社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (21)
2007－BIO－11 (21)
2007／12／20

複数 Web ページの重要文抽出および直感的理解を支援するための GUI の開発

柴田 裕子　山内 和子　石川 千里　高田 雅美　城 和貴
奈良女子大学大学院 人間文化研究科 情報科学専攻

概要：近年，Web 空間は情報収集における重要な情報源のひとつとなった．しかし，Web 空間には多種多様な情報が氾濫しているため，検索エンジンを用いて必要な情報を得ようとしても検索結果の数は膨大であり，ユーザが情報過多による混乱を起こす恐れがある．そこで本研究では，より効率的に必要な情報を取得するため，検索結果から得られる複数 Web 文書から重要文とキーワードを抽出するモデルを提案する．本稿では，モデルの提案と同時にユーザの直感的理解を視覚的に支援する GUI の開発を行った．本モデルでは，検索結果として得られる複数の文書に目を通さなければならないという人間にかかる負担を軽減させることを目的としている．

Effective Method for Web Exploration with a Novel GUI for Easy and Intuitive Understanding

Yuko Shibata, Kazuko Yamauchi, Chisato Ishikawa, Masami Takata, Kazuki Joe
Department of Information and Computer Sciences,
Graduate School of Humanities and Sciences, Nara Women's University

Abstract : Lately, Web space has been one of the most important resources for information exploration. However, users are sometimes confused because of various information floods in the Web space. In this paper, we introduce a method which extracts important sentences and keywords from plural Web documents provided from search results to acquire necessary information more effectively. Furthermore, we developed a GUI which supports visually easy and intuitive understanding for the users. Our method with the GUI reduces users' burden for web exploration drastically.

1. はじめに

近年，WWW（World Wide Web）などインターネット技術の発展により，我々がそこから入手可能な情報の量は爆発的に増大している．また同時に検索エンジンのような情報検索システムが広く利用されるようになり，探したい情報に関連する文書を瞬時に得ることが出来るようになった．しかし検索エンジンからの検索結果の数は膨大であるため，検索結果から文書を選択し，それらすべてを読んで理解するには時間がかかるという問題点がある．

そこで本稿では，複数 Web 文書から重要文とキーワードを抽出するモデルを提案し，同時にユーザの直感的理解を視覚的に支援する GUI の開発を行う．本モデルは，関連する複数文書の重要文や重要キーワードを抽出することにより，検索結果として得られる複数の文書に目を通さなければならないという人間にかかる負担を軽減させるという点から，情報探求者の求める情報の直感的理解を支援することを目的としている．現在様々な検索支援システムが存在するが，用語の定義文を自動的に抽出し，提示するシステム[1]，[2]など特定の種類のクエリを対象としているものが多く，多種多様な検索要求に対応できない．そこで，本稿ではクエリの種類を限定しない検索支援モデルを提案する．

次章ではモデルの概要について述べる．第 3 章ではモデルで使用した重要度計算の詳細を述べる．第 4 章では実装した GUI について述べる．第 5 章でモデルの実行速度に関する考察を述べたあと，第 6 章でまとめと今後の課題について述べる．

2. モデルの構造

本研究で提案するモデルの概要を図 1 に示す．

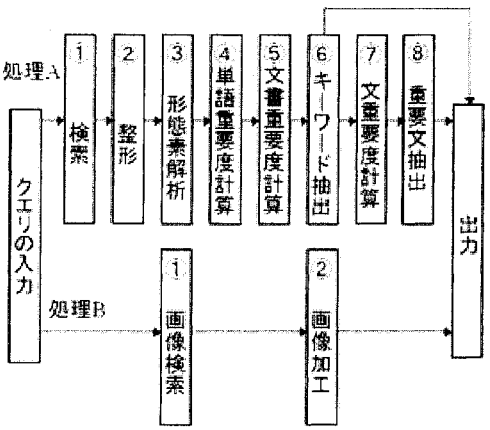

図 1：モデルの概要

クエリの入力後，処理 A と処理 B は並列して行う．それぞれの処理について詳細を以下に示す．

A-1) **検索**：検索クエリを受け取り，Yahoo! JAPAN[3]による検索を行い，検索結果の上位から指定した件数の Web ページとタイトルを取得する．フレーム構造のページや PDF ファイルは対象外とする．重要文抽出に用いる Web ページへのリンクをユーザに提示する．

A-2) **整形**：取得した各ページの HTML タグを除去し，文章を「.」「？」などで区切り，文単位に分割する．また文字実体参照である「&」「>」などを元の文字の「&」「>」などに変換する．各ページを整形したものを文書と定義する．

A-3) **形態素解析**：整形を終えた文書に対し，形態素解析を行う．形態素解析には茶筌[4]を用いる．

A-4) **単語重要度計算**：取得された文書全体において，形態素解析で名詞と判別された単語に関して出現頻度，文書頻度などを用いて重要度を計算する．ただし，格助詞や代名詞，「こと」「もの」などの非自立な名詞は処理の対象外とする．詳細は 3.1 節において述べる．

