属性型命題の真偽・確信度をウェブを用いて推定する

杉山　賛太[†1]　　松吉　俊[†2,†3]　　佐藤　理史[†3]

[†1] 名古屋大学工学部 電気電子情報工学科,
[†2] 京都大学大学院 情報学研究科,　[†3] 名古屋大学大学院 工学研究科

1. は じ め に

ウェブを用いた情報検索には次の 2 種類の場合があると考えられる。

(1)　わからないことを調べる場合
(2)　不確かな情報の真偽を確認する場合

例えば、(1) は、「木星の第 1 衛星が何であるか」を調べる場合である。多くの人は、検索エンジンに「“木星” “第 1 衛星”」と入力し、その検索結果から答えを探すだろう。検索結果のウェブページリストの中から適当に 1 つを選び、そこに記述されている情報から、「イオ」が木星の第 1 衛星であることを知るであろう。

(2) の例は、「木星の第 1 衛星がイオである」という命題の真偽を確認する場合である。これを行う際に、たまたま見つけた 1 つの情報に基づいて命題の真偽を決定することは危険である。なぜならば、ウェブには誤った情報も数多く存在するからである。それゆえ、人間は、ヒット数、スニペット、権威のあるウェブページなど様々な情報を複合的に用いて、命題の真偽を確認している。しかしながら、この作業を効果的に支援するツールは存在しない。

このような背景に基づき、我々は、ウェブを用いて命題の真偽と確信度を自動的に推定する手法を提案する。本研究では、その最初の題材として、ある対象の属性について述べる属性型命題を扱う。

2. 属性型命題の検証

2.1 属性型命題

属性型命題とは、次の形で表される、ある対象の属性について述べる命題のことである。

$<$ **対象** $>$ の $<$ **属性** $>$ は $<$ **属性値** $>$ である。

以下に、属性型命題の例を示す。

(A)　**ドイツ**の**首都**は**ベルリン**である。
(B)　**木星**の**第 1 衛星**は**エウロパ**である。

本研究では、属性型命題は、対象 (o)、属性 (p)、属性値 (v) の 3 つのキーワードの組 “o, p, v” として与えられるものとする。

2.2 真　　偽

属性型命題は、必ず、真か偽の値をとる。例えば、上記の命題 (A) は真であるが、命題 (B) は偽である (木星の第 1 衛星はイオである)。本研究では、属性型命題の真偽を推定する手法を提案する。

Magnini ら [1)] は、質問応答において、得られた解の妥当性を検証するために、質問文にその解を埋めこんで命題を作成し、その命題の「真」らしさを数値で算出している。ここでは、命題がどれほど「真」らしいかのみが問題にされる。なぜならば、彼らは、命題の集合に対して相対的な順位を付ければよいだけだからである。

それに対して、我々の問題設定は、1 つの命題の真偽を確認することである。この場合、真の命題に対してそれが真であるというだけでは不十分であり、偽の命題に対してそれが偽であるという必要がある。

それゆえ、本研究では、真の命題に対して “T (真である)” と推定し、偽の命題に対して “F (偽である)” と推定する手法を検討する。

2.3 確　信　度

本研究では、単に命題の真偽値を推定するだけではなく、推定した真偽値がどの程度信頼できるかを示す確信度も算出する。

我々が用いる確信度は、次の 3 段階である。

確信度 2　かなり信頼できる
確信度 1　信頼できる
確信度 0　信頼できない

続く 3 つの章において、真偽値と確信度を推定する、次の 3 種類の手法について説明する。

(1)　スニペットに基づく手法
(2)　ヒット数に基づく手法
(3)　上記 2 つを組み合わせた二次元法

3. スニペットに基づく手法 (スニペット法)

本章では、検索エンジンから得られる検索結果のテキスト部分 (スニペット) を利用して、属性型命題の真偽値と確信度を推定する手法について述べる。

3.1 真偽値の推定

対象 o と属性 p からなるクエリーを検索して得られたスニペット $S_{o,p}$ 中に、属性値 v が多く現れている場合、属性型命題 “o, p, v” は、真である可能性が高いと思われる。この考え方に基づいて、属性型命題の真偽値を推定する。

関数 $C(x, S_{o,p})$ と関数 $\mathrm{P}(o, p, v)$ を、以下のように定

義する。

$$C(x, S_{o,p})\ :\ S_{o,p}\text{中の } x \text{ の出現回数}$$

$$\mathrm{P}(o,p,v) = \frac{C(v, S_{o,p})}{\min(C(o, S_{o,p}), C(p, S_{o,p}))} \quad (1)$$

