人までである。

3. システムの課題

本研究では、特定のオンラインゲームを一つ選択し、一般に公開されているウィキ形式のゲーム攻略サイトから HTML を解析し、投稿データに付与するためのメタデータとして、4447 個のキーワードを収集した。実験的に、収集したキーワードを全て用いて、キーワード検索をしたところ、306,591 件の投稿データが収集された。

収集した投稿データを分析した結果、選択したオンラインゲームに関係する投稿データは、34,464 件であり、他は別のデジタルコンテンツに関する投稿データであった。

Twitter API の制限から、全てのキーワードを用いてデータを収集するためには多くの時間を要する。投稿データの内容を分析したところ、「○分後に△△イベントが開始」というような、投稿データが含まれることから、キーワード検索 API による収集だけでは、速報性に欠けてしまうことがわかった。リアルタイムな情報を取得するために、ストリーミング API を利用する必要性があるが、収集した投稿データの投稿者の数は、11574 人存在し、API で監視できる最大人数を超えている。したがって、監視対象となる投稿者に対して順位付けを行い、より多くの投稿データをリアルタイムに取得できるようにする必要がある。

投稿者に関しては、以下のような特徴があることがわかった。

- 投稿者の名前、またはプロフィールに、オンラインゲームの名称を含んでいる。
- プロフィールに、「○○(オンラインゲームのサーバ名)でプレイしています。」のような文体での自己紹介をしている。

また、Twitter の投稿データは、くずれた表現やスペルミスなどが存在するため、キーワードの不一致からメタデータの付与に失敗してしまい、LOD として利用できなくなる問題がある。

4. 関連研究

槇ら[2]は、Wikipedia を用いた LOD を提案しており、I-Scover という Web サイトに登録された論文データを対象に、メタデータの付与を行い、論文のクラスタリングによって検索効率の向上を図っている。しかし、Twitter の投稿データを対象にオンラインゲームのキーワードをメタデータとして付与する場合、同じキーワードが別のオンラインゲームのキーワードに含まれている場合があり、別のオンラインゲームの投稿データに対してメタデータを付与してしまう。また、本論文では前述のように、Twitter の投稿データは表記のくずれやスペルミスがあるため、キーワードの不一致からメタデータの付与に失敗してしまうことから、本論文の提案するシステムへの適用は難しい。

5. 提案手法

5.1. システム概要

本研究で提案するシステムの概要を図 1 に示す。構成された LOD は Web 上に公開する。

Study of Twitter data conversion to Linked Open Data
†Shuhei Asada, ‡Kazuo Ohzeki, ‡Yutaka Hirakawa
†Electrical Engineering and Computer Science, Shibaura Institute of Technology, Tokyo, Japan
‡Information Science and Engineering, Shibaura Institute of Technology, Tokyo, Japan

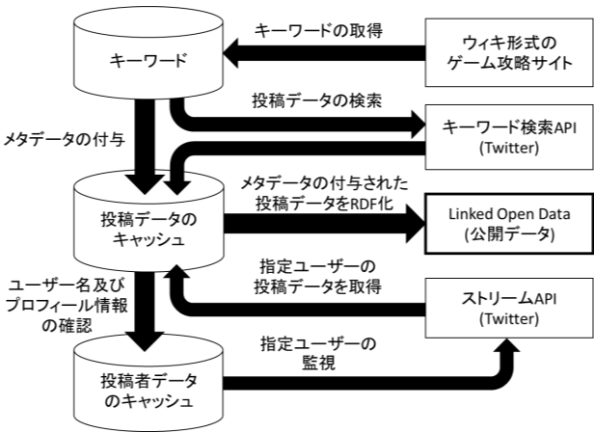

図 1 システムの概要図

5.2. Twitter の投稿データの収集

Twitter の投稿データの取得手段として、キーワード検索 API、及びストリーム API を利用する。ストリーム API で監視するユーザーについては、ウィキ形式の攻略サイトから収集したキーワードを用いてキーワード検索 API を利用した検索を行い、投稿データ、投稿者の名前、プロフィールのいずれかに、選択したオンラインゲームの名称を含む投稿データを収集し、その投稿者を監視ユーザーに指定する。

