多数決法と文章形態を考慮した検索システムの信頼性判断

中井川 祥[†]　　横井 健[†]
東京都立工業高等専門学校[†]

1　はじめに

今日、インターネットにはさまざまな人が発信した情報が混在しており、それらの情報には信頼度の差が存在する。そのため、あいまいな知識を検索する際、得られた情報の信頼度を判断することは困難である。従来、情報の信頼度を網羅的に判断する手法として多数決法[1]が提案されている。しかしながら、上記の情報群には知識の裏づけがないにも関わらず、多数決の結果に影響を及ぼす情報が含まれている可能性がある。そこで本研究では先の多数決法に加え、論理的な文章展開がなされている情報は信頼度が高いと考え、文章中の接続詞や語調などの文章形態にも着目した情報の信頼度判定システムの構築を目指す。

2　従来手法

検索エンジンはクエリマッチングやページランクなどの手法によって検索結果を決定している。山本らが提案した多数決法[1]は、まず検索によって得られた Web ページを検索した情報に対し肯定派・否定派に分類を行う。ページの分類は、「非～」「不～」などの接頭辞や、「～である」「～ないだろう」などの語尾を情報の肯定・否定を判断する語句として用いる。これらの言葉を「真偽判断スペル」と呼ぶ。

次に、分類結果を用いて式(1)に示すような情報の信頼度 TScore を算出する。

$$\mathrm{TScore}=\sum_{i=1}^{n}\sigma_i\times(P_i+1)\times\left(\frac{1}{2}\right)^{m_i} \quad \cdots(1)$$

P_i=ページランク，σ_i=+1(肯定),－1(否定)

m_i=真偽判断スペル,n=全検索結果数

σ_iは、n 件の検索結果の i 番目のページが、「肯定」なら$\sigma_i=1$、「否定」なら$\sigma_i=-1$、「どちらともいえない」なら$\sigma_i=0$ とする。また、Google が提供している PageRnak[2]をp_iとしている。

TScore の値は正負によって知識の肯定・否定を表し、値の大小によって知識の信頼度を表している。

3　提案手法

本研究では山本らの研究成果に加え、文書の形態と論理性も考慮して、知識の信頼度を判定する。

「～だ」「～である」などといった確言的表現は信頼度を上げ、「～だろう」「～はずだ」などといった推量的表現は信頼度を下げるとし、前者を確言スペル、後者を推測スペルとし分けて考え(1)式中の真偽判断スペルに関する項を次のように変更する。

$$(2/1)^{m_i}\rightarrow m_i^1+(1/2)^{m_i^2} \quad \cdots(2)$$

また、近年 Blog や掲示板などの普及によって、知識の裏づけのない確言的文章がよく用いられるようになった。そこで、論理的な文章展開の度合いを評価するために、接続詞や接続助詞の使用頻度に着目する。論理的な文章は信頼度が高いという仮定に基づき、接続詞、接続助詞を抽出し、その出現回数を信頼度判断に利用する。また、先の接続詞、接続助詞の中に逆説的表現が奇数回出現する場合は、真偽の肯定と否定が逆になるので、真偽判断スペルによって決定されたσ_iの正負を逆転する。

Trustworthiness Judgment of Information by the Majority Rule and its Context

† Sho NAKAIGAWA

† Takeru YOKOI

† Tokyo Metropolitan College of Technology

以上のような変更を加えた新たな式(3)を信TScore'とし、検証を行う。

$$TScore' = \sum_{i=1}^{n}\left\{\sigma_i \times (P_i + 1) \times \left(m_i^1 + \left(\frac{1}{2}\right)^{m_i^2}\right)\right\} \times \sqrt{1 + x_i} \quad \cdots (3)$$

x_i=接続表現の数,m_i^1=確言スペル数,m_i^2=推測スペル数

なおx_iは、一つのWebページ当たりの接続表現の出現回数である。

4　実験と考察

まず、真偽判断スペルを用いたページの分類システムが正確に動作するか検証した。検索キーワードとして「納豆　ダイエット」をYahoo!Japanに入力し、検索結果上位50件のページのスニペッドを人手と提案システムでそれぞれ「肯定」・「否定」・「どちらでもない」に分類し比較を行った。表1に結果を記す。

表1　真偽判別プログラムの結果

	肯定	否定	どちらでもない
システム	14	6	30
人手	26	4	20

表1よりほぼ正確に判別できたといえる。スニペッドは検索キーワードの付近をまとめたものなので、スニペッド内に肯定的表現がないページは「どちらでもない」に分類されている。

信頼度判断には、いくつかの解候補があるキーワードを利用する。今回は「邪馬台国」についてそれぞれ「九州」「畿内」「大和」「岩手県」で検証を行った。表2に結果を記す。

表2　邪馬台国のTScoreとTScore'

	九州	畿内	大和	岩手県
TScore	23.75	21.00	13.00	3.50
TScore'	34.9	28.85	15.27	4.50

表2より、邪馬台国問題について九州説の信頼度が一番高く、岩手県説は否定的な意見が信頼できるという結論が得られた。TScoreとTScore'を比較すると提案手法のTScore'は値が大きくなり、差が広がったことにより知識の信頼度の差がより明確になった。

次に、接続表現のページ当たりの平均出現回数(x_iの平均値)を表3に記す。

表3　接続表現の平均出現回数

	九州	畿内	大和	岩手県
x_iの平均値	0.14	0.08	0.14	0.04

表2と表3を見ると、TScoreからTScore'へ値を大きく伸ばした九州説に対し、畿内説の値はあまり伸びておらず、x_iの平均値も小さいことから、九州説を指示しているページの方が論理的である場合が多く、畿内説を指示しているページはそうでなかったと言える。接続表現の数が多いにも関わらず値が小さい大和節は、論理的な否定文章が多く、値を伸ばせなかった。

今回検証した邪馬台国問題は多くの議論がかわされており、論理的な文章が多くあったにも関わらず、表3のような結果になってしまった。本研究で用いた検索エンジンによって得られるスニペッドは、キーワードにマッチした前後の文章のみを提示システムであるので、接続表現の抽出には不十分であったとからと考えられる。

5　まとめ

本研究では、従来提案されている多数決法に基づいた、新たな信頼度判定手法を提案した。

また、今回の実験では確認できなかったが、TScoreとTsocre'との信頼度の順に違いが出てくることも考えられるので、今後キーワードを増やし検証を続ける予定である。

参考文献

[1]山本祐輔「ページ特性を考慮したWeb検索結果の集約とページ生成時間分析による知識の信頼性判断支援」

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, 'The PageRank Citation Ranking: Bringing Order to the Web', 1998,