そして、真偽値を以下のルールによって推定する。

$$\begin{array}{llll} \text{if} & \mathrm{P}(o,p,v)\ \geq\ 0.2 & \Longrightarrow & \mathrm{T} \\ \text{otherwise} & & \Longrightarrow & \mathrm{F} \end{array} \quad (2)$$

ここで、境界点 0.2 は、分析用データに基づいて設定した値である。

3.2 確信度の算出

式 (1) は、$(C(o,S_{o,p}), C(p,S_{o,p}), C(v,S_{o,p})) = (5{,}4{,}1)$ に対しても、$(C(o,S_{o,p}), C(p,S_{o,p}), C(v,S_{o,p})) = (42{,}40{,}10)$ に対しても、同じ 0.25 という値を返すが、前者は、出現回数が少なすぎるので、信頼できない。

一方、ルール (2) は $(C(o,S_{o,p}), C(p,S_{o,p}), C(v,S_{o,p})) = (42{,}42{,}9)$ に対して T と推定するが、これに対する $\mathrm{P}(o,p,v)$ は、境界付近の 0.21 であり、あまり信頼できない。

このような問題を考慮して、真偽値の確信度を算出する。

本研究では、確信度の算出に、二項分布の信頼区間を利用する。実際の計算にあたっては、試行の数を $\min(C(o,S_{o,p}), C(p,S_{o,p}))$、平均を $\mathrm{P}(o,p,v)$ であると近似して、95%信頼区間を算出する [2)]。信頼区間の下限、上限を、それぞれ $\mathrm{P_L}(o,p,v)$、$\mathrm{P_U}(o,p,v)$ とする。

信頼区間は、式 (1) の分母の平方根におおよそ反比例するので、出現回数が少ない場合、広くなる。このような場合、確信度を低く設定すればよい。

次のルールに基づき、確信度を決定する。以下、$\mathrm{P}(o,p,v)$、$\mathrm{P_L}(o,p,v)$、$\mathrm{P_U}(o,p,v)$ を、それぞれ P、$\mathrm{P_L}$、$\mathrm{P_U}$ と略記する。

(1) P が確率値とみなせない場合
 if $\mathrm{P} > 1.0 \Longrightarrow$ 確信度 0

(2) 信頼区間が非常に広い場合
 if $0.5 \leq \mathrm{P_U} - \mathrm{P_L} \leq 1.0 \Longrightarrow$ 確信度 0

(3) 信頼区間が広い場合
 if $0.3 \leq \mathrm{P_U} - \mathrm{P_L} < 0.5$
 (a) 境界付近ではない
  if $\mathrm{P_L} \geq 0.15 \Longrightarrow$ 確信度 1
 (b) 境界付近
  if $\mathrm{P_L} < 0.15 \Longrightarrow$ 確信度 0

(4) 信頼区間が狭い場合
 if $0 \leq \mathrm{P_U} - \mathrm{P_L} < 0.3$
 (a) 境界付近ではない
  if $\mathrm{P_U} \leq 0.2 \Longrightarrow$ 確信度 2
  if $\mathrm{P_L} \geq 0.2 \Longrightarrow$ 確信度 2
 (b) 信頼区間の端が境界に少しかかる
  if $0.2 < \mathrm{P_U} < 0.25 \Longrightarrow$ 確信度 1
  if $\mathrm{P_U} \geq 0.25\ \&\ 0.15 < \mathrm{P_L} < 0.20$
   $\Longrightarrow$ 確信度 1
 (c) 境界付近 $\Longrightarrow$ 確信度 0

4. ヒット数に基づく手法 (ヒット数法)

本章では、検索エンジンから得られる、クエリーのヒット数を利用して、属性型命題の真偽値と確信度を推定する手法について述べる。

4.1 真偽値の推定

対象 o と属性 p を含むページの集合のうち、大部分のページが属性値 v を含む場合、属性型命題 "o,p,v" は、真である可能性が高いと思われる。この考え方に基づいて、属性型命題の真偽値を推定する。

関数 $\mathit{hits}(x)$ と関数 $\mathrm{Q}(o,p,v)$ を、以下のように定義する。

$$\mathit{hits}(x)\ :\ x \text{ のヒット数}$$

$$\mathrm{Q}(o,p,v) = \frac{\mathit{hits}(o\ \&\ p\ \&\ v)}{\mathit{hits}(o\ \&\ p)}$$

