Web情報検索における先進的言語処理技術の大規模評価

新里 圭司　柴田 知秀　黒橋 禎夫
京都大学大学院 情報学研究科
{shinzato, shibata, kuro}@nlp.kuee.kyoto-u.ac.jp

1　はじめに

「情報爆発」という言葉で形容されるように，現在ウェブ上には，膨大な量の情報が発信されており，その中から求める情報を得るためには検索エンジンの利用は必要不可欠である．しかしながら，既存の検索エンジンは検索クエリ中の語の出現や，ページへの被リンク数を主な手がかりにクエリとページの関連度を測るため，いわゆる調査型 [1] と呼ばれるクエリに対しては高い精度で有用なページを提示できるとは言い難い．

既存の検索エンジンの抱える問題として以下の点が挙げられる．

自然文で表現されたクエリを扱えない: 既存の検索エンジンは，入力として数個のキーワードを想定しており，自然文で書かれた検索クエリ（以下，NLQ とする）は適切に扱えない．しかしながら，数個のキーワードのみでは，検索要求を表現することは明らかに不十分である．NLQ の利点として以下が挙げられる．

- ユーザが自分の要求をより直接的に表現できる
- クエリに含まれる単語間の関係をシステムが利用でき，より的確にユーザの情報要求を掴める

これらは検索エンジンの性能を改善する上で重要である．

同義表現を考慮した検索ができない: 例えば，“ケーキ職人” と “パティシエ” はほぼ同じものを表す表現であるが，既存の検索エンジンでは別のものとして扱われている．そのため，“ケーキ職人” というクエリを使って（“ケーキ職人” を含まない）“パティシエ” を含む文書を検索することはできない．このような同義関係にある表現は検索時の漏れの原因となる．

我々は研究用途を主眼においた検索エンジン基盤 TSUBAKI [4] を，日本語ウェブ文書 1 億件を対象に開発・運用しており，検索エンジンの抱える上記の問題点の解決，検索性能の向上を目的に以下のことを行っている．

- 検索クエリを構文解析し，情報要求の的確な把握
- 構文解析されたクエリおよび検索対象となる文書の構造レベルでのマッチング
- 同義表現による検索漏れの改善

このため，単語だけでなく，係り受け，同義表現についても索引付けしている．そしてこれらを用いた検索を，128CPU コア，100TB を越えるストレージの上で実現している．

本稿では，NTCIR3-WEB で構築されたテストコレクションを利用し，TSUBAKI で用いている各インデックスの効果について述べる．また，自然文による検索クエリの扱いについても述べ，その効果も報告する．

2　開放型検索エンジン基盤 TSUBAKI

2.1　インデックスデータ

TSUBAKI では，2007 年 5 月から 7 月にかけて情報通信研究機構にてクロールされた日本語ウェブ文書 1 億件を，代表表記化単語，同義表現，およびそれぞれの間の係り受けの計 4 種類で索引付けしている．例えば，“パティシエのはたらく店” という句からは以下の索引が抽出される．

代表表記化単語: パティシエ，働く，店
単語間の係り受け: パティシエ→働く，働く→店
同義表現（単語，句）: <パティシエ>，<働く>，<店>
同義表現間の係り受け: <パティシエ>→<働く>，
<働く>→<店>

同義表現およびその索引付けは 3 節で述べる．本節では，単語・係り受けインデックスについて述べる．

2.1.1　代表表記化単語インデックス

情報検索において，「検索漏れ」は重要な問題である．日本語の場合は，平仮名，カタカナ，漢字が利用されるため，検索漏れが起こりやすい．TSUBAKI では JUMAN の解析結果に含まれる「代表表記」を用いて索引付けすることで表記揺れの問題に対処してい

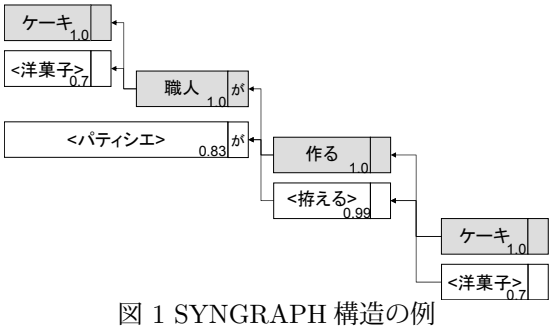

図 1 SYNGRAPH 構造の例

る．例えば，“とうがらし”，“トウガラシ”，“唐辛子” は代表表記を用いることで，“唐辛子” に集約されるため，検索漏れの影響の軽減が期待できる．

2.1.2 係り受けインデックス

クエリに適合する文書を高い精度で得るためには，単語の集合 (bag-of-words) で文書をモデル化しただけでは不十分である．以下の文はどちらも同じ単語から構成されるが，その構造が異なっている．

