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

2009－ＦＩ－ 93（17）
2009－ＮＬ－189（17）
2009／1／23

クエリの語句の重要度と係り受けを考慮した自然文検索

新里 圭司　黒橋 禎夫
京都大学 大学院情報学研究科
〒606-8501 京都府京都市左京区吉田本町
{shinzato,kuro}@i.kyoto-u.ac.jp

あらまし

本稿では自然言語で表現されたクエリに含まれる単語，係り受け関係を，重要度に応じて使い分けて文書検索を行う方法を提案する．重要度としては，(1) 検索結果中の文書に必ず含まれなければならない (必須)，(2) 含まれている方が好ましい (任意)，(3) 含まれていなくても良い (不要) の 3 段階を設け，単語，係り受け関係をいずれかに分類する．提案手法を NTCIR-3,4 で構築されたテストセットを利用して評価した結果，固有表現および一部の複合名詞内の係り受け関係については，「必須」として扱うことで検索性能が改善されることがわかった．また，名詞から意味が推測される動詞を「任意」として扱うことで，全動詞を検索に用いたり，削除したりする方法よりも高い性能が達成できることがわかった．

キーワード 自然文クエリ, 語・句の重要度, 係り受け関係

Natural Language Search Based on Term Importance and Dependency Relations

Keiji Shinzato　Sadao Kurohashi
Graduate School of Informatics, Kyoto University
Yoshida-honmachi Sakyo-ku, Kyoto 606-8501, JAPAN
{shinzato,kuro}@i.kyoto-u.ac.jp

Abstract

This paper describes the method that retrieves documents by switching words and dependency relations in a natural language query according to their importance. We defined three levels as term importance. The first is “required”, terms that must be in retrieved documents. The second is “optional”, terms that are preferable to be included in the documents. The third is “unnecessary”, terms that are not matter whether they are included in the documents. Words and dependency relations in a query are classified into any one of these levels. We evaluated the contribution of our approach using the NTCIR-3 and NTCIR-4 test collections. As a result, the search performance was improved by regarding dependency relations in named entities and some compound nouns as the required terms. The performance was also improved by regarding verbs whose meaning is inferred from nouns as optional terms.

Key Words natural language query, term importance, dependency relation

1 はじめに

現在の検索エンジンは，検索要求を 2，3 のキーワードにより表現するキーワード検索が主流であり，TREC などの検索システムの評価を行うワークショップにおいてもキーワード検索に焦点がおかれてきた．しかしながら，検索要求が複雑になるほど，ユーザの検索要求を数個のキーワードで表現することは難しくなるため，キーワード検索には限界があると考えられる．また，近年の音声インターフェースの高度化を鑑みると，音声インターフェースを介して検索エンジンにアクセスすることも近い将来実現すると考えられる．音声インターフェースを介した検索では，人間が検索要求を表現するのに適した自然言語で表現されたクエリの入力が想定される．以上のことを考えあわせると，検索要求が自然言語で表現された自然文クエリによる情報検索技術は重要である．

自然文クエリのもう一つの利点は，クエリに含まれる語の係り受け関係を検索に利用できることである．これらの関係を利用することでより高い精度でクエリと関連する文書を検索することが期待できる．

自然文検索を考えたとき，既存のウェブ検索エンジンのように AND 検索を行ったのでは用いる表現が多いため，検索結果が得られない，もしくは得られる文書が極端に減ってしまうという問題が起こる．一方で，TREC などで研究されているシステムでは，語の重みと OR 検索により文書検索を行っており，これらのシステムでは，語の重みによりクエリ中の主要な表現とそれ以外の表現を区別しようとしている．しかしながら，自然文クエリのように複数の主要な表現が含まれると，それらの組み合わせによりクエリとは関係ない文書がクエリと適合する文書よりも関連度が高くなるという問題が生じる．

そこで本手法では，語句の重要度を重みづけにより表すのではなく，語句が検索結果中の文書に含まれていなくてはならない（必須），含まれていた方が好ましい（任意），含まれていなくても構わない（不要）の 3 段階で表わし，重要度ごとに検索での利用方法を変化させることで，既存の AND 検索や重みベースの手法よりも高い検索性能を実現する．