5.3. 監視ユーザーの順位付け

ストリーム API で監視するユーザー数が 5000 人を超える場合は以下の基準を用いて、ユーザーの順位付けを行い、上位 5000 ユーザーを監視対象とする。

A) キーワード検索 API によって取得された投稿データの数の多さ
B) フォロワーの数の多さ

5.4. LOD の作成

取得した投稿データに対して、ウィキ形式の攻略サイトから取得したキーワードのいずれかを含む場合は、該当するキーワードをメタデータとして投稿データに付与し、RDF 化する。該当するキーワードを含むかどうかに関して、文字列の類似度を算出して判断する。文字列の類似度に関してはメジャーな手法である Jaro-Winkler 距離を用いる。

6. 評価実験

監視ユーザーの順位付けにおいて、A)の基準を用いたとき、キーワード検索 API での取得結果に対して、79%の投稿データをリアルタイムに取得することができた。また、B)の基準の場合は 36%であった。したがって、A)を用いて提案システムを実装し、Web 上に公開した。

図 2 は 2015 年 4 月 1 日から 2015 年 7 月 31 日における、セッション数の推移である。破線部の 4 月 22 日付近において、セッション数が増加しており、以降定期的にセッション数が増加している。これは、選択したオンラインゲームのアップデート及び、メンテナンスが実施された日付と重なっている。また、図 3 は 2015 年 4 月 22 日のオンラインゲームのアップデート実施日におけるセッション数の推移であり、アップデートの終了時刻である 16 時 30 分以降から、セッション数が増加している。これらの結果から、Twitter より、選択したオンラインゲームに関して、速報性の高い投稿データを取得し、LOD として有用なデータを提供できていると考えられる。

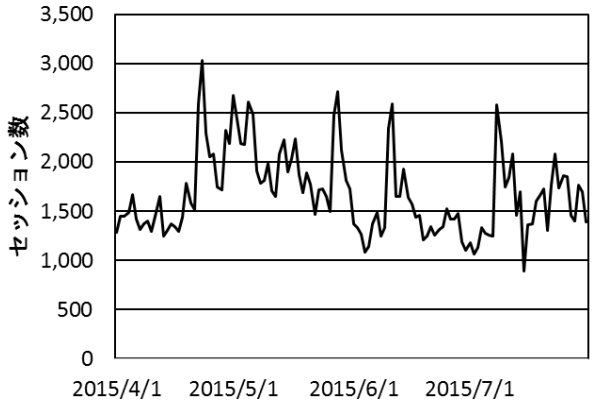

図 2 セッション数の推移

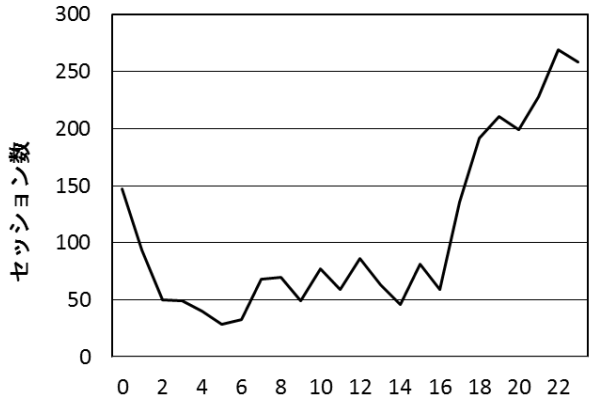

図 3 ゲームアップデート実施日における
セッション数の推移

7. まとめと今後の課題

本研究では Twitter の投稿データに着目し、自動的にオンラインゲームなどのデジタルコンテンツに関する投稿データを LOD として利用するためのシステムを提案し、実装及び評価をした。結果として、速報性の高い、有用な LOD を構築することができた。

今後の課題としては、投稿データの収集に関して、ヒューリスティックな判断を用いて自動収集しているため、より正確に分類するための手法を適用していく。

参考文献

[1] Twitter, https://twitter.com/
[2] 槇俊孝, 若原俊彦, “LOD を用いた論文のクラスタリングとメタデータの自動付与の試み”, 電子情報通信学会技術研究報告, vol.114, no.32, LOIS2014-8, pp.91-95, 2014 年 5 月.