S1: 日本はドイツの車を輸入する
S2: ドイツは日本の車を輸入する

単純に単語を索引としただけでは，両者の違いを捕らえることができない．そこで構造を反映している係り受けに注目し，索引とする．これにより，両者が異なる文であることが検索時に考慮される．

TSUBAKI では，KNP の解析結果より係り受けインデックスを作成している．

2.2 文書スコアリング

TSUBAKI では，OKAPI BM25 [2] を利用して，クエリ Q と文書 D の関連度 $score_{rel}(Q,D)$ を求めている．具体的には，以下の式を用いている．

$$score_{rel}(Q,D) = \sum_{q \in Q} w_q \times BM25(q,D)$$

$$BM25(q,D) = w \times \frac{(k_1+1)F_{Dq}}{K+F_{Dq}} \times \frac{(k_3+1)F_{Qq}}{k_3+F_{Qq}}$$

$$w = \log \frac{N - DF_q + 0.5}{DF_q + 0.5}, K = k_1((1-b) + b\frac{l_D}{l_{ave}})$$

ここで q は検索クエリ Q から抽出された索引であり，単語または係り受けに該当する．F_{Dq} は，文書 D 中での q の出現頻度，F_{Qq} は Q 中での q の出現頻度，w_q は，Q の解析時に与えられる q の重み，N は検索対象となっている文書数（1.0×10^8），DF_q は q の文書頻度，l_D は D の文書長（単語数），l_{ave} は平均文書長である．w_q については 4 節で述べる．また，k_1,k_3,b は OKAPI のパラメータであり，$k_1 = 1$,$k_3 = 0$,$b = 0.6$ としている．

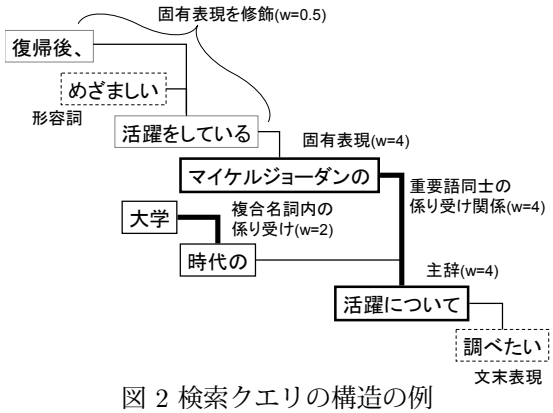

図 2 検索クエリの構造の例

3 同義表現インデックス

2.1.1 節で述べた代表表記化単語インデックスは平仮名・カタカナ・漢字の表記揺れは解消できるが，“ケーキ職人” と “パティシエ” のような表現の揺れは解消できない．そこで，国語辞典とウェブテキストから，表現間の同義関係，上位下位関係を自動獲得し，文書中に現れる表現と同義・上位下位関係にある表現も索引付けする．

単純に同義関係・上位下位関係にある表現を索引付けしたのでは組み合わせ爆発を起こすため，Shibata ら [3] の提案する SYNGRAPH データ構造の各ノードを索引とする．SYNGRAPH データ構造とは，同義グループに ID を与え，同義関係にある表現同士を効率よくまとめた（つまり，表現のずれを吸収した）データ構造である．図 1 は，「ケーキ職人が作るケーキ」を SYNGRAPH 構造に変換した例である．図中，影のついた表現は変換元の表現 (基本ノード)，白抜きの表現は同義関係および上位下位関係にある表現の集合に与えられた ID（SYN ノード）である．ID としては，<パティシエ>のように，<>で囲まれた表現が与えられる．また，表現の脇にある数字は，元の表現との類似度を表している．図は “ケーキ” の上位概念にある SYN ノードとして<洋菓子>が，“<洋菓子>職人” と <パティシエ>が同義関係にあることを示している．

TSUBAKI では基本ノード，SYN ノードおよびそれらの間の係り受けを索引としている．これにより，“パティシエ” で検索しても（“パティシエ” を含んでいない）“ケーキ職人” を含む文書を検索できる．

4 検索クエリの解析

先述したように，数個のキーワードによる検索要求の表現は不十分であり，より正確に表現するには，自然文の利用が必要である．しかしながら，NLQ の問題として，検索に不要な表現と重要な表現が玉石混合になっている点がある．図 2 は NLQ の例であるが，細

い線で囲まれた表現，および細い線で表わされた係り受け関係は，クエリに関連する文書を検索する上で不要と考えられる．一方で，太線で囲まれた表現や太線で示された係り受け関係は，重要であると考えられる．