2 関連研究

クエリ中の語句の重要度を明確に設定する研究としては，Allan ら [1] の研究がある．Allan らは「クエリ内には，クエリと適合する文書に含まれていなくてはならない語が存在する」という仮定のもと，検索結果中の文書に含まれていなくてはならない表現（論文では *core term* と呼ばれている）を，語の出現頻度を利用して自然文クエリから抽出している．しかしながら，Allan らは，文書検索の一つの手がかりとして *core term* を用いているだけに留まっており，実際に *core term* 単独で検索は行っていない．本研究では，*core term* に相当する表現を使って検索を行い，検索結果が十分得られない場合は，検索条件を徐々に緩めることで検索結果を増やすという手法をとっており Allan らとは異なる．

自然文クエリから重要語・句を抽出する方法としては，固有表現，複合名詞に注目する手法 [3]，機械学習に基づく手法 [2] などがある．Callan らはクエリから固有表現，複合名詞を抽出し，それらについてフレーズ検索を行うことで検索性能の改善を行っている．本手法では，複合名詞を構成する語同士のつながりの強さに着目し，フレーズ検索として扱った方が良い複合名詞とそれ以外の複合名詞を区別しており，この点で上記の手法と異なる．また，Bendersky らは機械学習により自然文クエリから重要な名詞句を抽出する手法を提案しており，抽出された名詞句の重みを大きくすることで検索性能が向上した報告している．我々の手法は学習データを用いずに，固有表現解析結果および名詞と名詞の共起頻度情報をもとに重要語・句を抽出しており，この点で Bendersky らの手法と異なっている．

現在は自然文による問い合わせが可能な検索エンジンも公開されており，有名なものとしては Poerset[1]がある．Powerset では，“What did steve jobs say about the iPod?” のような自然文クエリを用いた検索が可能であるが，検索対象は英語版 wikipedia に限られており，自然文を用いたウェブ検索を行うことはできない．

3 係り受け関係の利用

自然文クエリの利点として，クエリ内の語と語の係り受け関係を利用できるという点があり，係り受け関係を利用することでより高精度な検索が期待できる．本研究で用いる係り受け関係とは，内容語と内容語の 2 項関係であり，格助詞は考慮しない．また複合辞は直前の内容語に連結される．図 1 はク

[1] http://www.powerset.com/

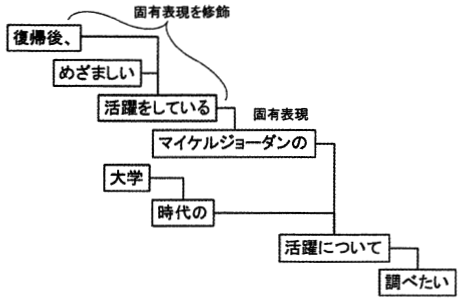

図 1: 検索クエリの構文解析結果

エリ「復帰後，めざましい活躍をしているマイケルジョーダンの大学時代の活躍について調べたい」を構文解析した結果であるが，(マイケルジョーダン，活躍)，(大学，時代) などが係り受け関係として抽出され，検索に利用される．

クエリと文書の関連度の計算にはOkapi BM25 [8]を利用する．多くの場合，Okapi BM25 はクエリ中の単語と文書の関連度を計算するために用いられるが，ここでは係り受け関係についても関連度を計算できるよう拡張する．$T_{q_{word}}$ をクエリ q から抽出された単語の集合，$T_{q_{dpnd}}$ を係り受けの集合としたとき，文書 d とクエリ q の関連度を以下の式で求める．

$$R(q,d) = (1-\beta) \sum_{t \in T_{q_{word}}} BM_{25}(t,d) + \beta \sum_{t \in T_{q_{dpnd}}} BM_{25}(t,d)$$

ここで β はスコアリングに係り受け関係をどの程度利用するかを調節するパラメータである．$BM_{25}(t,d)$ は以下の式で定義される．

$$BM_{25}(t,d) = \frac{(k_1+1)F_{dt}}{K+F_{dt}} \times \frac{(k_3+1)F_{qt}}{k_3+F_{qt}}$$
$$w = \log \frac{N-n+0.5}{n+0.5}, K = k_1((1-b)+b\frac{l_d}{l_{ave}})$$