A-5) **文書重要度計算**：A-4 において求められた単語の重要度を用いて文書の重要度を計算する．詳細は 3.2 節において述べる．

A-6) **キーワード抽出**：A-4 において求められた重要単語をキーワードとし，求められた各文書のキーワードを文書重要度順に並び替え，ユーザに提示する．

A-7) **文重要度計算**：A-5 において求められた文書の重要度が上位 3 文書である文書のすべての文について，単語の重要度を用いて文の重要度を計算する．詳細は 3.3 節において述べる．

A-8) **重要文抽出**：A-7 において求められた文の重要度のうち，各文書上位 3 文を重要文としてユーザに提示する．

B-1) **画像検索**：検索クエリを受け取り，Yahoo! JAPAN による画像検索を行い，検索結果の上位から指定した件数のサムネイル画像を取得し，取得したものから随時画像加工を行う．

B-2) **画像加工**：B-1 で取得したサムネイル画像に影を付けるなど加工を施す．

3. 重要度計算の詳細

3.1 単語重要度の計算

単語の重要度は各文中の単語における tf*idf 値を採用する．tf*idf 値は，ある文書中に単語 w が現れる頻度$tf(w)$ 値と，ある文書群の中で対象単語が現れた文書の数$df(w)$ 値とを組み合わせて計算される値で，$tf(w)$値はある単語に関して，その単語を多く含む文書ほどその単語について詳しく説明しているものと考えるものであり，一方$df(w)$ 値は多くの文書で使用されている単語より，少ない文書で使用されている単語の方が重要性の高いと考えるものである．これらにより tf*idf 値は文書中のある単語が，どの程度その文書特有の単語であるかを示す．ある文書群の文書総数が N であるとき，各単語における tf*idf 値の計算式は式(1)のようになる．

$$tf * idf(w) = tf(w)\log\frac{N}{df(w)} \quad (1)$$

本稿では対象が Yahoo! JAPAN にインデックスされている全 Web 文書であるので Yahoo! JAPAN を巨大な

文書データベースとみなし，文書総数 N を Yahoo! JAPAN の総インデックス数，$df(w)$ 値を Yahoo! JAPAN での検索ヒット数とした．

また，検索クエリに含まれる単語は重要度を 2 倍にし，

$df(w)$値が 1 以下のときは tf*idf 値を 0 とする．続いて各文書における全単語のうち，重要度が高いもの上位 10 単語を重要単語リストとして保持する．それ以外の単語は重要度を 0 とする．

3.2 文書重要度の計算

3.1 節で定義した重要度をもとにして，文書 i の重要度 DI_i を式(2)で計算する．

$$DI_i = \sum_{w \in TL} \{TI_w * tf_i(w)\} * \frac{\sum_{w \in TL} tf_{wi}}{\sum_{w \in DT_i} tf_{wi}} \quad (2)$$

$$TI_w = tf * idf(w) \quad (3)$$

ここで TL は重要単語リストに含まれる単語集合，DT_i は文書 i に含まれる単語集合である．

式(2)の右辺前半部分は，重要単語リストの単語を多く含む文書ほど DI_i の値を高くする．さらに後半部分で，文書 i に含まれる総単語数と，その中で重要単語リストに含まれている単語数の割合を掛ける．これにより相対的に重要な単語を含む文書ほど DI_i が高くなる．

次に，クエリに含まれるすべての単語が含まれている文がひとつ以上存在する場合，DI_i に α を乗ずる．これは α の値を適切に設定する事で，クエリと文書の関連付けを強化できると考えられるからである．本稿では，α =1.2 とした．

3.3 文重要度の計算

3.2 節で求められた文書重要度が高い文書上位 3 文書それぞれにおいて，文書 D に含まれる k 番目の文 k の重要度 SI_{Dk} を式(4)により求める．

$$SI_{Dk} = \frac{1}{k} * \sum_{i \in TL} (TI_i * tf_{iDk}) * \prod_{q \in Q} (tf_{qDk} + 1) \quad (4)$$

ここで Q はクエリに含まれる単語集合，tf_{iDk} および tf_{qDk} はそれぞれ単語 i とクエリに含まれる単語 q の文 k における出現頻度である．式(4)は，文書の冒頭に位置する文ほど重要度が高くなり，かつ重要な単語やクエリに含まれる単語を多く含む文ほど重要度が高くなるようになっている．

また，文の長さが一定の値 C より短い文にはペナルティを与える必要がある．これは文の長さが一定値以下の文には含まれる情報が極端に少なく，重要文として抽出されるには不適切であるという考えに基づいている．本稿ではしきい値 C を 20(文字)とした．さらに，クエリに含まれる単語をすべて含む文の重要度を極めて大きな値に設定する．これは重要文を抽出する際にクエリに含まれる単語すべてを含む文が抽出されやすくするためである．

