1G2-2

複数文書の単語情報を用いたWebページの自動要約

Automatic Summarization of Web pages using Term Information in Multiple Documents

井上 龍太郎[*1] 内海 彰[*2]
Ryutaro Inoue Akira Utsumi

[*1]電気通信大学大学院 電気通信学研究科 システム工学専攻
Department of Systems Engineering, Graduate School of Electro-Communications, The University of Electro-Communications

[*2]電気通信大学 電気通信学部 システム工学科
Department of Systems Engineering, Faculty of Electro-Communications, The University of Electro-Communications

This paper proposes a method to rank and summarize Web pages based on the importance score of the contained words calculated by TF, DF, and variance of frequency. Using the propsed method, a Web search support system is developed which shows a ranking of relevant Web pages to a query and creates a summary of the top-ranked page. The effectiveness of the system is evaluated by 50 users with more than 200 queries.

1. はじめに

今日，Web を使って情報を得る手段としてサーチエンジンが頻繁に用いられている．しかし，その検索結果には不必要な情報が混入することが少なくなく，また，結果として表示されるものは URL やスニペットと呼ばれるクエリ周辺の断片的な文章であり，実際にページの内容を確認しないと自分の目的に合った内容であるか判断し難い．そのため，必要な情報を得るまでには時間や労力を割かなければならない．

そのとき，ユーザーの求めている情報そのものや，求めている情報について詳しく記載されているページを自動的に提示することができれば情報を得るまでに必要なユーザーの負担を大幅に減らすことができると考えられる．

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] などの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，本研究ではクエリの種類を限定しない検索支援システムを提案する．本システムは Web 検索の結果の文書集合からクエリと関係の強い単語を抽出し，それを用いて検索要求に最も合致する Web ページを提示する．さらに，先に抽出した単語の情報を用い，そのページから検索要求に応じた要約文を生成し，提示することにより検索支援を行う．

2. 検索支援システム

本研究で構築したシステムの概要を図 1 に示す．なお，4 から 6 の重要度計算部の詳細は次節で扱う．

1. **ページ取得部：** 検索クエリを受け取り，Google による検索を行い，検索結果の上位から指定した件数を取得する．PDF ファイルは対象外とする．
2. **整形部：** 取得した各ページからタグで定義されたタイトル文を抽出する．次にタグを除去し，文章を改行,「。」,「？」などで区切り，文に分割する．ページからタグを除去したものを文書と定義する．

連絡先: 井上 龍太郎, 電気通信大学大学院 電気通信学研究科 システム工学専攻, 〒 182-8585 東京都調布市調布ヶ丘 1-5-1, 0424-42-5258, ryu@utm.se.uec.ac.jp

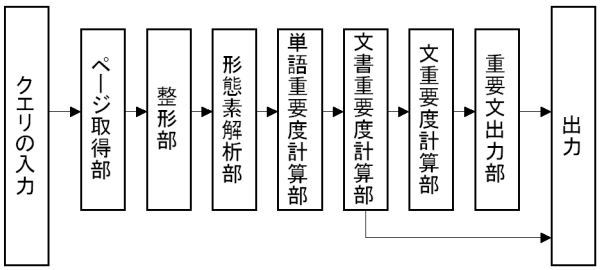

図 1: システムの概要

3. **形態素解析部：** 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[*1]を用いる．
4. **単語重要度計算部：** 取得し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ついて，出現頻度，文書頻度，分散などを用いて単語の重要度を計算する．ただし，「もの」,「こと」などの非自立の名詞や代名詞は処理の対象外とする．
5. **文書重要度計算部：** 手順 4 で計算された単語の重要度を用いて文書の重要度を計算し，文書をランキングする．そして上位 2 件の文書へのリンクをユーザーに提示する．
6. **文重要度計算部：** 最も重要度の高い文書に含まれる全ての文について，単語の重要度を用いて文の重要度を計算する．
7. **重要文出力部：** 重要度の上位 3 文それぞれから重要度がある閾値以下の文にあたるまで，文を順番に出力し続ける．閾値は 0 に設定されており，平均して全部で 8 文前後が出力される．

3. 重要度計算部の詳細

3.1 単語重要度の計算

文書集合に含まれる単語 i について文書集合全体における出現頻度 tf_i，単語 i を含む文書の数 df_i，各文書での出現頻度に

*1 URL;http://chasen.naist.jp/hiki/ChaSen/

対する分散 var_i を用いて，式 (1) で単語 i の重要度 TI_i を計算する．

$$TI_i = \frac{tf_i \cdot df_i}{\log(var_i) + 1} \quad (1)$$

式 (1) は「多くの文書に満遍なく頻出する単語は重要である」という仮定に基づいており，そのような場合に高い値を取るようになっている．ただし，df_i が 1 のときは TI_i を 0 とする．また，クエリに含まれている単語は重要度を 2 倍にする．