F_{dt} は文書 d 中での t の出現頻度，F_{qt} は q 中での t の出現頻度，N は検索対象となっている文書数，n 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．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．

4 語句の重要度

本節では自然文クエリ中の語句の重要度について述べる．TREC などで従来より研究されてきたシステムは，*tfidf* などの統計量にしたがって求められた語の重みにより，クエリの主題に関連する表現と関連しない表現の区別を試みている．しかしながら，このような重みに基づく方法では，自然文クエリのように，複数の主要な表現が含まれたクエリの場合，主要な表現同士の相互作用により，クエリと全く関係ない文書がクエリと適合する文書より検索結果の上位にランクされるという問題が発生する．これは，クエリの主題と関連する表現と，関連しない表現の間に明確な線引きがなされていないためであると考えられる．

そこで本手法では，語句の重要度として以下の 3 段階を設け，検索での利用方法を区別する．

必須: 文書に含まれていなくてならない．さらに，必須の全表現が W 語以内という近接条件を課すことも考えられる

任意: 文書に含まれていた方が良い

不要: 文書に含まれていても，いなくてもどちらでも良い

先述したように，本研究では単語に加え係り受け関係を検索に用いている．係り受け関係も考慮した AND 検索を考えることができるが，それでは多くの文書を得ることができない．そのため，基本的には**単語は必須，係り受け関係は任意**と見なす，しかし後述する語句の重要度判定処理により，必要に応じて単語，係り受けの重要度は変更される．今後は，「必須」「任意」「不要」に分類された単語，係り受けを，必須表現，任意表現，不要表現と呼ぶ．検索には不要表現は一切利用せず，文書の収集は必須表現のみを，文書のスコアリングは必須表現，任意表現を利用して行う．

5 重要度の判定処理

重要度判定処理により，固有表現と強連結複合名詞，固有表現の修飾句，名詞からその意味が推測される動詞，クエリ末尾表現と機能的表現のいずれかに該当する単語，係り受け関係は重要度が変更される．以下では，ここに挙げた表現の認識方法と，その重要度について述べる．

5.1 固有表現と強連結複合名詞

クエリ中に固有表現が含まれている場合，ユーザが期待することは，固有表現がそのまま現われている文書が検索されることである．例えば「日本銀行」が含まれていた場合，「日本の銀行」のように「日

本」と「銀行」が離れて現われている文書ではなく，「日本銀行」がそのまま現われている文書をユーザは求めていると考えられる．そのため本研究では，固有表現内の係り受け関係を任意表現から必須表現へ変更する．これにより，固有表現をそのまま含んでいる文書を検索することが可能になる．

固有表現の他にも，そのまま検索結果に含まれていてほしい表現が存在する．その例として，複合名詞「赤ちゃんポスト」が考えられる．「赤ちゃんポスト」がクエリに含まれていた場合も，先程の「日本銀行」の場合と同様に，ユーザは「赤ちゃん」と「ポスト」が離れて現われている文書ではなく，「赤ちゃんポスト」がそのまま出現している文書を求めていると考えられる．本研究では「赤ちゃんポスト」のような複合名詞を，**強連結複合名詞**と呼ぶ．一方で「京都旅行」のように，「京都旅行」を含んでいなくても，「京都」と「旅行」を含んでいる文書が得られればユーザが満足すると思われる複合名詞も存在する．このため，「赤ちゃんポスト」と「京都観光」を強連結複合名詞と，普通の複合名詞にわける基準が必要となる．

以上より問題は，$|N|$ 語から構成される複合名詞 N $(n_1, ..., n_{|N|})$ のつながりの強さをどのように求めるかである．ここでは，n_i と n_{i+1} のつながりが強ければ，「n_i の n_{i+1}」という形で文書に出現しにくいだろうという仮説に基づき，つながりの強さを以下のスコアで求める．

$$Score_{strength}(N) = \frac{1}{|N|-1} \sum_{i=1}^{|N|-1} \frac{DF(n_i \text{の } n_{i+1})}{DF(n_i n_{i+1})}$$

ここで $DF(X)$ は日本語ウェブページ一億件から求めた表現 X の文書頻度を表す．本手法では，スコア $Score_{strength}(N)$ が閾値 T_p を超える場合，N をつながりが強い複合名詞と見なす．T_p の値は経験的に求めた 300 を用いている．

この手法により，つながりが強いと判定された複合名詞内の係り受け関係は，任意表現から必須表現へと変更する．

5.2 固有表現の修飾句

修飾句は被修飾語の意味を限定するために用いられるが，固有表現の場合，すでにある 1 つのエンティティに限定されるため修飾句によりその意味を限定する必要はない．このため固有表現の修飾句は無くても問題ないと考えられる．そこで本手法では，

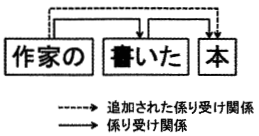

図 2: 新しく追加された係り受け関係

固有表現の修飾句内の語を必須表現ではなく任意表現として扱う．先程示した図 1 においては，「マイケルジョーダン」を修飾している句「復帰後，めざましい活躍をしている」内の各単語が任意表現として見なされる．

固有表現の修飾句の認識は，クエリの構文解析結果，固有表現解析結果を利用して行う．

5.3 名詞からその意味が推測される動詞

本手法では，名詞からその意味が推測される動詞を任意表現として見なす．例えば，表現「作家の書いた本」は，しばしば「作家の本」として「書く」が省略されることがある．しかしながら，「書く」が省略された状態でもその意味は「作家の書いた本」と同じである．これは動詞「書く」の意味を名詞「作家」から推測できるためと考えられる．本手法では，このような動詞は検索結果中に含まれていなくても構わないと考え任意表現として扱う．

このような名詞と動詞の関係を大規模ウェブコーパスより自動獲得する．獲得に用いるアイデアは，名詞 n と動詞 v が係り受け関係を持ちやすいならば，名詞 n から動詞 v の意味は推測されやすいだろうというものである．

名詞 n と動詞 v の係り受け関係の持ちやすさは，以下のスコアにより求める．

$$Score_{cooc}(n, v) = P(n, v) \cdot \log_2 \frac{P(n, v)}{P(n) \cdot P(v)}$$

ここで $P(\cdot)$ はウェブから収集した日本語 16 億文から求めた，名詞，動詞，名詞から動詞への係り受け関係の出現確率である．$Score_{cooc}(n, v)$ の値が閾値 T_v を越える場合，動詞 v の意味は名詞 n から推測できると考える．閾値 T_v の値としては，経験的に求めた 1×10^{-6} を用いている．自然文クエリ中の動詞 v と係り受け関係にある名詞のうち，ひとつでも上記のスコアが閾値を越える場合，動詞 v を任意表現としてみなす．

動詞 v が任意表現として見なされた場合，動詞 v の係り元と係り先の間に新しい係り受け関係を追加

クエリ末尾表現パターン: <形容>*<説明述語>* <文書>*について*<欲求述語>;

<形容>: (詳しい|詳細だ);

<説明述語>: (説明|書く|記述|記載|記す|述べる)(する)* (いる|ある|れる|られる)*;

<文書>: (ウェブ|WEB)* (文書|ページ|HP|情報|文章|テキスト);

<欲求述語>: (知る|探す|調べる|見る|見つける|読む)(たい|いる);

図中のA|Bは「AまたはB」を表す．また，*は直前の表現が省略可能であることを表す．

図 3: クエリ末尾表現パターン

する．図2は「作家の書いた本」の構文解析結果に新しく係り受け関係を追加した例である．これにより,「作家の本」を含む文書であっても高いスコアを得ることが可能となり，クエリと文書の間の表現の揺れを吸収することが期待できる．

5.4 クエリ末尾表現および機能的表現

自然文クエリの末尾には,「〜を知りたい」「〜文書を探したい」のような表現が現れやすい．これらの表現は検索とは関係ないため，あらかじめクエリ末尾表現パターンを準備し，パターンにマッチした表現内の単語，係り受けは不要表現として扱う．図3にクエリ末尾パターンを示す．パターン中の各単語の表記のゆれは考慮されており,「知りたい」「しりたい」のどちらでもマッチする．

