関連記事の判定に関する検討

奥　雅博、鷲崎誠司、田中智博
NTT情報通信研究所
{oku, suzaki}@isl.ntt.jp, tomohiro@nttnly.isl.ntt.jp

1．はじめに

インターネットの普及によって、誰もが容易に大量の情報にアクセスできるようになってきている。しかし、その大量の情報の中から必要な情報を自由に選び出すことができなければ、その大量の情報は十分に活用されているとは言えない[1]。このような問題を解決するため、情報検索の分野では多くの研究が行われている。これらの研究では主に検索速度や検索精度の向上に主眼が置かれている。特に検索精度向上のための方策として、単語ベクトルを用いた類似度計算に基づく部分一致検索や、利用者が選択した適合情報を検索質問に反映させる関連性フィードバックなどが研究されている。これに対して検索結果をどのような優先順位で利用者に提示するか、あるいは複数の検索結果をどのように統合して利用者に提供するかといった、情報の提供に関する研究は重要性が指摘されているにもかかわらずあまり行われていないのが現状である[2]。

我々は、情報の提供にあたっては、得られた検索結果を関連のあるグループごとにまとめて利用者に提供することが望ましいと考え、「関連がある」とはどういうことなのかについて新聞記事を対象に検討を進めている。新聞記事には「社会面に関連記事」といった「関連がある」記事の存在を明示したものが存在する（1面トップ記事に多い）。人間は「社会面に関連記事」という情報だけで社会面に存在する多数の記事の中から関連のある記事を容易に同定することができる。本稿では、その同定に用いられている情報について検討し、それをもとに行った評価実験の結果を報告する。

2．関連記事の判定

2．1　見出し間の関連

ある記事とその関連記事とを比較すると、それらの見出し間にも何らかの関係があると推定される。そこで、見出し間の関係を調べるために94年4～6月の3カ月間の新聞記事を対象に以下の2つの調査を行った；

（調査1）同一の記事に対する複数の新聞社間の見出しの比較。
朝日新聞の各日の1面トップ記事と同一の事件を扱っている毎日新聞および読売新聞の記事との見出しの比較を行った。

（調査2）ある記事の見出しとその関連記事の見出しとの比較。
朝日新聞を対象に、各日の1面トップ記事とその関連記事（「x面に関連記事」と明示されているもの）との見出しの比較を行った。

なお、両調査とも比較結果を以下の5つに分類した；

（1）見出し　…　見出し全体が一致／同義／類似していることによって判断される場合。すなわち、見出しそのものが記事間で非常に似ている場合。

（2）見出し語句　…　見出し全体では異なっているが、見出しに含まれる語句の単位で一致／同義／類似していることによって判断される場合。すなわち、見出し全体では異なっているが、見出しに用いられている語句が似ている場合。

（3）組み合わせ　…　複数の見出しや見出しの語句を組み合わせた単位で一致／同義／類似していることによって判断される場合。

（4）記事記述　…　見出しでは一致が見られず、記事本文の記述によって判断される場合。すなわち、見出しからだけでは同一事件を扱っている（調査1）／関連記事である（調査2）ことが判断できない場合。

（5）背景知識　…　見出し中の語句が辞書的な意味では同義や類似にはならないが、背景知識によって判断される場合。

図１に調査１と調査２の結果を示す。横軸は上記の比較結果の５分類、縦軸は調査期間３カ月間の合計件数である。図１より以下のようなことがわかる；

（調査１）同一事件を扱った記事では新聞社が異なっていても見出し全体が類似している。
ある事件を報道するにあたっては見出しそのものが類似する。これは見出しによって伝えるべき内容（事件の重要なポイント）は新聞社を問わず同一であり、事件を客観的に捉えると見出しそのものが類似するためであろう。

（調査２）１面トップ記事とその関連記事とでは見出し全体というよりもその一部が類似している。
調査１の結果に対して、１面トップ記事とその関連記事との間では、関連記事が１面トップ記事のある側面を捉えたものであるため、見出し全体が異なるのは当然のことであろう。また、関連記事であることが読者にわかるように、見出しのどこかに１面トップ記事との関連を示す語句を用いているということが類推される。

以上のことから、関連記事を見つけるには見出しの一部が類似しているものを見いだせばよいということが言える。

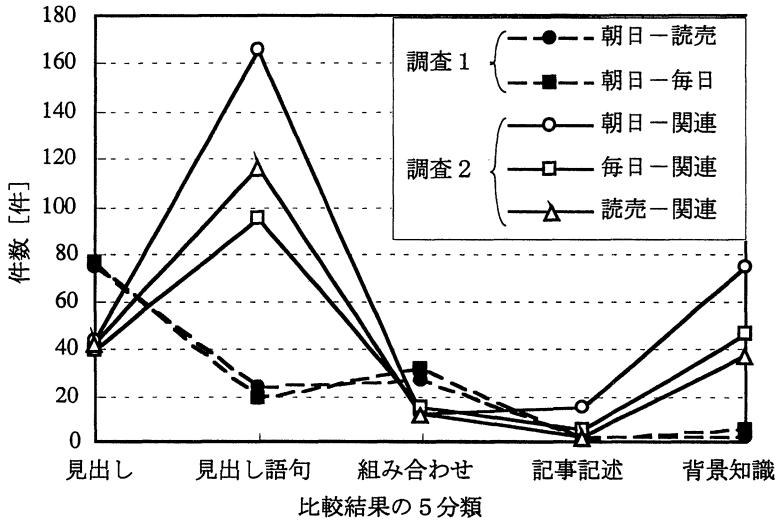

図１：　調査１、２の結果

２．２　見出し間の類似判定の単位

関連記事であるか否かを判定する際に見出し間の類似判定を行う必要があるが、見出しの類似を見るための単位として最適なのは何であろうか？情報検索や情報分類の分野、特に空白で単語単位に区切られている英語をはじめとする欧米語族を対象とする場合には、内容が類似していることを示す基本単位として単語を用いることが多い。しかし、日本語のように単語を区切って書く習慣のない言語を対象とすると、単語を切り出すために高精度な形態素解析システムと大量の語彙的知識が必要となる。単語によって内容の類似が計れるのならば、自立語を構成することが多い文字種（漢字、カタカナ）によってもある程度内容の類似が計れると考えられる。また、漢字文字、カタカナ文字を単位とすることで、単語切り（形態素解析）も不要となるし、大量の語彙的知識も不必要となる。

そこで本稿では、文献［３］と同様な考え方に基づき、自立語を構成することが多い文字種（漢字、カタカナ）を意味の基本単位としてとらえ、これを基本にして見出しの類似を定量的に扱うことを試みる。

3．評価実験

2．2節の考え方に基づいて、2つの新聞記事の見出し間で同一の漢字およびカタカナが文字単位でどの程度現れるかをカウントすることによって関連記事の判定を行う評価実験を実施した。

3．1　評価実験の手順

94年4～6月の3カ月間の新聞記事を対象に以下の手順で評価実験を行った；

（1）「x面に関連記事」という記述がある1面トップ記事を選択する。

（2）1面トップ記事の見出しに含まれる漢字、カタカナを抽出する。

（3）x面に掲載されている関連記事を同定し（この同定は人手によって行う）、その見出しに含まれる漢字、カタカナのうち、（2）で抽出した文字と同一の文字の数をカウントする。

（4）x面に掲載されている関連記事以外の記事（以下、非関連記事）の見出しに含まれる漢字、カタカナのうち、（2）で抽出した文字と同一の文字の数をカウントする。

（5）以上の処理を実験対象期間の各日について行う。

3．2　結果と考察

評価実験の結果を図2に示す。図2から、1面トップ記事の見出しに含まれる文字と同一の文字が、非関連記事よりも関連記事の方に多く出現することがわかる。このことは、1面トップ記事の見出しと別記事の見出しとを文字レベルで比較することによって、その記事が1面トップ記事の関連記事であるか否かを判定できることを示している。

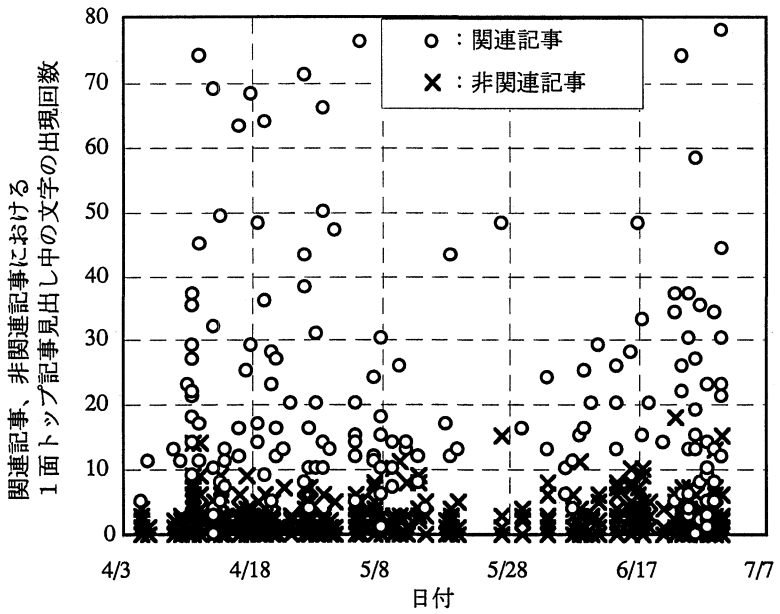

図2：　評価実験の結果

次に、再現率と適合率を用いて関連記事判定の精度を評価する。再現率、適合率は以下のように定義される；

○再現率＝（実際の関連記事のうち、実験によって関連記事であると判定された件数）／（実際の関連記事の件数）。

○適合率＝（実際の関連記事のうち、実験によって関連記事であると判定された件数）／（実験によって関連記事であると判定された件数）。

実験による関連記事の判定は、「１面トップ記事の見出しに含まれる文字と同一の文字が、関連記事の見出しに閾値の数以上、存在した場合」とした。

図３に再現率、適合率の閾値による変化を示す。図３より、閾値を５とした場合（新聞記事の見出し間で５個以上の漢字、カタカナの文字が同一である場合）を関連記事とすると、再現率＝約９０％、適合率＝約６５％であり、閾値を１０とすると、再現率＝約７５％、適合率＝約９０％である。

今回の結果は、３カ月間の１面トップ記事とその関連記事から得られた結果であり、無作為に取り出した２つの新聞記事を文字レベルで比較して関連記事であるか否かを上記の精度で答えられるとは限らない。しかし、文字レベルの比較だけでもかなりの精度で関連記事を判定できる可能性があることがわかった。

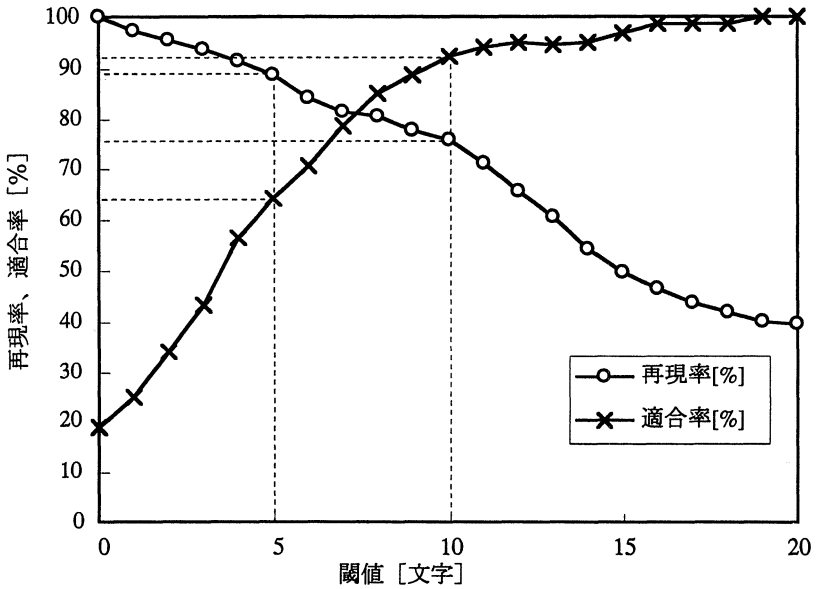

図３：　再現率、適合率の閾値による変化

４．おわりに

本稿では、新聞記事における関連記事を判定する手法として、文字を単位とした手法について述べた。実際の新聞記事を調査した結果、ある記事と別の記事とが関連記事であるか否かは両者の見出しの一部がどの程度類似しているかに依存することが多く、これをもとに関連記事の判定が可能であることがわかった。さらに本稿では、見出しの一部の類似を判定するのに漢字文字、カタカナ文字を単位として用いた評価実験を行い、新聞記事の見出し間で５個以上の漢字、カタカナの文字が同一である場合を関連記事としたとき、再現率＝約９０％、適合率＝約６５％の精度で判定できることを示した。

【参考文献】

［１］住田、三池、“知的情報検索の動向”、人工知能学会誌、Vol.11、No.1、pp.10-16 (1996).

［２］城風、羽生田、木下、“統計的シソーラスを用いた分散型ネットワークニュース検索システム”、信学技報AI95-24 (1995).

［３］渡辺、竹内、村田、長尾、“χ^2法を用いた重要漢字の自動抽出と文献の自動分類”、信学技報NLC94-25 (1994).