

観光口コミの Web 情報との内容一致度による事実性判定

○鈴木 麻耶

福本 淳一

立命館大学大学院情報理工学研究科 立命館大学情報理工学部

1 はじめに

観光口コミは、主に旅行の計画を立てる際の情報収集として用いられる。しかし、口コミの中には本当かどうか分からないものや憶測で書かれているものも存在する。このような嘘の情報を信用すると、計画に支障が出る可能性がある。また、情報が信頼できるかどうかは受け取り側で判断する必要がある。そこで本研究では、口コミに書かれている内容と Web 情報を照合することで事実性の判定を行う。本研究によって、ユーザの旅行の充実度をあげることが出来ると考える。

2 関連研究

2.1 事実性確認

既存研究として、伊木らの複数の信頼性指標を用いて情報の信頼性を算出し、ユーザに提示する研究 [1] や Dong らの矛盾のあるファクトに着目し、web ページソースの信頼性を計算する研究 [2] などがある。これらは、対象の情報に対して、信頼できるかを計算している。しかし、事実であるかは判定されていない。

本研究では、他の媒体にも同じことが書かれているかを確かめることで情報に対して事実であるかの判断を行う。

2.2 文間類似度

同じことが書かれているかを判断する従来手法として文間類似度計算手法である、Word Mover's Distance (WMD)[3] がある。WMD は 2 つの文間の距離を定義できるを用いて、単語分散表現を特徴量に単語の意味を反映させた文書間の類似度を計算している。しかし問題点として、複数文にわたって書かれている場合は対応出来ない。文の一部が同じだと類似度が高くなるなどがある。さらに本研究では口コミ文に書かれている内容が web 情報に完全に含まれていれば事実確認としては十分であるため、クチコミ文に書かれていない部分まで類似度を計算する必要はない。

本研究では、クチコミ文に書かれている内容のみの文間類似度計算手法を提案する。

3 事実確認の判定対象

口コミの中には、投稿者にしか分からない情報など事実確認が不可能な口コミが存在する。そこで本研究では、事実が確認出来る文のみを判定対象として選択する。対象は以下の 2 つになる。対象となるのは、「観光地の情報」を説明している文である。例えば、「専用駐車場があります」や「食べ歩きをすることが出来ます」などの観光地についての基礎情報や見どころ、名産品、交通情報、お得情報が書かれている文が対象となる。ただし、行動や状況を説明しているものは対象外とする。

4 内容一致度による事実確認手法

4.1 提案手法概要

口コミに書かれている情報が事実であるかを、同じ内容が Web にも書かれているかで判定を行う。他にも同じ内容が書かれていることは、事実である可能性が高いと仮定している。提案する事実確認手法の概要を述べる。まず文に含まれる複数の内容を分割する。次に、一致する内容が最も含まれている Web ページを探し、内容一致度を算出する。最後に内容一致度から事実性があるか判定を行う。具体的な処理については以降の章で述べる。

4.2 内容の分割

口コミ文に含まれる複数の内容を分割する。内容の分割は、係り受け解析を用いて [主語, 述語], [修飾語, 被修飾語] となる単語の組の配列を作成することで行う。例えば「城内の階段は急なので注意が必要です。」という文の場合、['城内', '階段'], ['階段', '急'], ['注意', '必要'] の 3 つの単語の組が生成される。これにより文に含まれる「城内に階段がある」「階段が急である」「注意が必要である」の 3 つの内容が分割出来ている。

表 1: 実行結果

No.	口コミ文	一致度が最大となったスニペット	一致度
1.	国宝に指定されている 5ヶ所の城の一つの彦根城です	2020/12/18 ... 全国で現存する 12 天守の中の一つで、2015 年 (平成 27 年) 7. ... 彦根城・松本城・犬山城と並び、国宝に指定されている 5 城の一つとしても貴重なお城です ...	0.66
2.	天守閣内へは靴を脱いで上がります	脱衣場以外は浴場しかなく、石鹸やシャンプーなどのアメニティもないのですが、... 高さの天守閣のような形をした塔があり、誰でも靴を脱いで上がることができます。	1.0

4.3 Web 情報との内容一致度計算

以下の手順で内容一致度の計算を行う。

1. 口コミ文から検索クエリを生成し、Web 検索
2. 上位 20 件の Web ページのスニペット (Web ページの要約文) を取得
3. 口コミ文と同様に内容の分割を行い、単語の組を生成
4. 生成された口コミ文の単語の組とスニペットの単語の組を一つずつ比較し、単語の組が一致しているかを判定
5. 「一致した単語の組の数 ÷ 全ての単語の組の数」で内容一致度を計算

4.4 事実性の判定

4.3 章で算出した内容一致度から事実性の判定を行う。閾値を設定し、内容一致度が閾値よりも高ければ「事実性あり」、低ければ「事実性なし」と判定する。

5 内容一致度の評価実験

評価実験として、「事実性あり」と判定された口コミ文が、本当に内容が一致している Web ページが存在したかの評価を行った。また閾値は、「0.66」としている。実験データは、観光口コミサイトのじゃらん投稿された伊勢神宮と彦根城の口コミ文 144 件×スニペット 20 件の 2880 件を使用した。実験データの一部を表 2 に示す。

比較手法として WMD を使用する手法でも同様に評価を行った。また閾値は「0.4 以下」としている。

6 結果と考察

評価結果を表 3 に示す。提案手法の適合率は、78% となり比較手法より 10% 上回る結果となった。

表 2 の No.3 より、多少の表現のずれがあっても判定出来ていることが分かる。

表 2 の No.1 では、内容が一致しているにも関わらず一致度が低くなっている。原因としては内容の分割の際に内容のない単語の組が多く生成されていることが考えられる。

適合率に対して再現率は、30% となり比較手法より 30% 下回る結果となった。提案手法では、類似単語が出現しないと一致と判定されないためこのような結果になったと考える。

しかし WMD では、2.2 章でも述べたように口コミ文に書かれていない文まで類似度を計算しているため、スニペットに口コミの内容が完全に含まれていても類似度が低くなっているものが多く存在する。

今後の課題は、WMD のように類似単語が多く出なくとも一致度計算を行えるよう手法を改善することであると考える。

表 2: 評価結果

	適合率	再現率	F 値
提案手法	0.78	0.3	0.43
比較手法 (WMD)	0.68	0.6	0.63

参考文献

- [1] 伊木惇, 亀井清華, 藤田聡ほか. レビューを対象とした信頼性判断支援システムの提案. 情報処理学会論文誌, Vol. 55, No. 11, pp. 2461–2475, 2014.
- [2] X.Dong, E.Gabrilovich, K.Murphy, V.Dang, W.Horn, C.Lugaresi, S.Sun, and W.Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proc. VLDB Endow.*, Vol. 8, No. 9, p. 938–949, may 2015.
- [3] K.Matt, S.Yu, K.Nicholas, and W.Kilian. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.