Informativeな要約によるWebサイトの情報源としての活用

岡崎直観（共同研究者: 川原尊徳）

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，要約が広く利用されている．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]．これらの手法は，検索結果の適合性の判断に用いるindicative（指示的）な要約に適している．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，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．

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

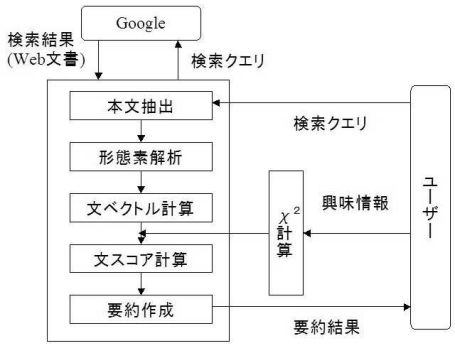

図 1: システム概略

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できるようなシステムを考える．

図1は，検索エンジンで得られたWebサイトのコンテンツから文を単位に情報を切り出し，ユーザにとって必要と思われる文を出力するシステムの概要である．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 10 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 (ここでは TFIDF[3] 値を用いた) と興味情報 (ユーザの興味のある単語，興味のない単語）から文のスコアを算出し，上位のものからユーザに提示する．単語 w の重要度は以下の式で定義した．

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

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ある．第二項はユーザから与えられた興味情報と偏って共起する語に高い重要度を与えるものあり，共起回数のカイ二乗値 [4] を使った．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．G をユーザが興味のある単語の集合として，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上げることができる．

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．

$$score(S) = \sum_{i \in S} score(w_i) \qquad (2)$$

要約の出力方法であるが，抜き出す文の話題が偏らないようにするため，MMR-MD (Maximal Marginal Relevance) [5] という尺度を導入した．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り，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 結論

図 2 にシステムの動作画面を示した．ユーザがクエリを入力すると，検索された Web サイトの内容の中で重要と思われる文が出力される．このとき，ユーザは興味のある語，興味のない語を入力して，

図 2: システムの利用画面

システムに興味情報を明示的に与えるか，ブラウザのキャッシュを興味情報として利用するか選択できる．ユーザはこの要約を読みながら，興味が沸いた語を新たに興味情報に追加したり，クエリを変更しながら自分の欲しい情報を獲得していく．

プロトタイプシステムでは，文書ダウンロードから本文抽出までを約 10 秒，要約出力までを 20 秒程度で処理している．検索結果中に大きなコンテンツがあると，単語や文の数が増大してしまい，要約作成までの所要時間が長くなることがある．

現在，本システムの実装ならびに評価を進めている段階であり，定量的評価の結果などは今後明らかにしていく予定である．

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. In *Proc. Human Language Technology Conference*, 2001.

[2] Eniat Amitay and Cecile Paris. Automatically summarising web sites - is there a way around it? In *Proc. of 9th International Conference on Information and Knowledge Management (CIKM 2000)*, pp. 173–179, 2000.

[3] Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of are Information by Computer*. Addison-Wesley, 1989.

[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, 2002.

[5] J. Goldstein, V. Mittal, J. Carbonell, and M. Kantrowitz. Multi-document summarization by sentence extraction. In *Proc. of ANLP/NAACL Workshop on Automatic Summarization*, pp. 40–48, 2000.