続いて，重要度により全単語を降順に並べ変え，上から全体の 5%分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 節で定義した単語の重要度を用い，文書 j の重要度 DI_j を式 (2) で計算する．

$$DI_j = \sum_{i \in TL} (TI_i \cdot tf_{ij}) \cdot \frac{\sum_{i \in TL} tf_{ij}}{\sum_{i \in DT_j} tf_{ij}} \quad (2)$$

ここで TL は重要単語リストに含まれる単語集合，DT_j は文書 j に含まれる単語集合，tf_{ij} は単語 i の文書 j における出現頻度である．

式 (2) の右辺の前半部分により重要単語リストの単語を多く含む文書ほど DI_j の値が高くなる．さらに後半部分で，文書 j に含まれる総単語数とその中で重要単語リストに入っ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多く含む文書ほど重要度が高くなる．

次にクエリと文書との関連付けを強化するために，クエリに含まれる全ての単語を含む文が 1 つ以上存在する文書の重要度を 1.2 倍にする．また，文書のタイトル文が重要単語リスト中の単語を 1 つも含まない場合，その文書の重要度を 0.5 倍する[*2]．これは文書のタイトルが文書集合に共通する話題と全く関係のない単語から構成されている場合，その文書の重要度を下げるためである．

3.3 文重要度の計算

文書重要度計算部により計算された最も重要度の高い文書 D における文 k の重要度 SI_{Dk} を式 (3) で計算する．

$$SI_{Dk} = \sum_{i \in TL} (TI_i \cdot tf_{iDk}) \cdot \prod_{q \in Q} (tf_{qDk} + 1) \quad (3)$$

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 の文書 D の文 k における出現頻度である．式 (3) は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，クエリに含まれる単語を全て含む文があった場合，その文の重要度を極めて大きな値に設定する．これにより要約文を出力する際にその文が必ず選択されるようになる．このような文が複数ある場合，文書中での文番号 k の逆数を掛けることにより調整を行い，冒頭に位置する文ほど重要度が高くなるようにする．

4. 評価実験

本システムが文書重要度計算部で計算するランキングや最終的に出力される要約文を評価するために 2 種類の評価実験を行った．

*2 これらの係数は 4.1 節の実験により調整し，決定した．

表 1: ランキングの評価結果

	本システム	Google
得点合計	51.8	50.9
相関係数	0.599	0.345
正解データのランク	3.00	3.89

4.1 ランキングの評価

文書重要度の計算方法について，人手により作成した正解データを用いて評価を行った．

評価方法

実験者が用意したクエリ 9 件について，1 件あたり 7 人の被験者に Google での検索結果の上位 15 件のページがそれぞれどの程度クエリに合致した内容であるか，0 点（全く合致していない）から 4 点（完全に合致している）の 5 段階で評価してもらった．この評価データを基にランキングの正解データを作成し，以下の項目において本システムと Google との比較を行った．

- **得点合計：** 本システムおよび Google が出力したランキング上位 3 ページの得点の合計．4 点 ×3 ページ ×7 人 =84 点が満点となる．
- **相関係数：** 15 件のページに対する正解データのランクと本システムおよび Google が出力したランクとの順位相関係数．
- **正解データのランク：** 本システムおよび Google のランキング 1 位のページの正解データにおけるランク．

結果と考察

9 件のクエリについての評価を平均した結果を表 1 に示す．全ての項目において Google より優れた結果となった．本システムの手法によって，人間による評価により近いランキングを行うことができたと言える．

Google との主な違いとして，相関係数が大幅に高くなっていることから，特にクエリに対し適切ではないページを正しく判別できていると考えられる．

4.2 システム全体および要約文の評価

検索クエリを限定せずに，様々な検索要求における本システムのランキングおよび要約文の適切さの評価を行った．

評価方法

Web 上に実装した本システムを用いて，被験者に自由にクエリを選んで検索してもらった．そして，本システムが出力した結果を以下の項目について評価してもらった．

- **ランキングについて**: 本システムが出力したページ上位 2 件と Google での上位 2 件の計 4 件のページをランダムな順番で提示し，クエリに合致した内容かどうかを 0 点（全く合致していない）から 4 点（完全に合致している）の 5 段階で評価してもらった．それぞれ 2 件分の得点の合計をシステムおよび Google の評価点とした．
- **要約について**: システムが出力した要約文について，以下の 3 項目を評価してもらった．
  1. 知りたいことが含まれているか（包含度）: 0 点（全く含まれてない）から 4 点（十分に含まれている）の 5 段階評価．

