4Q-1

統計的手法を用いた命題的知識の真偽判断システム

松岡雅也† 奥村紀之†
香川高等専門学科情報工学科†

1. はじめに

近年，検索エンジンの普及に伴い多くの情報を Web 上から取得する事が可能となり，現在 Web 検索は情報収集の一般的な手法になっている．しかし，日々増加していく Web 上の情報の中には，情報提供者の誤解や世間に出回ったデマの影響により，正しくない知識を発信している場合がある．その事が原因で，ユーザが正しくない情報を取得してしまう事が検索エンジンを用いた情報収集における問題となっている．そのために Web 情報の信憑性を統計的に分析することで，自身が取得している知識の真偽を測定することが必要であると考えられる．本稿では Web 上の情報を分析する際に，どの様な条件を元に信憑性を判断したか、及び分析による傾向の考察について記述する．

2. 研究目的

本研究の目的は，ユーザ自身の持つ知識が，本当に正しい情報であるかを，機械的に判断させることである．その際に命題がインターネット内で，どれほど確からしい情報であるかを解析するシステムを考案する事が必要となる．本研究では検索エンジンを用いて Web 情報を取得し，命題がどれだけ正しいと考えられるか検討することで，命題の Web 上における信憑性について考察する．

3. 検索エンジン Google

分析を行うに当たって，Web 情報を取得するために，検索エンジン Google を使用する．命題一文(例:「織田信長は自殺した」)をキーワードとして検索し，表示されたサイトのうち上位 50 件の文章を取得する．但しこの 50 件は試験的に定めたもので，今後の研究でパラメータを検討する.そして命題の否定文である「織田信長は自殺してない」も分析対象とし，同様に文章を取得する．本研究では，例文「織田信長は自殺した」と「信長の奇行を諫めるために彼は自殺した」では前者の方が正しい事が考えられる.この事に着目して，重要な単語同士(例:信長と自殺)が近ければ近いほど信憑性が高いと仮定して分析を行い，命題の真偽を判定出来るか検討する．

4. 分析手法の検討

本稿で行う分析では，命題における重要な単語を二つのみ取得する事とする.検索によって取得した文章から，重要なキーワードが二単語とも存在する文章のみ抽出し分析対象とする．そして文章ごとの単語間距離を取得し考察を行う．また文章群それぞれに構文解析を行うことで文章の内容を分析する.分析結果に応じたスコア付けを行うことで，出力された結果全体から命題が Web 上でどの程度の信憑性を持っているかを数値的に見る事が可能となる．

5. Web 情報性分析

本節では，命題に対する Web 情報の信憑性分析の手法としてスコアリング方式を提案する．

5.1. スコアリング

スコアリング方式は，情報の様々な事象を定量化し，スコアを設けることで分析結果を数値的に見る手法である，スコアリング方式を用いた手法による研究は広く行われている[2].
本稿で行うスコアリングでは，下記の様な文章の特徴と信憑性の関係を定量として起用する．

(1)単語間の距離が短いほど命題について記されている可能性があり，逆に長い程命題とは違った内容を記している可能性がある
(2)係り受け関係が正しければその文章は命題について関わっている可能性があり，正しくなければその文章は命題について関わっていない可能性がある．係り受け解析器 Cabocha にて解析を行い，係り受け関係になっているか分析する．
(3)否定文ならばその命題は正しくない情報である可能性がある．
(4)疑問文ならばその命題は正しいと言えない情報である可能性がある

「A System of Judge Truth or Falsity of Propositional Knowledge Using Statistical Method」
†「Masaya MATSUOKA」
‡「Noriyuki OKUMURA」
Kagawa National College of Technology, Department of Information Engineering

上記の条件によって，スコアに変化を設ける．本稿では試験的に初期値を 50 と定め，(1)では単語間距離が 5 以下であれば 1.2 倍，10 以下ならば 1.1 倍，50 以上離れていれば 0.9 倍し，(2)を満たしていれば 1.2 倍，(3)を満たしていれば 0.6 倍，(4)を満たしていれば 0.8 倍とスコアに変化を設ける．そして出力された結果を検討することで命題の Web 上における真偽判断の考察を行う．但しスコア付けにおいて試験的に定数倍を施してあるが，最終的に学習によって係数を確率させていくため，本実験の定められた数は検証のための設定値である．

