2G-1

Webからの情報抽出・検索システムにおける文書検索へのフィードバック適用の効果

濱口　佳孝[†]　池野　篤司[†]　山本　英子[‡]　井佐原　均[‡]

沖電気工業株式会社[†]　　独立行政法人情報通信研究機構[‡]

あらまし

本稿で報告するシステムは、Webを情報ソースとして文書検索を行い、検索された文書から抽出された人名などのＮＥをランキングして提示する。このシステムの文書検索部に pseudo feedback の手法を適用した場合の、抽出・検索されるＮＥの精度への影響の評価を行った。この時、関連語を得るための一次文書検索に Web 文書を一定単語長で分割したものを用いた場合、その分割長により精度の変化が見られた。東京大学殿のホームページを対象とした人名検索で評価した結果、Web 文書を 100～200 語程度で分割した場合に性能向上が認められた。

1. はじめに

我々は、検索したいＮＥ（固有表現）の種類がユーザによって指定されることを前提として、入力文にマッチした種類のＮＥ（人名、技術名など）を Web ページのような非定型文書から検索するシステムの開発をすすめている。

このシステムはまず文書検索を行い、検索された文書の上位のものから指定のＮＥを抽出する。そして出現頻度及び、入力文中の単語との文書中での出現位置の関係に基づきスコアリングし、情報検索結果として提示する。

そのため、ＮＥの検索の精度向上には、文書検索の精度向上が必要となる。文書検索の精度向上の手法としては、１度文書検索を行い、上位の文書検索結果により検索条件を修正する pseudo feedback が効果があることが報告されている[1)]。

また、新聞記事についてパッセージに分割した索引を用いて feedback を行うと精度が上がるという報告がある[2)]。Web ページは長文であったり、複数の話題が１ページに含まれることも多く、この文書を分割する試みは効果が期待できる。

今回、これらの手法を文書検索に適用した場合の、Web ページからのＮＥ抽出・検索結果の精度への影響の評価を行った。

2. 実験概要

システム全体は図１に示したような構成を取る。文書 DB１には一定数の自立語を含むように分割した Web 文書を収めている。これを対象とした文書検索結果に特徴的な単語を関連語として抽出し、検索条件に追加する。文書 DB２には文書全体を収め、これを修正された検索条件により文書検索を行うことで、Web のページ単位での検索結果を得ている。

ここで、各文書検索での評価値には、Web からの情報抽出に有効である OKAPI[3)]に単語の文書中での繰り返しやすさを考慮した正規化を導入した手法を用いている[4)]。また、この文書検索の実装にはエンジンとして、情報処理振興事業協会(IPA)殿が実施した独創的情報技術育成事業の研究成果である、汎用連想検索エンジン GETA を使用させていただいている。

関連語抽出での重み付けには、一次文書検索結果中での出現数と、文書 DB 中での idf を用い

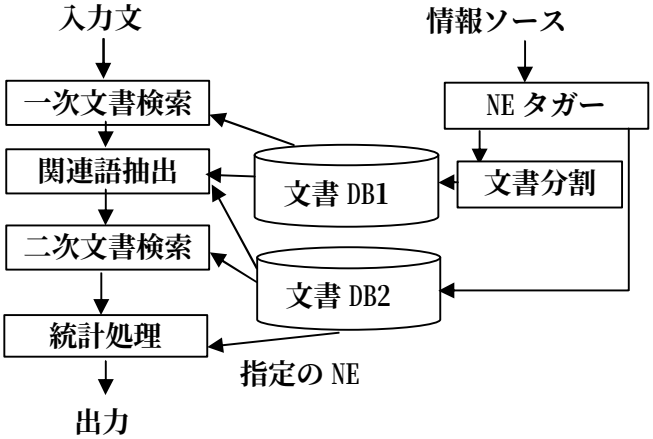

図１：実験システム

†OKI Electric Industry Co. Ltd.
‡National Institute of Information and Communications Technology

ている。一次文書検索と二次文書検索で対象 DB が異なるため、idf の基となる文書 DB による結果の相違についての比較実験も行った。

評価の入力文及び正解のデータは、東京大学殿の産学連携提案テーマデータベース [5]の情報・通信分野の、Web ページ収集時直近の 205 データから作成した。入力文としては、ここから手で選んだ 10～40 語の主要文、人名検索の正解としてはそのテーマの担当の方の名前を用いた。情報ソースは東京大学殿の Web ページ(http://www.u-tokyo.ac.jp/)から上記の産学連携提案テーマデータベースを除いた約 36 万ページを用いた。

