クラスタリングに基づくWeb文書の要約法に関する研究

川原尊徳[†]，岡崎直観[‡]，石塚満[‡]

[†]東京大学 工学部電子情報工学科
[‡]東京大学大学院 情報理工学系研究科 電子情報学専攻

1 はじめに

Web サイトを情報源として活用するための一つのアプローチとして、要約が利用されている。Google [1]では、検索結果に各 Web サイトの要約を含めている。GoogleNews [2]や NewsInEssence [1] では、Web 上のニュースサイトから記事を自動的に取得し、その要約を提供している。RSS (RDF Site Summary) を利用し、自らサイトの概要をメタデータとして記述する動きも急速に広まっている。

サーチエンジンを活用した Web サイトの要約手法としては、検索したサイトの内容を先頭から表示するもの [3]や、クエリーの出現箇所周辺の内容を表示するもの [4]など、単純なヒューリスティックを用いるものが代表的である。その他の手法としては、あるサイトに張られているハイパーリンクのアンカーテキストを検索エンジンを用いて収集し、そのサイトの要約を作成する手法が提案されている [2]。これらの要約は、検索結果の適合性の判断に用いる indicative（指示的）な要約と呼ばれる。一方、複数の Web サイトの内容をまとめた informative（報知的）な要約を作成するための研究は、非常に少ないのが現状である。

そこで、ユーザが検索エンジンを使って収集した複数の Web サイトから、ユーザにとって有用である箇所を推定しながら、その内容を網羅的にまとめる手法を提案する。複数の Web サイトの informative な要約を作成し、ユーザの情報欲求を要約で直接的に満たすことを目指す。

2 提案手法

検索エンジンで収集した複数 Web サイトの要約手法共通の課題として、多様な文書形式への対応や要約システムの応答速度が挙げられる。しかし、informative な複数 Web サイト要約を目指には、要約の網羅性とユーザの情報要求の把握を重視する必要がある。要約の網羅性は、複数文書自動要約に共通の課題であるが、収集した Web サイトの中に含まれる話題を認識し、出来るだけ多くの話題を要約に含めるのが望ましい。

また、ユーザは自分の情報要求をクエリという形で表現するが、自分の欲しい情報を得るための適切なクエリを

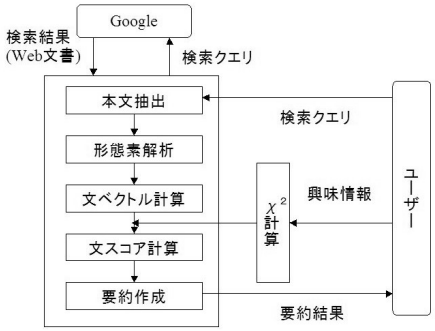

図 1: 概略

正確に表現できることは稀であり、幾つかのクエリを試しながら自分の知りたい情報に近づいていく。検索と連動した informative な要約としては、ユーザの興味や文脈に基づいて要約を作成し、このプロセスを支援することが望ましい。そこで、ユーザが明示的に指定した語や過去の閲覧履歴をユーザの興味情報として利用し、ユーザが知るべき情報を提示するような要約を試みる。

システムの概略を図 1 に示す。 まずユーザからの検索クエリで Google 検索を行い、検索結果を 20 件ダウンロードしてくる。そこから本文を抽出し、本文から計算される単語の重要度 (TFIDF[3] 値) と興味情報 (ユーザの興味のある単語、興味のない単語) を統合して文のスコアを計算し、上位のものからユーザに提示する。

内容的に関係ないリンクや広告を取り除くためにつかった手法は主に３つである。本文中の単語にリンクを張ることはあっても文全体にリンクを張ることは少なく、広告の文などは逆に文全体に張ってあるので、文全体にリンクが張られているものを除く。短すぎる文塊は取り除く。[5]文

[1]http://www.google.com/finalstrutff
[2]http://news.google.com/finalstrutff
[3]例えば goo (http://www.goo.ne.jp/) など finalstrutff
[4]例えば Google (http://www.google.co.jp/) など finalstrutff

Web Documents Summarization Based on Clustering

[†]Takanori Kawahara
Dept. of Information and Communication Engineering,
School of Engineering, University of Tokyo

[‡]Naoaki Okazaki, Mitsuru ISHIZUKA
Dept. of Information and Communication Engineering,
Graduate School of Information Science and Technology,
University of Tokyo