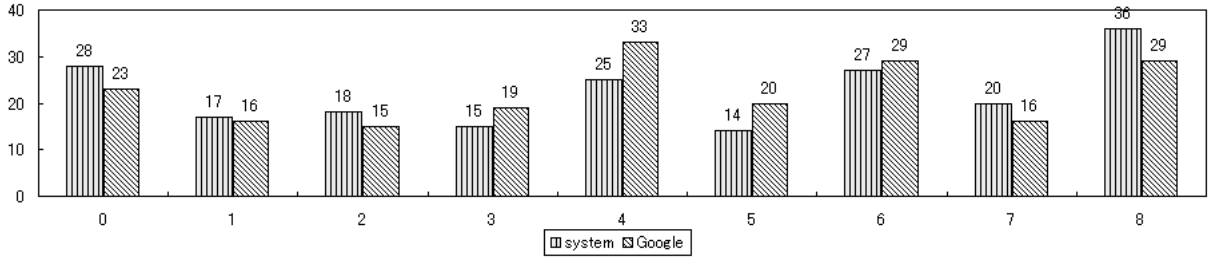

図 2: ランキングの評価点のヒストグラム

表 2: 本システムと Google のランキング結果

	本システム	Google
評価点の平均	4.29	4.27
評価点の標準偏差	2.79	2.57
評価点の最頻値	8(36 回)	4(33 回)
6 点以上だったクエリ数	83	74
2 点以下だったクエリ数	63	54

2. 読みやすさ：0 点（読みにくい）から 4 点（読みやすい）の 5 段階評価.
3. 総合評価：0 点（適切な要約文ではない）から 4 点（適切な要約文である）の 5 段階評価.

被験者は Web 上で募集し，約 50 人の参加者から得た 200 件の評価データについて集計を行った.

ランキングの評価結果と考察

本システムが出力した合計 400 ページのうち 330 ページが Google 上位 2 件の結果とは異なるページであった．よって，ほとんどのクエリにおいて Google とは異なるページを出力していたと言える.

ランキングに関する本システムと Google との比較結果を表 2，得点のヒストグラムを図 2 に示す．評価点の平均は Google と同等の結果となったが，Google に比べ 6 点以上だったクエリ数が多く，最頻値が 8 点であることから，本システムのランキング結果はおおむね良好であると言える．しかし，得点の分布は適切にランキングできた場合とそうではない場合の両極端であった.

システムの評価点が低くなった原因の 1 つに，検索結果に単語の羅列だけのようなノイズとなるページが混在している場合がある．単語の情報だけではこのようなページがノイズかどうか判別できないため，タグの情報などを用いて予め排除をしておく必要があると考えられる．また，検索クエリが画像や表などを求めるものであった場合や，クエリが明らかに不適当な場合などは本手法だけでは対応できず，評価点が低くなっていた.

続いて，ページごとに評価点を見ていくと，掲示板や blog は総じて評価点が低くなっていた．これらのページにはクエリと関係のない様々な話題が含まれていることが多く，ノイズとなる場合が多い．しかし，感想や評判を求める検索クエリの場合は blog なども評価点が高く，一概に排除はできない．よって，基本的にこれらのページは排除するが，クエリの内容によっては排除しないなどの方法を採る必要があると考えられる.

要約文の評価結果と考察

結果を表 3，得点のヒストグラムを図 3 に示す．良評価のクエリとはシステムの評価点が 4 点以上だったときの 122 件のクエリである.

表 3: 要約文に関する評価結果の平均値

	包含度	読みやすさ	総合評価
全てのクエリ	1.77	2.50	1.90
良評価のクエリ	2.42	2.85	2.40

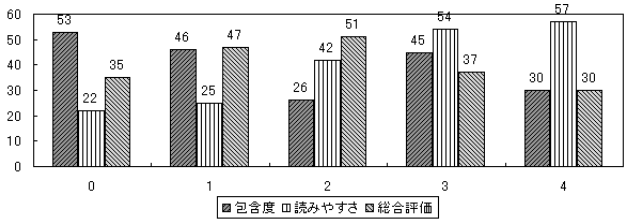

図 3: 要約文の評価点のヒストグラム

全てのクエリにおける平均に比べ，良評価のクエリにおける平均が全ての項目について大幅に向上している．ランキングが適切で，良いページを正しく判定できた場合は要約文の評価も高くなると言える．このことから要約の方法自体に大きな問題はなく，むしろ要約元のページの選定が重要であると思われる.

検索クエリが日付を解答とする場合であれば，日付に関する記述を抽出するなどの解答のタイプに応じたパターンマッチングを併用することにより，さらに適切な要約文を作ることができると考えられる.

5. おわりに

Web 検索結果の文書集合における単語の出現情報を用いて重要な単語を抽出し，それを用いて重要文書を決定した後，要約文の出力を行うシステムを提案した．本システムの手法がユーザーの検索要求に合致した要約文やページを提示するのに有効であることが示された．今後の課題として，実行速度の向上が必要である．現在，15 件の検索結果に対して 15 秒ほどの処理時間が掛かり，うち 11 秒がページのダウンロードに要する時間である．これは，複数スレッドにより同時にダウンロードすることで処理時間を大幅に短縮することが可能であると考えられる．また，本システムはページ中の文章しか考慮していないが，画像や表などに含まれる情報も考慮することによりさらに精度良くランキングを行うことができると思われるので，そのような手法を開発することを考えている.

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. 電子情報通信学会論文誌 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).