これらによる人名の検索により、正解が５位以内に入る率での評価を行った。

3. 評価結果

feedback により検索条件に加えられる単語数を 10 語とし、一次文書検索における検索文書数と、一次文書検索に用いるデータベースに投入する時に Web ページを分割するのに用いた自立語数を変数として、精度評価を行った結果のグラフを図２に示す。

一次検索文書数が 10 ページの場合であれば、一次文書検索対象を 100～200 語程度の自立語を含む程度に分割された文書とした場合に、従来 5 位までに正解が入らなかったうち約１割を救済するに相当する精度向上が見られた。

一次文書検索対象を分割されない Web ページとした場合の pseudo feedback では、人名検索の５位正解率は 43.2～46.8%となり、一次検索文書数がいずれの場合でも、feedback を行わない場合よりも精度はかえって落ちている。

一次検索文書数に関しては、10 文書、5 文書では大きな差は見られないが、30 文書では効果が小さくなった。

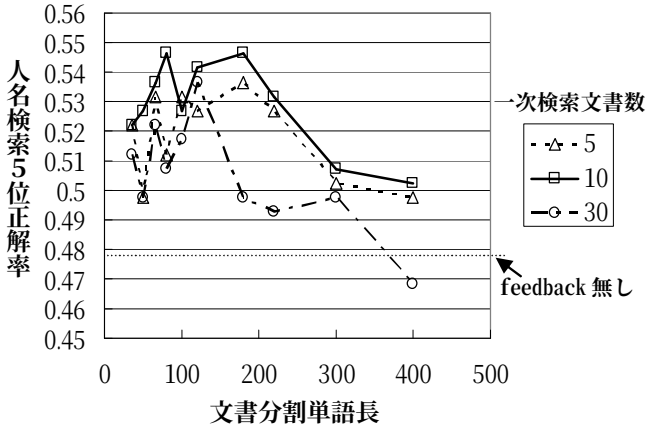

図２：文書分割と NE 検索精度

feedback で検索条件に加える単語数については、10～20 語で効果が見られ、それより少ない場合は効果が小さくなり、2 語ではほとんど効果が得られなかった。

なお、関連語抽出における重み付けで、idf としていずれの文書 DB の情報を用いるかについては、有意な差は見られなかった。

4. 考察

以上述べたように、Web 文書からのＮＥ抽出・検索における全文検索に pseudo feedback を適用にあたり、文書を分割することが情報検索の精度向上に効果があった。

分割単語長が 100 語以下では分割長などによる精度が大きく変動するが、これは関連語抽出の対象となる単語数が少なくなり統計的な母集団が小さくなるためと予想している。統計的手法に必要な語数を保つための、検索文書数と分割語数などの決定の目安となると考えている。

分割語数、一次検索文書数、feedback に使う単語数など各変数は今回の実験では経験的に決定して固定している。これらの傾向と決定手法や、適応的な手法についてさらなる分析が必要と感じている。また、複数の記事を含む Web ページの記事分割の手法を適用した場合などと、比較評価を行う必要がある。

参考文献

1) 酒井哲也, Gareth J.F. Jones, 梶浦正浩, 住田一男: 確率モデルに基づく日本語情報フィルタリングにおけるフィードバックによる検索条件展開および検索精度評価, 情報処理学会論文誌 Vol.40 No.5 pp.2429-2438 (1999)

2) 佐藤光弘, 伊藤快, 野口直彦: 松下電器産業における IR タスクへの取り組み, IREX ワークショップ予稿集 pp.69-74 (1999)

3) Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M. and Gatford, M. Okapi at TREC-2. *Proc. Text Retrieval Conference (TREC-2).*(1993)

4) 濱口佳孝, 池野篤司, 井佐原均: Web からの情報抽出・検索システムにおける全文検索, 情報処理学会研究報告 Vol.2004, No.93, pp.9-14 (2004)

5) 東京大学産学連携提案データベース, http://www-db.ccr.u-tokyo.ac.jp/