[5]文塊とは、 ... 、<td>...</td>などのタグで囲まれた文の集合のことをいう。finalstrutff

塊の中で（名詞、未知語）と（その他の品詞の語）との比によって重要でないと判断された文塊は取り除く。

単語 w の重要度を以下の式で定義した。

$$score(w) = (1-\alpha)TFIDF(w) + \alpha\chi^2(w) \quad (1)$$

第一項は文書からの情報のみから計算される客観的な単語重要度、第二項はユーザからの入力を元に計算される主観的な単語重要度である。

検索クエリで表現しきれない情報要求は表面上重要でない重要な単語に関連することが多い。これを表す指標として語の以下のような共起の統計情報 [4] をつかう。

χ^2 は単語集合 G に対して理論確率 $p_g(g \in G)$ 、語 w と語群 G の共起の総数 n_w 、語 w と語 $g \in G$ の共起頻度を $freq(w,g)$ とすると、

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g}$$

で表される。これは語群 G との共起の偏りを表す統計量で、値が大きいほど語 w は語群 G と選択的に共起することとなる。この χ^2 という統計量を導入することで、本当は重要単語との関連性が高いにもかかわらず出現頻度が低い単語のスコアをあげることができる。

文 S のスコアは含まれる単語を w_i として以下のように定義した。

$$score(S) = \sum_i score(w_i) \quad (2)$$

最終的な要約の出力方法であるが、文ごとにクラスタリングし、なるべく多くのクラスタから重要な文を取り出すと、話題が偏らない要約が作成できる。しかし、文の中の単語を基底とする文ベクトルの内積によって計算される類似度が一定以上であるときに同一のクラスタとなる、最小距離法でクラスタリングを行うと、文同士の類似度は 0 かそれに近い値になりやすく孤立クラスタが多くなってしまうことと、文全部を対象にするとあまりに組み合わせの数が多すぎて計算量が現実的でなくなってしまうことが問題となる。そこで MMR-MD[5](Maximal Marginal Relevance) を使うことにした。MMR-MD とは検索要求の適合度と情報の新規性 (すでに選択されたものとの異なり度) をともに考慮する尺度である。

紙面の都合で式は省略するが、MMR-MD を使うと、すでに出力してある要約と類似している文はスコアが低くなるため、何度も類似の文を繰り返すような冗長な要約を回避することができる。

3 実装

システムの動作画面は図 2 のようになる。文書ダウンロードから本文抽出 (初期動作) まで約一分、要約出力までは約 2.5 分であった。10 件のみを処理する場合は初期動作 10 秒、要約出力までは 20 秒であった。ユーザは検索クエリとともに、興味のない語、興味のある語を入れる。

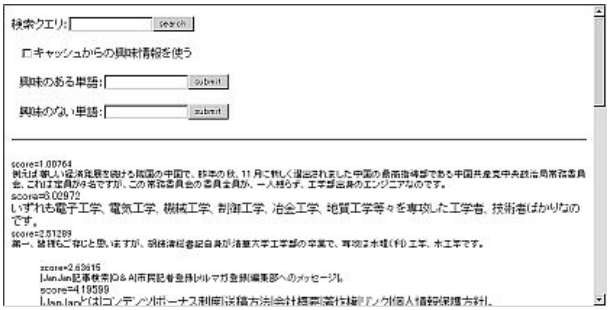

図 2: システムの利用画面

4 結論

HTML 文書から要約の対象とすべき本文を抽出するというところはかなりうまくいっている。

所要時間は検索結果のデータ量に比例するのでキーワードや件数によって大きく所要時間が変わる。

実際に黒船が来たときの通訳の名前を調べてみた。検索クエリを「黒船　通訳　名前」とし、興味のある単語に「通訳　名前」をいれたところ、「通詞（つうじ）とは、通訳者のことで、通弁とも言いました。」という文が得られ、通詞を興味のある単語に加えると「浦賀奉行所与力・香山栄左衛門がオランダ語通詞・堀辰之助、立石得十郎を従えて小舟で漕ぎつけ」という記述を得ることができ、目的を達成することができた。またそれに加え、「アメリカ側にはオランダ語を話せる人がいて、英語→オランダ語→日本語の順で通訳をしました。また、森山栄之助という阿蘭陀通詞は、英語も少しは話せたので、」「●森山栄之助は長崎のオランダ語通詞で、ペリーと幕府が開港条約を結ぶときにオランダ語の通訳をした人です。」などという記述も得られ、関連した知識も得ることができた。

参考文献

[1] Dragomir R. Radev, Sasha Blair-Goldensohn, Zhu Zhang, and Revathi Sundara Raghavan. Newsinessence: A system for domain-independent, real-time news clustering and multi-document summarization.

[2] Paris C. Amitay E. *Automatically Summarising Web Sites - Is There A Way Around It?* 9th International Conference on Information and Knowledge Management (CIKM 2000).

[3] G. Salton and M. McGill. *Introduction to modern Information Retrieval*. McGraw-Hill, 1983.

[4] 松尾豊, 石塚満. 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム. 人工知能学会誌、Vol.17、No.3, pp.213-227.

[5] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction.