単語の反復度と共起頻度に基づく関連記事の提示方法

A method for retrieving related articles using both word adaptation and co-occurrence histogram

島田 諭† Satoshi Shimada
佐藤 哲司† Tetsuji Satoh

1. はじめに

ウェブにおける情報発信のコストは低く，単独では内容が完結していない断片的な記事でも気軽に公開される傾向にある．このような断片的な記事は，関連する記事を合わせて読むことで内容が把握できることから，関連記事の提示方法が大きな課題になっている．

本稿では，記事集合中に出現する単語の反復度および共起頻度を用いて，ある一つのトピックを形成していると思われる単語集合を求め，その単語集合を含む記事を関連記事として提示する手法について述べる．

2. 単語の重要度の算出

本稿では，「記事中の複数の単語によって表現・形成される概念」をトピックと定義する．一つの記事は複数の単語からなり，一つの記事中に表現されているトピックは必ずしも一つとは限らないとする．また，記事集合の中で，互いに関連するトピックは，一部が重複する単語集合によって表現されるものとし，この関係を利用して記事の中からトピックを形成していると思われる単語集合を抽出する．

トピックを形成する単語の候補は，専門用語を含めて漢字，カタカナ，英数字で書かれていることが多いといえる．また，トピックを語の集合として扱うため，個々の単語の抽出精度より，文の長さや表現の違いなどの影響を受けずに安定して単語が抽出できることのほうが重要になる．このため，本稿では次に述べる方法で単語の抽出および重要度の算出を行なう．

・単語の抽出…文字種の変わり目を利用
　漢字とカタカナからなる語，および英数字と一部の記号からなる語を抽出する．ただし，1 文字の漢字またはカタカナからなる語，2 文字以下の英数字からなる語，数字のみの語，およびURLとして解釈できる書式の文字列を除く．
・単語の重要度の算出…反復度を利用
　ある文書において一度出現した単語が再度出現する度合いが語の種類と密接な関係を持っていることが知られている[1]. 本稿では[1]に示された以下の反復度(adaptation)を用いる．

$$Adaptation(w) = df_2(w)/df(w)$$

$df(w)$：単語 w が 1 回以上出現する文書の数
$df_2(w)$：単語 w が 2 回以上出現する文書の数

例えば，「つくばエクスプレスが開業して便利になった」というトピックを含む記事があるとする．上記の方法では「つくばエクスプレス」という固有名詞は抽出されず，図 1 のように「エクスプレス」「開業」「便利」「秋葉原」「駅前」といった単語が抽出される．これらの語の集合がトピックを形成する．トピックを形成する単語集合は一通りではなく，「開業」「便利」の組み合わせや「駅前」「秋葉原」の組み合わせなど，抽出した単語の部分集合もトピックとなりうる．

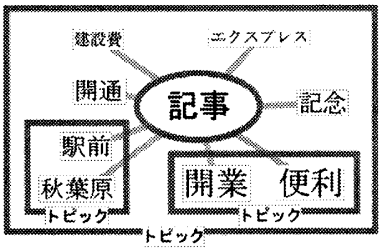

図 1 トピックを形成する単語の例

3. 単語集合を用いた関連記事の提示手法

本稿では，「一つ以上のトピックを共有する記事」を関連記事と定義する．

情報検索システムを評価する際の尺度として，検索結果がどれだけ所望の情報をカバーしているかを示す網羅性（exhaustivity）と，どれだけ所望の情報に絞った内容を含むかを示す特定性（specificity）という尺度が知られている[2]. 本稿ではこの尺度を応用し，抽出した単語の反復度を利用して，それぞれの単語を「網羅性を示す単語」と「特定性を示す単語」とに分類する．ある記事と別の記事との間で共起する単語について，特定性を示す単語に高いスコアを，網羅性を示す単語に低いスコアを与え，記事間の関連度を求める．関連度の高い上位数件の記事を関連記事として提

† 筑波大学大学院図書館情報メディア研究科，
Graduate School of Library, Information and Media Studies, University of Tsukuba.