また自立語であっても機能的である表現は検索に不要と考えられる．このような機能的表現を削除するため，不要語リストを整備し，リストに登録されている語，および登録されている語を含む係り受け関係を不要表現として見なす．不要語リストには，形式名詞，指示詞，疑問詞，副詞に加え,「ある」「なる」「使う」が登録されており，クエリ末尾パターン同様に表記のゆれが考慮される．

6 評価実験

6.1 実験設定

NTCIR-3, 4 [4,5] で構築されたテストセットを利用して評価実験を行った．NTCIR-3,4 では同じ文書セット（jp ドメインから収集された 11,038,720 文書）に対して検索課題が用意されており，課題数はNTCIR-3が47，NTCIR-4が80である．検索課題の例を図4に示す．実験には<DESC>タグで囲まれた自然文で表現された検索クエリを利用した．

```
<TOPIC>
<NUM>0008</NUM>
<TITLE CASE="b">サルサ，学ぶ，方法</TITLE>
<DESC>サルサを踊れるようになる方法が知りたい</DESC>
:
</TOPIC>
```

図 4: NTCIR-3, 4 で定義される検索課題の例

検索は文書の収集と，スコアリングの2段階からなる．文書の収集には必須表現のみを用い，文書のスコアリングには必須表現，任意表現の両方を利用する．収集およびスコアリングのどちらの場合についても，国語辞典とウェブ文書から自動獲得した同義語 [9] を考慮する．また，必須表現のウィンドウサイズは予備実験で最も性能が良かった 75 単語を利用する．

各検索課題には,「高適合」「適合」「部分的適合」「不適合」の4段階で評価された文書が提供されており，本実験では「高適合」「適合」「部分的適合」を正解,「不適合」を不正解とみなした．評価尺度としては，MRR, P@10, R-prec, MAP の4種類のを用いた[2]．各尺度のスコアは，検索結果から未判定文書を除いてから計算した．MRR の値は検索結果の上位 10 件を対象に計算した．また，R-prec および MAP の値を計算するため検索結果は上位 1000 件を出力した．1000 件得られない場合は，得られるまで検索条件を緩めた．検索条件の緩め方については，次節で述べる．

クエリの解析には JUMAN [7] および論文 [10] で提案されている固有表現認識手法が組み込まれた KNP [6] を利用した．

6.2 提案手法と他の手法との比較

重要度に従って自然文クエリ内の語句を検索で明確に使い分けることにより，どの程度精度が向上するかをみるため，以下の手法と比較実験を行った (表1)．なおこれら全ての手法で，5.4 節で述べた不要表現は削除されている．

OR 検索: 単語のみを利用した OR 検索．表1中で単語の列が△になっているが，これは単語を

[2] MRR 以外の値は，http://trec.nist.govtrec_eval で配布されている評価ツールを利用した．

表 1: 提案手法と他の手法の比較

モデル名	OR 検索		AND 検索		AND 検索 (近接を考慮)		AND → OR 検索		AND → OR 検索 (係り受け有)			提案手法					
検索に用いる表現と近接	近接	単語	近接	単語	近接	単語	近接	単語	近接	単語	係り受け	近接	単語 NE・強連結複合名詞	単語 普通	単語 冗長動詞・NE の修飾句	係り受け NE・強連結複合名詞内	係り受け 通常
パラメータ	無	△	無	○	有	○	有	○	有	○	△	有	○	○	△	○	△
												有	○	○	△	△	△
					無	○	無	○	無	○	△	無	○	○	△	△	△
												無	○	△	△	△	△
							無	△	無	△	△	無	△	△	△	△	△
MRR	0.400		0.536		0.531		0.518		0.544			**0.560**					
P@10	0.294		0.342		0.338		0.355		0.348			**0.387**					
R-prec	0.210		0.167		0.163		0.224		0.219			**0.228**					
MAP	0.151		0.115		0.115		0.177		0.179			**0.186**					