5.2. 検討事項

分析結果から，スコアと単語間距離の関係をグラフ化すると，距離が 9 付近でスコアが減少しているのがわかる(図１)．しかし仮定より，文章間の距離が短いほど，高いスコアを出力する筈であるが，仮定通りのスコアが出力されていない場合がある．本項では仮定通りのスコアにならない文章を考察し，原因について考察する．

(1)重要なキーワードが複数ある場合

重要な単語がどちらも存在する一文の中には，「“信長”は“信長”の部下によって“自殺”した」の様にその単語が複数ある場合が考えられる，その際に距離の取得に間違いが生じたため，スコアに影響を受けた．二つの単語の位置関係を正確に取得する手法を検討する必要がある．

(2)情報を補足する文章がある場合

比較的単語間距離が長い場合でも『信長は「必ず 死骸を敵に渡すことなかれ」と遺言し、自殺したそうです』の様に理由等，情報を補足する文章の場合，スコアは高く表示されていた．単語間距離と信憑性の関係において，今後着目して考察していく事が必要である．

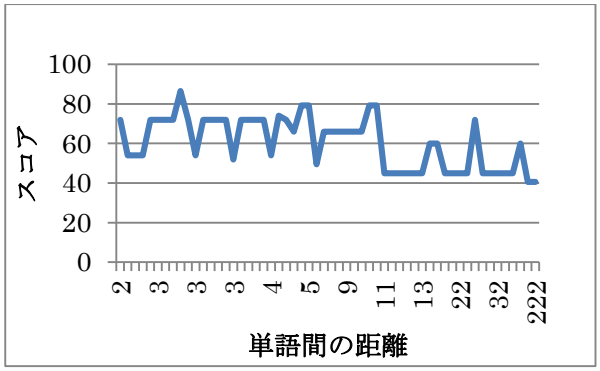

図１:単語間距離とスコアの関係
例:「織田信長は自殺した」

6. フレーズ検索と信憑性分析

命題の信憑性判断材料の一つとして，フレーズ検索数を起用した．フレーズ検索とは検索エンジンの持つ検索方法の 1 つで、単語でなく語句で検索する方法である．本研究では検索数を取得することでスコアとの関係を検討していく．

7. 考察

フレーズ検索と信憑性分析によるスコアリングによって出力された数値化された命題の信憑性と，その命題の否定文を表す命題の信憑性を比較した(表２)．その結果スコア平均，単語間距離の平均には大きな差は見られなかったが．フレーズ検索数においては大きな差が出力された．これは検索エンジンが通常の検索において，入力したキーワードを分解して単語として検索するので，検索対象が殆ど同じであるからである．

表２：命題の Web 上における信憑性分析結果

命題「キジは国鳥である」			
スコア平均	距離平均	検索ヒット数	フレーズ検索
64.36839	7.754839	66100	1550
命題「キジは国鳥ではない」			
スコア平均	距離平均	検索ヒット数	フレーズ検索
63.61678	6.811189	52100	1

8. まとめ

本稿では命題の真偽を，Web 情報を基に分析する手法として，スコアリング方式を考察し，単語間距離とスコアの関係や検索ヒット数と信憑性の関連性について考察した．今後は命題を分析し，パラメータを機械的に学習させる事が目標となる．また，本研究では命題に対して重要な単語を二つ定めたが，重要な単語が三つの場合(例：「“信長”は“本能寺の変”で“自殺“した」)など、命題の種類によってどの様な分析方法を行うかを考察する事が必要であると考えられる．

謝辞

本研究の一部は研究費(23720222)の助成を受けたものである。

参考文献

[1]森　恒: 情報の信頼性分析に向けた評価データおよびプロトタイプシステム WISDOM，社団法人，情報処理学会　研究報告，2007-NL-180(18).
[2]佐藤永欣: 協調サーチエンジンにおける tf・idf 法に基づく分散スコアリング，情報処理学会シンポジウム論文集，Vol.1999，No.18，Page91-96