3P-5

# Web 上のニュース記事を対象とした信頼度の提案

奈倉　良介†　　関　洋平†　　青野　雅樹†

豊橋技術科学大学†

## 1. はじめに

近年，インターネットの発達により大量の情報が Web 上にあふれている．ユーザーが得ようとする情報は，必ずしも正しい情報だけではない．悪意のある者による嘘や，ニュース記事においては，誤報が含まれることもある．そのような場合，正しくない情報を分類し，排除することができればユーザーは信頼に足る情報だけを得ることができる．本研究では， Web 上のニュース記事を対象として，記事を単位とした情報の確かさ，すなわち信頼できる度合を表す信頼度を判定する方法を提案する．

## 2. 関連研究

情報の確かさを表す研究として，Rubin ら[1] は，新聞記事の“著者の視点”，“著者の焦点”，“記事の時制”に着目し，それらが情報の確かさに与える影響を 4 段階の評価によって表した．評価の対象は，文を単位として行っている．

Google News[2]では，ニュース記事の配信元を特定の基準によりランク付けし，ランクの高い配信元から配信されたニュースはトップページから到達し易い所に表示することを提案している．これは，配信元の信頼度を設定することで，信頼度の高い配信元のニュース記事の閲覧できるようにしていると考えられる．本研究では，これらの提案とは異なり，記事自体の信頼度を求める手法を提案する．

## 3. 提案手法

ニュース記事の信頼度を得るために以下の 3 つの手順を用いた．

まず，イベントに対する記事がある配信元によって配信された時，異なる配信元でも同様な記事が数多く配信された場合には，その記事の情報は確かであると考え，その“**共通性の度合**”を計算した．

次に，記事中に“乗客 100 人”や“トラック 3 台”など数量表現が現れる時がある．その，数量表現が他の配信元の記事との間で矛盾していた場合，すなわち，数量属性が同じであるにもかかわらず，値が異なっていた場合，その記事の情報の確かさは落ちるものとした．逆に，数量表現が他の配信元の記事との間で同じであった場合には，情報の確かさは増すと考え，“**一貫性の度合**”を計算した．

最後に，2 つの基準により，他の配信元の記事と共通性が少なく，数量表現の一貫性にも欠けると判断された記事に対して，言葉の意味に着目し，“**信頼度**”を計算する．具体的には，「～かもしれない」や「～な可能性」といった，推測を表す言葉や「～という」などの伝聞を表す言葉が現れた際に，その推測や伝聞の度合を元に 4 段階のスコアをつける．以下では，それぞれの手続きについて詳しく説明する．

### 3.1. 他の配信元との記事内容の共通性

まず，信頼度判定の対象となる記事(対象記事)について，文を単位として分割する．タイトルも 1 つの文として扱う．次に，対象記事が配信された時間の 2 時間前から対象記事が配信された時間までを対象記事の“**同一時間帯**”と定義し，その時間帯内に配信されたすべての記事から，対象記事と同じ配信元の記事を除いた“**同一時間帯記事集合**”を作成する．そして，対象記事の各文に対して同一時間帯記事集合内の全ての記事の文と比較を行う．比較は，*LSI*（*Latent Semantic Indexing*）を利用して次元を 1/3 に削減した特徴量に基づくコサイン類似度を採用し，ある閾値を超えたものを類似している文と定義した．

別の配信元からの記事が類似した文を含むとき，その記事の配信元はその文について対象記事の配信元と同じ内容を扱っているとする．こうして，対象記事中のすべての文に対してそれぞれいくつの配信元が同じ内容を扱っているかの割合を計算し，すべての文についての平均を記事内容の配信元間での共通性の度合とした．これを表すと式(1)のようになる

$$f(n,t)=\frac{\sum_{k=1}^{n}(S_k/(t-1))}{n} \qquad (1)$$

ここで，$S_k$は文 $k$ における類似する配信元の数，$t$ は同一時間帯記の全配信元の数，$n$ は対象記事の文の数となっている．$t-1$ となっているのは対象記事の配信元は除かれるためである．

A Method of Rating the Trust for News Documents on the Web

†Ryosuke Nagura, Yohei Seki and Masaki Aono
†Toyohashi University of Technology

### 3.2. 他の配信元との数量表現の一貫性

まず，記事中より“乗客”などの数量属性と“100 人”などの数値の組を係り受け解析器 CaboCha[3]を利用して取り出す．次に，記事同士を比較する際に数量属性の一致したものの数値を比較し，値が等しいならば一貫性が認められるので正の値を，値が異なっていたならば矛盾が生じているので負の値を加える．そうして得られた値を，記事に対応する重みを与えるために対象記事の文の数で割った値を数量表現の“**一貫性の度合**”とした．これを表すと式(2)のようになる

$$g(i,c,n)=\frac{i-c}{n} \qquad (2)$$

ここで，$i$ は数量表現の値が等しい表現の数，$c$ は数量表現が矛盾していた表現の数である．

式(1)と(2)より求められた値を足したものが閾値より上回るものは，同一時間帯において同じ内容の記事が報道されており，かつ数値的な矛盾が大きく発生していないことを意味する．本研究では，これらの記事を，信頼できる記事と判定する．