4. GUI プロトタイプの実装

本稿で開発する GUI は，Web 検索結果から得られる複数文書を要約するという特徴から，Web 上で使用することが考えられる．さらに視覚的に情報の直感的理解を支援するためには，インタラクティブな動きによる情報の提供が不可欠と考え，Flash を用いて実装を行った．GUI の画面構成を図 2 に示す．

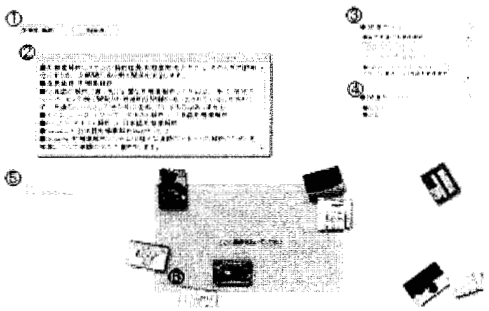

図 2：GUI の画面構成

各部位の詳細は以下のようになっている．

① **クエリ入力部**：検索したいクエリを入力し，Search ボタンを押すことによって画像検索および Web 検索を行う．

② **重要文表示ウィンドウ**：検索された複数の Web ページから抽出した重要文を表示する．右上に 3 つのボタンが配置されており COPY ボタンをクリックすると重要文がクリップボードにコピーされ，URL ボタンをクリックすると関連ページ表示ウィンドウが表示され，KEYWORD ボタンをクリックすると関連キーワード表示ウィンドウが表示される．ウィンドウ上部のバーをドラッグすることで自由に移動することができる．

③ **関連ページ表示ウィンドウ**：重要文抽出に用いられた Web ページのタイトルと URL を表示する．右上の×ボタンを押すことでウィンドウを非表示にすることができる．再度表示させるに

は重要文表示ウィンドウの URL ボタンをクリックする必要がある．

④ **関連キーワード表示ウィンドウ：**3 節の計算によって求められたキーワードを表示する．キーワードをクリックすることでキーワードをクエリとして Yahoo! JAPAN で検索を行うことができる．右上の×ボタンを押すことでウィンドウを非表示にすることができる．再度表示させるには重要文表示ウィンドウの KEYWORD ボタンをクリックする必要がある．

⑤ **画像表示部：**クエリから検索された画像のサムネイル画像を右から左へ，水の中を流れるように表示させる．画像はドラッグすることができる．また，ダブルクリックすることで元画像を開くことができる．

⑥ **画像保持部：**画像表示部を流れる画像をここに置くと，画像がその場に固定され保持することができる．関連性の高い画像を保持することにより，提示された重要文と合わせることができ，より直感的な理解がしやすいと考えられる．

5. 考察

Web 上で実行されるシステムに必要とされるのはタスク処理時間の短さである．現在，15 件の検索に対し 20 秒ほどの処理時間がかかり，そのうち 12 秒が Web ページのダウンロードに必要な時間である．これは複数スレッドにより同時に複数の Web ページをダウンロードすることで処理時間を短縮することが可能と考えられる．さらに，7 秒が茶筌による形態素解析に必要な時間である．これは，茶筌より高速な処理が望める和布蕪[5]を使用することによって解決できると考えられる．

6. まとめ

Web 検索結果から得られる文書集合に対し，単語の出現頻度を用いて重要な単語を抽出し，それを用いて重要文を抽出することにより，膨大な量の文書すべてに目を通さなければならないという人間にかかる負担を軽減させるモデルを提案した．また，Flash を用いたユーザの直感的理解を視覚的に支援する GUI を提供することにより，重要文を抽出する間にかかる時間にもクエリに関する情報を得ることが可能となった．

今後の課題として，実行速度の向上が不可欠である．また，本モデルでは Web ページの本文のみを処理の対象としているが，画像に含まれる情報を考慮したり，<b>や<i>など強調に用いられる特定の HTML タグに囲まれた単語の重要度を高くすることにより，より精度を高められると思われる．さらに，文書集合をクラスタリングすることで関連性の高い文書群を作成でき，それら文書群から抽出される重要文はより精度が高いと考えられるので，今後適用していく予定である．

参考文献

[1] 藤井敦，石川徹也：World Wide Web を用いた事典知識情報の抽出と組織化．電子情報通学会論文誌 D-II,Vol.J85-D-II, No.2, pp.300-307,(2002).
[2] 桜井裕，佐藤理史：ワールドワイドウェブを利用した用語説明の自動生成．情報処理学会論文誌 Vol.43, No.5, pp.1470-1480 (2002).
[3] http://yahoo.co.jp/
[4] http://chasen-legacy.sourceforge.jp/
[5] http://mecab.sourceforge.net/