○：必須表現として利用する，△：任意表現として利用する

任意表現として扱うことを意味している．

AND 検索: 単語のみを利用した AND 検索．検索結果が 1000 件に達しなくても何もしない．表中の単語の列が○になっているが，これは単語を必須表現として扱うことを意味している．

AND 検索（近接を考慮）: 単語が 75 単語以内 (出現順序を考慮しない) に現われていなければならないという近接条件のもとで文書を検索する．この「75」という数値は，予備実験において最も性能の高かった値である．検索結果が 1000 件に達しない場合は，近接条件を外し，AND 検索を行う．表 1 のパラメータの欄は，このように検索条件を緩める過程を表している．

AND → OR 検索: 上記の AND 検索（近接を考慮）の後，検索結果が 1000 件に達しない場合は OR 検索を行う

AND → OR 検索（係り受け関係有）: 上記の AND → OR 検索を行う際，文書のスコアリングに係り受け関係を利用する．

上記の比較手法に対して提案手法は，固有表現 (NE)，強連結複合名詞内の係り受け関係を必須，それ以外の係り受け関係を任意，名詞から意味が推測される動詞および NE の修飾句内の単語を任意，それ以外の単語を必須として扱う．さらに必須に属す単語が 75 単語以内に現われていなければならないという近接条件を課す（表 1 の提案手法のパラメータの一番上の欄）．この条件で検索結果 1000 件得られない場合は，表 1 に示した手順で検索条件を緩める．

表 1 より既存の検索エンジンで行われている AND 検索や，Okapi などの尺度により語句の重要度を決定し OR 検索する従来手法よりも提案手法の方が性能が良いことがわかる．また AND → OR 検索 (係り受け有) に比べ提案手法の方がどの評価尺度においても性能が高い．このことから，名詞から意味が推定される動詞や，NE の修飾句を任意表現として扱ったり，NE，強連結複合名詞内の係り受け関係を必須表現として扱うことが検索において有効であることがわかる．

OR 検索と AND 検索を比べると，MRR,P@10 においては AND 検索の方がスコアが高いことがわかる．一方，AND 検索は全ての必須表現を含む文書でないと検索されないため，多くの検索結果を得ることができず，R-prec, MAP の値が低くなっていることがわかる．

表 1 の提案手法のパラメータは最適値となっており，残りの実験ではパラメータの値を変化させた時，検索性能がどのように変化するかを調べている．

6.3 文書スコアリングにおける係り受けの効果

文書をスコアリングする際，係り受け関係をどの程度重視すればよいかを調べるために，重み β の値を変化させる実験を行った．β の値を 0 から 1 まで 0.1 刻みで変化させた時の各値における評価尺度の値を表 2 に示す．表中の β の値が 0 の時はスコアリングに係り受け関係を考慮しない場合，1 の時は係り受け関係のみを使ってスコアリングを行った場合の検索性能をそれぞれ示している．β の値が 0.0 から 0.5 の範囲に各尺度の最大値があることから，係り受け関係から計算されるスコアを単語よりも少し重みを下げて利用した方がいいことがわかる．この結果から，他の実験においては $\beta = 0.2$ を用いている．

表 2: 文書スコアリングにおける係り受けの重み β と検索性能

β	MRR	P@10	R-prec	MAP
0.0	0.549	0.380	**0.230**	0.184
0.1	0.554	0.377	0.225	0.182
0.2	0.560	**0.384**	0.224	0.183
0.3	0.553	0.381	0.222	0.183
0.4	0.555	0.377	0.221	0.184
0.5	**0.571**	0.376	0.224	**0.186**
0.6	0.560	0.371	0.222	0.182
0.7	0.561	0.365	0.219	0.179
0.8	0.547	0.359	0.215	0.173
0.9	0.538	0.355	0.210	0.169
1.0	0.516	0.334	0.191	0.148

6.4 固有表現の修飾句の扱いと検索性能

固有表現の修飾句を必須表現，任意表現，不要表現として用いた場合について検索性能を調査した．実験結果を表 3 に示す．表より修飾句に含まれる語，係り受け関係は必須表現として用いない方が検索性能が良いことがわかる．