そして、真偽値を以下のルールによって推定する。

$$\begin{array}{llll} \text{if} & \mathrm{Q}(o,p,v)\ \geq\ 0.3 & \Longrightarrow & \mathrm{T} \\ \text{otherwise} & & \Longrightarrow & \mathrm{F} \end{array}$$

ここで、境界点 0.3 は、分析用データに基づいて設定した値である。

4.2 確信度の算出

3.2 節と同様の手法により、真偽値の確信度を算出する。ここでは、試行の数を $\mathit{hits}(o\ \&\ p)$、平均を $\mathrm{Q}(o,p,v)$ であると近似して、95%信頼区間を算出する。信頼区間の下限、上限を、それぞれ $\mathrm{Q_L}(o,p,v)$、$\mathrm{Q_U}(o,p,v)$ とする。

次のルールに基づき、確信度を決定する。以下、$\mathrm{Q}(o,p,v)$、$\mathrm{Q_L}(o,p,v)$、$\mathrm{Q_U}(o,p,v)$ を、それぞれ Q、$\mathrm{Q_L}$、$\mathrm{Q_U}$ と略記する。

(1) Q が確率値とみなせない場合
 if $\mathrm{Q} > 1.0 \Longrightarrow$ 確信度 0

(2) 信頼区間が非常に広い場合
 if $0.5 \leq \mathrm{Q_U} - \mathrm{Q_L} \leq 1.0 \Longrightarrow$ 確信度 0

(3) 信頼区間が広い場合
 if $0.2 \leq \mathrm{Q_U} - \mathrm{Q_L} < 0.5$
 (a) 境界付近ではない
  if $\mathrm{Q_L} \geq 0.25 \Longrightarrow$ 確信度 1
 (b) 境界付近
  if $\mathrm{Q_L} < 0.25 \Longrightarrow$ 確信度 0

(4) 信頼区間が狭い場合
 if $0 \leq \mathrm{Q_U} - \mathrm{Q_L} < 0.2$
 (a) 境界付近ではない
  if $\mathrm{Q_U} \leq 0.3 \Longrightarrow$ 確信度 2
  if $\mathrm{Q_L} \geq 0.3 \Longrightarrow$ 確信度 2
 (b) 信頼区間の端が境界に少しかかる
  if $0.3 < \mathrm{Q_U} < 0.35 \Longrightarrow$ 確信度 1
  if $\mathrm{Q_U} \geq 0.35\ \&\ 0.25 < \mathrm{Q_L} < 0.3$
   $\Longrightarrow$ 確信度 1
 (c) 境界付近 $\Longrightarrow$ 確信度 0