示する．

以上により，記事集合中で頻繁に言及されているトピックを含む記事に対しては，より代表的な記事が関連記事として提示される．一方，記事集合中で言及される頻度が低いトピックを含む記事に対しては，周辺的なトピックを含む記事が関連記事として提示される．また関連記事の探索が容易になるよう，どんな記事にも一定数の関連記事が提示されるようにした．

表 1　共起頻度計算のため単語に与えるスコア

反復度の範囲	文書数の閾値	スコア
$0.6 \leqq df_2/df$	$df_2 > 2$	10
$0.35 \leqq df_2/df < 0.6$	$df > 9$, $df_2 < 19$	1
（上記以外）		0.01

反復度の範囲について，比較的小規模な記事集合から抽出されたすべての単語を予備調査した．その結果，概ね上記の範囲ごとに性質の異なる単語が分布していたことから，本稿では表 1 に示す範囲で区切ってスコアを与える．ただし，関連記事を提示するという目的に対して，記事集合全体で出現文書数の非常に少ない語は適さないため，$df < 3$ となる単語には低いスコアを与える．同様に，df_2 が十分に大きくなるとされる[1]，記事集合において非常に一般的な語（例えば質問回答サイトにおける「質問」や「場合」など）や常套句についても，$df_2 \geqq 20$ となる単語には低いスコアを与える．なお，20 という値は記事集合に依存する．

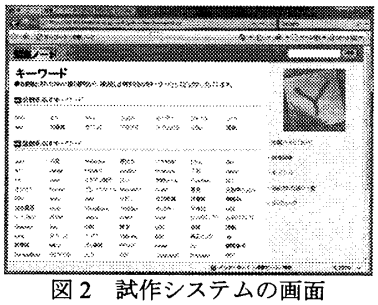

図 2　試作システムの画面

Yahoo!知恵袋データのカテゴリ「デジタルカメラ」「パソコン，周辺機器」「家電，AV 機器」の記事を用いて提案手法を実装した(図 2)．記事集合全体の特徴を簡潔に示すために，反復度に基づいて求めた「網羅性を示す単語」と「特定性を示す単語」を提示している．

2005 年 10 月分の質問と回答，あわせて 43,834 件の記事から，56,031 個の単語が抽出された．これらの単語の中から「網羅性」および「特定性」を示す単語を求めたところ，表 2 に示す単語が提示された．いずれも，一般的な語を除くことができている．

表 2　提示された単語の例

網羅性を示す単語　14 個 $(0.35 \leqq df_2/df < 0.6)$ かつ $(df_2 \geqq 20)$
DVD セル iPod フォント ルーター フォーム シート amp 万画素 サウンド アダプタ オブジェクト イラレ 置換
特定性を示す単語　109 個 $(0.6 \leqq df_2/df)$ かつ $(df_2 \geqq 3)$
quot 行目 ttfCache 肥大化 COUNTIF ドラム Set A10 Range PC9801 MsgBox Protocol 入社 Ghost int color イヤホン端子 スジ 未読メール Function SLI オンオフ Format グレースケール MicroATX CHAR 畳用 丸数字フォント PSX Auto font リボン 自然乾燥 （以下省略）

このうち「万画素」という単語について，提案手法を用いて関連単語および関連記事を求めたところ，表 3 が得られた．「万画素」を含む記事は 136 件あり，頻繁に言及されているトピックが提示できているといえる．

表 3　「万画素」の関連単語および関連記事

関連単語
デジカメ 画素数 写真 購入 粒子 絵柄 ドット 画質
関連記事　上位 8 件
1. プリンタによります ピクセル数はあくまで点の数ですので
2. 殆どがＬ判なら 500 万画素のほうが 1 個の素子の大きさが
3. L 判のプリントで画素数の違いが判らなくなるのは 800 万画素
4. 先々月にパナソニックのＮＶ-Ｇ Ｓ250 を購入した物です
5. 初心者です 教えていただきたいので質問いたします
6. コンデジで 500 万画素と 800 万画素はどのくらい差が
7. まず「高画素＝高画質」ではないしＬ判サイズであれば
8. この質問であればかなり的を絞り込んだ回答が可能です

4. おわりに

本稿では，記事集合中に出現する単語の反復度および共起頻度を用いて，記事集合の特徴を示す単語や関連記事を提示する手法について述べた．提示する関連記事の適切性については，主観評価を含めて今後，検証していく．

さらに，共起頻度計算のために単語に与えるスコアについて，記事集合の規模や性質に従って適応的に決定する手法についても検討していく予定である．

本研究の実施にあたっては，ヤフー株式会社が国立情報学研究所に提供した Yahoo!知恵袋データを利用した．

参考文献

[1] 武田善行, 梅村恭司：キーワード抽出を実現する文書頻度分析, NL-146-5, pp.27-32 (2001).

[2] 酒井哲也：よりよい検索システム実現のために：正解の良し悪しを考慮した情報検索評価の動向, 情報処理, Vol.47, No.2, pp. 147-158 (2006).