そこで，検索クエリに対して，固有表現解析・構文解析を行った後，不要な索引の削除および索引の重みづけを行う．

不要な索引の削除: 検索クエリに含まれる以下の索引を削除する．

- 文末表現パターンにマッチする表現
- 検索クエリ中の全形容詞・動詞[1]

文末表現パターンとしては，“～を探したい” や “～について書かれた文書が欲しい” などを用いる．

索引の重み付け: 次の 4 種類の索引に対して，クエリと文書の適合度を計算する際に考慮する重みを与える．

- 固有表現に含まれる索引の重視:

  固有表現はクエリに適合する文書を検索する上で重要なため，固有表現内の索引 (図中，“マイケルジョーダン”) に対し，重要度を表す重み $\alpha(>1)$ を与える．さらに固有表現内の係り受けは，検索結果中の文書に必ず含まれるように検索条件を設定する．
- クエリの主辞に含まれる索引の重視:

  クエリの主辞は，クエリに適合する文書を検索する上で重要である．そこで，主辞に含まれる索引に対して重み $\beta(>1)$ を与える．図の例では，“活躍” に対して重み β が与えられる．
- 複合名詞内の係り受けの重視:

  “大学” から “時代” への係り受けを複合名詞 “大学時代” の近似と見なし，複合名詞内の係り受けに対して重み $\gamma(>1)$ を与える．
- 固有表現を修飾している索引の軽視:

  図 2 に示したクエリの場合，“復帰後活躍している” という句は，“マイケルジョーダン” に関する文書を検索する上で冗長である．そこで，固有表現を修飾する句に含まれる索引に対して，重み $\delta(<1)$ を与える．

重み α, β, γ , δ の値として，4, 4, 2, 0.5 を用いている．また，上記のどの場合にもあたらない索引については，重みを 1 とする．

[1]検索に重要な動詞もあるため全部削除することは多少乱暴である．不要な動詞の自動判別は今後の課題である．

検索キーワード: コンピューターウィルス, 予防, 対策
検索自然文: コンピューターウィルスの予防方法や対策法について説明している文章を探したい

図 3 検索クエリの例

5 評価実験

評価実験として，基本形，代表表記化単語，同義表現およびそれらの係り受けの効果を検証した．また，4 節で述べたクエリ解析処理の効果についても確認した．

5.1 テストコレクションと評価方法

検索性能の評価には，NTCIR3-WEB で構築されたテストコレクションを用いた．このコレクションには，“.jp” ドメインから収集された 1,000 万ページ，47 検索クエリ，各課題に対する文書の適合判定データが含まれている．検索課題は，図 3 に示すように，検索要求を表現した検索キーワード，および，要求を 1 文で述べた検索自然文 (NLQ) が含まれている．今回の評価実験では NLQ を用いた．

文書の適合性は，高適合，適合，部分適合，不適合の 4 段階が設けられており，適合性判定は，NTCIR3-WEB に参加したシステムが出力した文書に対して行われた．そのため，TSUBAKI が検索した文書集合には，未判定文書が含まれる可能性があることに注意されたい．本実験では，高適合，適合，部分適合と判定された文書を適合文書として扱っている．

評価尺度としては，上位 10 件における精度 (P@10)，非補間平均精度 (AveP)，R 精度 (R-repc.) を用いた．これらは NTCIR3-WEB でも用いられている尺度である．非補間平均精度，R 精度の算出には，検索結果の上位 1,000 件を用いた．しかし，検索クエリによっては，AND 検索時に 1,000 件得られない場合もある．そのような場合は，改めて OR 検索を行い，その結果を AND 検索の後に追加した．

5.1.1 検索自然文を用いた場合の検索性能

表 1 に，AND 検索，OR 検索した際の評価結果を示す．係り受けを考慮することで，全評価尺度において精度が向上していることがわかる．また，OR 検索の方が AND 検索に比べ，係り受けの効果が大きいことがわかる．

代表表記化単語インデックスにより，表記のゆれが吸収されるため，カバレッジの向上が期待できる．そこで，実際に得られた関連文書数を調査した．その結果，基本形インデックスが 2,249 文書，代表表記化単語インデックスが 2,291 文書であった．この結果から，代表表記化単語インデックスがカバレッジの向上に有効に働くことがわかる．また，P@10 で基本形インデッ

表 1 実験結果

モデル	AND 検索			OR 検索		
	P@10	AveP	R-prec.	P@10	AveP	R-prec.
BF	**0.304**	0.127	0.174	0.262	0.119	0.174
BF+D	0.302	**0.135**	**0.177**	**0.281**	0.131	0.182
RF	0.283	0.124	0.170	0.249	0.122	0.177
RF+D	0.285	0.129	0.174	0.275	**0.132**	**0.184**
SG	0.275	0.114	0.154	0.194	0.094	0.139
SG+D	0.279	0.126	0.159	0.253	0.114	0.169