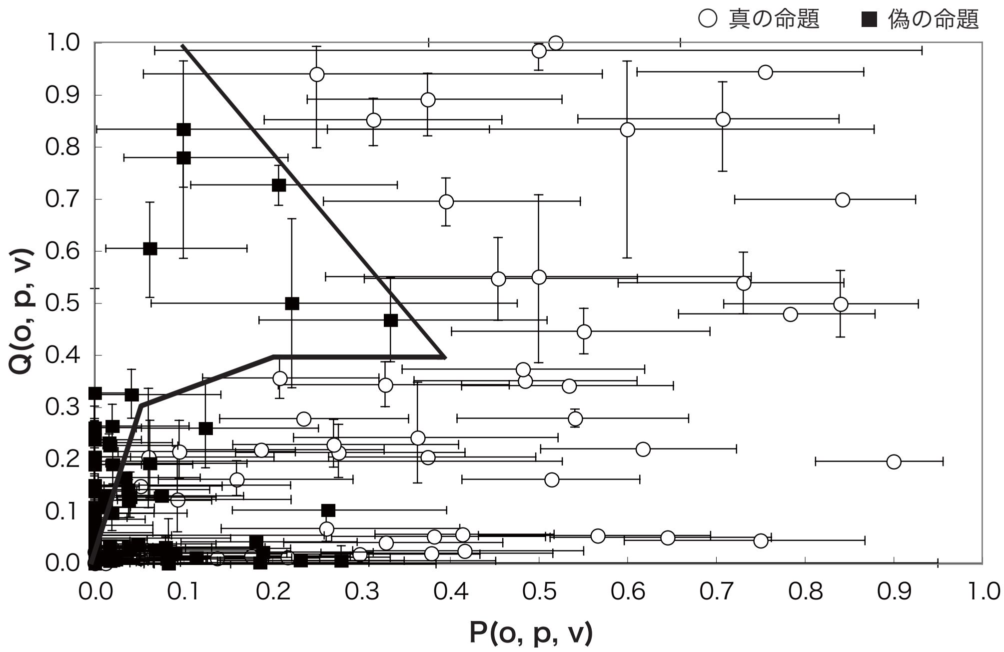

図 1 分析用データの属性型命題に対する散布図

5. 二 次 元 法

本章では、スニペット法とヒット数法を二次元的に組み合わせて、属性型命題の真偽値と確信度を推定する手法について述べる。

5.1 真偽値の推定

分析用データの属性型命題に対する (P,Q) の散布図を図 1 に示す。図 1 を見ると、真の命題は図の右に、偽の命題は図の左に偏って存在していることがわかる。したがって、これらの点をうまく分割することができる境界線を引くことができると考えられる。

この境界線として、曲線 $f(x,y)$ を次のように定義する。

$$f(x,y) = \begin{cases} 6x - y & (0 \leq y < 0.3) \\ 10x - 15y + 4 & (0.3 \leq y < 0.4) \\ y - 0.4 & (y = 0.4) \\ 2x + y - 1.2 & (0.4 < y \leq 1.0) \end{cases}$$

そして、真偽値を以下のルールによって推定する。

$$\begin{array}{lll} \text{if} \ \ f(\mathrm{P}(o,p,v), \mathrm{Q}(o,p,v)) \ \geq \ 0 & \Longrightarrow & \mathrm{T} \\ \text{otherwise} & \Longrightarrow & \mathrm{F} \end{array}$$

5.2 確信度の算出

3.2 節、4.2 節における手法を、二次元に拡張する。ここで、関数 $crossf(\mathrm{P},\mathrm{Q})$ を次のように定義する。

$(\mathrm{P_L},\mathrm{Q})$ と $(\mathrm{P_U},\mathrm{Q})$ を結ぶ線分、
もしくは $(\mathrm{P},\mathrm{Q_L})$ と $(\mathrm{P},\mathrm{Q_U})$ を結ぶ線分が、
曲線 $f(x,y)$ と交わる場合、$crossf(\mathrm{P},\mathrm{Q})$=1
そうでない場合、$crossf(\mathrm{P},\mathrm{Q})$=0

次のルールに基づき、確信度を決定する。

P が確率値とみなせない場合 $(\mathrm{P} > 1.0)$
$\Longrightarrow$ 確信度 0

Q が確率値とみなせない場合 $(\mathrm{Q} > 1.0)$
$\Longrightarrow$ 確信度 0

Q が大きい場合 $(0.3 \leq \mathrm{Q} \leq 1.0)$

（1） 境界付近ではない
if $crossf(\mathrm{P},\mathrm{Q}) = 0 \Longrightarrow$ 確信度 2

（2） 境界付近
if $crossf(\mathrm{P},\mathrm{Q}) = 1 \Longrightarrow$ 確信度 1

Q が小さい場合 $(\mathrm{Q} < 0.3)$

（1） P が大きい場合 $(\mathrm{P} \geq 0.3)$
- （a） 境界付近ではない
  if $\mathrm{P_L} \geq 0.3 \Longrightarrow$ 確信度 2
- （b） 境界付近
  if $\mathrm{P_L} < 0.3 \Longrightarrow$ 確信度 1

（2） P が小さい場合 $(\mathrm{P} < 0.3)$
- （a） “T” で、Q がそれほど小さくない場合
  if $f(\mathrm{P},\mathrm{Q}) \geq 0$ & $0.15 < \mathrm{Q} < 0.3$
  - （i） 信頼区間が広い場合
    if $f(\mathrm{P_L},\mathrm{Q}) < 0 \Longrightarrow$ 確信度 0
    if $\mathrm{Q_U} \geq 0.3 \Longrightarrow$ 確信度 0
    if $\mathrm{Q_L} \leq 0.15 \Longrightarrow$ 確信度 0
  - （ii） P と Q がかなり近い場合
    if $|\mathrm{P} - \mathrm{Q}| \leq 0.05$
    $\Longrightarrow$ 確信度 2
  - （iii） P と Q が近い場合
    if $0.05 < |\mathrm{P} - \mathrm{Q}| \leq 0.1$
    $\Longrightarrow$ 確信度 1
  - （iv） そうでない場合