### 3.3. 語の意味に基づいた信頼度

まず，式(1)と(2)の結果から閾値より下回った記事に対して，確実さの度合いを 4 段階で評価するための手がかり語句のリスト（表 1）[1]を作成した．次に，リストに対応した語句を対象記事の文中より抽出し，最低のスコアを持つ手がかり語句を利用して，すべての文にスコアをつけた．リストの中に“可能　性　高い”と“可能　性”といった重複があるが，係り受け解析により，形態素の数の多い係り受けを持つものを優先している．さらに，タイトルにおいて“～か”といった推測の形が現れた際に，タイトル以外の文の状況によりタイトルのスコアに 3 か 2 の値を振られている．最後に，対象記事のすべての文に割り振られたスコアから，ルール(図 1)に基づいて記事の信頼度のスコアを計算し，これを最終的な信頼度とした．

**表１．　不確実さの手掛りとなる語句の評価リスト**

| スコア | 語句 |
|---|---|
| 4 | [方針 表明] |
| 3 | [伝える] [と いう] [と 言う] [報じる]<br>[予定] [有力] [模様] [見通し] [方針]<br>[なる だ] [みる られる] [する 考え]<br>[情報 ある] [構え] [可能 性 強い]<br>[可能 性 高い] [意欲] [見込み] [展望 する]<br>[想定 する] |
| 2 | [希望] [可能 性] [予測 する] [予想 する]<br>[計画] [狙い] [かも する] |

[1] スコア 4 が“信頼できる”，スコア 1 が“信頼できない”ものとする．また，リストにはスコア 1 は存在しない．

1. スコア 4 を対象となる記事の仮の信頼度とする．
2. タイトルと重要文(タイトルと最も類似した 2 つの文)に付いたスコアの内，最も低いスコアを仮の信頼度として更新する．
3. タイトルと重要文以外の文での不確実さを考慮して，タイトルと重要文以外の文に 3 つ以上仮の信頼度より低いスコアがあるならば，そのスコアを記事の仮の信頼度として更新する．
4. 多くの不確実さをあらわす文がある場合を考慮して，記事の仮の信頼度と等しいスコアの文が，全体の文の数の半数より多く出現したならば，記事の仮の信頼度から 1 を引く．
5. また，記事の仮の信頼度より小さいスコアの文が，全体の文の数の半数より多く出現したならば，そのスコアから 1 を引いたものを記事の仮の信頼度として更新する．
6. この時点での仮の信頼度の値が記事の信頼度とする．

**図 1.　記事信頼度評価ルール**

## 4.　実験

実験データとして，2005/11/13～2005/12/13 までの 1 ヶ月間に 7 つの配信元によって配信されたニュースから，11 個のトピックに関連した記事を人手により分類し，記事集合を作成した．

実際に信頼度が低く判定された記事の一部を以下に表す．

タイトル(抜粋)：…で国内線旅客機が墜落か
本文(抜粋)：…乗客乗員およそ 114 人を乗せた国内線の旅客機が離陸直後に消息を絶ち，墜落したものとみられています…(中略)…搭乗していた可能性もあり…

この出来事の場合，他の配信元では乗員乗客 116 人と報道しているものが多く，矛盾した情報を含む記事が検出されていることがわかる．記事の信頼度は，タイトルの“～か”や本文中の“～とみられている”“可能性”などの推測を表す語の出現により，２と判定されている．

## 5.　おわりに

本研究では Web 上のニュース記事から記事の信頼度を判定する方法を提案した．

今後は，イベントが発生してからの時間がニュース記事に与える影響を考慮した信頼度モデルの作成や，数量的な矛盾のほかに，意味的な矛盾も取り扱うモデルの作成を検討している．

**参考文献**


1. V. L. Rubin, E. D. Liddy, and N. Kando, “Certainty Identification in Texts: Categorization Model and Manual Tagging Results,” In J.Shanahan, Y. Qu, J. Wiebe (eds), Computing Attitude and Affect in Text: Theories and Aplications, The Information Retrieval Series, Vol.20, Springer, Dordrecht, (2005), pp.61-74.
2. R. Ord, “Google News Patent Application - Full Text,” [online] 2005. [cited 2005-12-26], Available from: <http://www.webpronews.com/news/ebusinessnews/wpn-45-20050503GoogleNewsPatentApplicationFullText.html>.
3. 工藤拓，CaboCha/南瓜 [online]，奈良先端技術科学大学，2004. [cited 2005-12-26] Available from; <http://chasen.org/~taku/software/cabocha>.