(BF:基本形，RF:代表表記， SG:同義表現，D:係り受け)

表 2 クエリ解析処理の効果

処理	BF-D (AND)	BF-D (OR)
全処理有り	0.302	0.281
文末表現削除無し	0.234	0.272
重みの考慮無し	0.270	0.292
形容詞・動詞の削除無し	0.213	0.279
全処理無し	0.168	0.272

クスと比べた際，性能に差が見られるため，実際に出力された結果を調査した．その結果，適合文書として判断しても構わない文書が未判定になっていることが原因だった．この未判定文書の問題は，代表表記化単語インデックスだけでなく，同義表現インデックスにおいても発生している．

5.1.2 クエリ解析処理の効果

次にクエリ解析処理が，検索性能の向上にどの程度効果があるか調べた．具体的には，各処理をひとつずつ抜き，どの程度 P@10 の値が低下するか確認した．モデルには，前節の実験でもっとも P@10 が高かった BF+D（基本形 + 係り受け）を用いた．

実験結果を表 2 に示す．表より処理を抜くことで，多くの場合 P@10 のスコアが下がっていることから，クエリの解析処理が精度の向上に効いていることがわかる．OR 検索では文末表現削除以外の処理では P@10 の値の向上があまり見られないが，AND 検索ではすべての処理が精度の向上につながっている．もっとも効果的な処理は，形容詞・動詞の削除であり，P@10 の値が 0.089 減少している．

5.1.3 NTCIR3-WEB 参加システムとの比較

最後に NTCIR3-WEB においてもっとも性能の高かった GRACE [5] と BF+D(AND) の比較を行った．GRACE の特徴としては，擬似適合性フィードバック (pseudo-relevance feedback, 以下 PRF) を行っている点が挙げられる．比較実験は，キーワード検索，自然文検索の両方について行い，高適合，適合と判断される文書のみを適合文書として評価した．

表 3 に，キーワード検索，自然文検索における，BF+D(AND) および GRACE の性能 (AveP, P@10) を示す．表より，キーワード検索，自然文検索のどちらにおいても，GRACE（PRF 有）の性能が高いことがわかる．その一方で，GRACE(PRF 無) との性能を比較すると，キーワード検索においては BF+D(AND) の方が高く，自然文検索では GRACE の方が高い．この結果から，評価指標により若干性能が上下するが，BF+D(AND) の性能は従来のシステムと比較して同程度であることがわかる．そのため，本評価実験で行った検討は，十分に意味のあるスコアレベルで行われていると言える．

表 3 NTCIR3-WEB でもっとも性能の高かったシステム (GRACE) との比較

モデル	P@10	AveP
GRACE （キーワード検索，PRF 有）	0.221	0.151
GRACE （キーワード検索）	0.181	0.121
BF+D(AND) （キーワード検索）	0.206	0.111
GRACE （自然文検索，PRF 有）	0.234	0.155
GRACE （自然文検索）	0.209	0.132
BF+D(AND) （自然文検索）	0.187	0.091

6 おわりに

本稿では，TSUBAKI で用いてる各種インデックス（単語，同義表現およびそれぞれの間の係り受け）の効果を NTCIR3-WEB テストコレクションを用いて評価した．その結果，係り受け関係を利用することで，単語，同義表現単体で検索するよりも，検索精度が向上することが確認された．また，自然文で与えられた検索クエリを解析し，クエリ中の固有表現や主辞を重視したり，クエリに含まれる形容詞・動詞を削除することで，検索精度が向上することも確認できた．

参考文献

[1] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau. Okapi at TREC. In *Text REtrieval Conference*, pages 21–30, 1992.

[3] Tomohide Shibata, Michitaka Odani, Jun Harashima, Takashi Oonishi, and Sadao Kurohashi. SYNGRAPH: A flexible matching method based on synonymous expression extraction from an ordinary dictionary and a web corpus. In *Proc. of IJCNLP2008*, pages 787–792, 2008.

[4] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. TSUBAKI: An open search engine infrastructure for developing new information access methodology. In *Proc. of IJCNLP2008*, pages 189–196, 2008.

[5] Masashi Toyoda, Masaru Kitsuregawa, Hiroko Mano, Hideo Itoh, and Yasushi Ogawa. University of tokyo/ricoh at ntcir-3 web retrieval task. In *Proceedings of the 3rd NTCIR Workshop Meeting*, pages 31–38, 2002.