$\Longrightarrow$ 確信度 0

(b) “F” で、Q がそれほど小さくない場合

if $f(\mathrm{P},\mathrm{Q}) < 0$ & $0.03 < \mathrm{Q} < 0.3$

(i) 信頼区間が非常に狭い場合

if $\mathrm{P_U} < 0.07 \Longrightarrow$ 確信度 1

(ii) そうでない場合

$\Longrightarrow$ 確信度 0

(c) そうでない場合 $\Longrightarrow$ 確信度 0

6. 実　　験

前章までに述べた 3 種類の手法を用いて、属性型命題の真偽と確信度を推定する実験を行った。

6.1 実験条件

本実験では、検索エンジンとして、Yahoo!☆を用いた。検索エンジンから取得するスニペットは、1 クエリーあたり上位の 50 件とした。

本研究では、属性型命題のリストを作成するために、カードゲーム版「クイズ$ミリオネア」☆☆ の問題 1960 問を用いた。問題の形式は 4 択問題であり、賞金別に A(￥10,000)〜O(￥10,000,000) のクラスからなる。

分析用・評価用データ作成の手順を以下に示す。

(1) それぞれのクラスを 2 つのグループに分割する

(2) 各々のグループの中から、人手で属性型である問題を抽出する

(3) 問題を 3 つのキーワードの組 “o, p, v” に変換する

それぞれの問題は 4 つの選択肢を持つので、問題 1 題から 4 つの属性型命題 (1 つの真の命題と 3 つの偽の命題) が作成される。

最終的に、分析用データとして、属性型命題 228 個、評価用データとして、属性型命題 308 個を得た。

6.2 結　　果

分析用データおよび評価用データに対して、提案する 3 つの手法で真偽を推定する実験を行った。手法ごとの正答率を表 1 および表 2 に示す。ここで、baseline は、入力された命題全てに対して、確信度 2 で F と推定する。

分析用、評価用ともに、二次元法が最も高い正答率を示した。これは、P と Q の 2 変数を用いることで、より正確な真偽値の推定ができることを示している。

二次元法の推定結果を詳しく表 3 および表 4 に示す。これらの表において、二次元法が F と推定したときの正答率は、分析用データに対して 97.9%(142/145)、評価用データに対して 97.1%(165/170) と、非常に高い。これより、二次元法は、偽の命題を偽であると正しく判断することができる長所があると言える。

7. おわりに

本稿では、検索エンジンから得られるスニペットとヒット数を利用して、属性型命題の真偽値と確信度を推定する手法について述べた。提案手法は、二項分布の信頼区間を利用することにより、推定された真偽値に対して 3 段階の確信度を算出する。スニペット法とヒット数法を組み合わせて用いる二次元法は、評価用データの属性型命題に対して、確信度 2 もしくは 1 と推定されたもので、94.4%の正答率を示した。

今後の課題は、次の 2 つである。1 つは、二次元法における境界線および確信度算出のルールを改良することである。もう 1 つは、属性型以外の命題の真偽と確信度を推定する手法を検討することである。

☆ http://www.yahoo.co.jp/

☆☆ フジテレビ製作 http://wwwz.fujitv.co.jp/quiz/index.html

表 1 真偽の正答率 (確信度 2)

手法	分析用 (228 個)	評価用 (308 個)
baseline	75.0%(171/228)	75.0%(231/308)
スニペット法	96.7%(174/180)	95.0%(210/221)
ヒット数法	82.5%(175/212)	79.1%(189/239)
二次元法	**100%**(63/ 63)	**95.8%**(68/ 71)

表 2 真偽の正答率 (確信度 2 もしくは 1)

手法	分析用 (228 個)	評価用 (308 個)
baseline	75.0%(171/228)	75.0%(231/308)
スニペット法	93.5%(187/200)	93.9%(231/246)
ヒット数法	81.4%(180/221)	78.5%(204/260)
二次元法	**98.8%**(169/171)	**94.4%**(184/195)

表 3 二次元法に基づく推定結果 (分析用データ)

推定	確信度	真	偽	計
T	2	**27**	0	83
	1	**12**	0	
	0	**15**	29	
F	0	1	**12**	145
	1	2	**94**	
	2	0	**36**	
計		57	171	228

表 4 二次元法に基づく推定結果 (評価用データ)

推定	確信度	真	偽	計
T	2	**31**	3	138
	1	**12**	3	
	0	**29**	60	
F	0	0	**24**	170
	1	5	**104**	
	2	0	**37**	
計		77	231	308

参 考 文 献

1) Magnini, B., Negri, M., Prevete, R. and Tanev, H.: Is It the Right Answer? Exploiting Web Redundancy for Answer Validation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 425–432 (2002).

2) 竹内啓, 藤野和建: 2 項分布とポアソン分布, 東京大学出版会 (1981).