6.5 動詞の扱いと検索性能

次に動詞を名詞から意味が推測される語とそうでない語にわけ，それぞれを必須表現，任意表現，不要表現として用いた場合について検索性能を調査した．実験結果を表 4 に示す．提案手法では名詞から推測可能な動詞は任意，それ以外の動詞は必須として，動詞を区別して検索を行っているが（表中の推測不可：○，推測可：△），こうすることで，全動詞を検索に用いたり（推測不可：○，推測可：○），動詞を検索に全く用いない（推測不可：×，推測可：×）場合よりも高い性能が得られることが表よりわかる．推測可能な動詞を任意表現として扱うことで検索がうまくいった例として，クエリ「ブルーベリーなどに含まれている成分，アントシアニンの効果について詳しく知りたい」が挙げられる．5.3 節の処理により，「成分」から「含む」の意味が推測可能と判断されたため，「ブルーベリーの主成分であるアントシアニンは...」と書かれた「含む」が現われていない文書を検索することができていた．

しかしながら，今回の実験で最も高い性能を達成したのは，全ての動詞を任意表現として扱う（推測不可：△，推測可：△）方法であった．全ての動詞を任意として見なすことで，適切な文書が検索できるようになったクエリとしては「コペルニクスの地動説がキリスト教社会でどのように受容されていたかを調べたい。」がある．提案手法では，「社会」から「受容」の意味が推測されないと判定されていた．これは正しい動作であるが，一般に動詞には多数の同義語があるため，同じ事柄でも複数の言い回しを考えることができる．そのため，動詞を必須として扱ってしまうことは，他の言い回しで表現された文書が検索されなくなることを意味しており，これは検索性能の低下の原因となりえる．このような必須の副作用については，今後は，動詞の同義語を整備することで対応していく予定である．

表 3: 固有表現の修飾句の扱い方と検索性能

修飾句の扱い	MRR	P@10	R-prec	MAP
○	0.560	0.379	**0.229**	**0.186**
△	0.560	**0.387**	0.228	**0.186**
×	**0.561**	0.386	0.227	0.185

表 4: 動詞の扱い方と検索性能

動詞		MRR	P@10	R-prec	MAP
推測不可	推測可				
○	○	0.530	0.360	0.221	0.177
○	△	0.560	0.387	0.228	0.186
○	×	0.560	0.381	0.223	0.182
△	△	**0.561**	**0.389**	**0.231**	**0.190**
△	×	0.560	0.385	**0.231**	0.189
×	×	0.556	0.383	0.230	0.188

6.6 係り受け関係の扱いと検索性能

最後に固有表現，強連結複合名詞内の係り受け関係を必須表現，任意表現，不要表現として用いた場合について検索性能を調査した．

実験の結果を表 5 に示す．表より係り受けのタイプを区別することなく，単純に全ての係り受け関係を必須，任意，不要と用いる場合は，どの評価尺度においてもそれほど大きな差がないことがわかる．一方で，提案手法では，NE・強連結複合名詞内の係り受け関係を必須，それ以外の係り受けを任意とすることで，どの尺度においても最高性能を示していることがわかる．この結果から，係り受け関係を検索で利用する際は，単純に全部用いるのではなく，タイプごとに利用方法を変える必要があることがわかる．

表 5: 係り受け関係の扱い方と検索性能

係り受け関係		MRR	P@10	R-prec	MAP
NE・強連結 複合名詞以外	NE・強連結 複合名詞内				
○	○	0.537	0.362	0.221	0.181
△	○	**0.560**	**0.387**	**0.228**	**0.186**
×	○	0.523	0.378	0.219	0.179
△	△	0.526	0.377	0.221	0.181
×	△	0.518	0.376	0.221	0.180
×	×	0.515	0.375	0.222	0.179

7 おわりに

本稿では自然言語で表現されたクエリに含まれる語句の重要度の判定方法および，重要度に応じた語句の検索での利用方法について述べた．本手法では語句の重要度として，「必須」「任意」「不要」の 3 段階を設け，クエリ内の語句をいずれかに分類する．「不要」と判定される表現は，クエリの末尾に現われる「〜について知りたい」や「〜という文書を探している」などであり，「任意」と見なされる表現は，固有表現の修飾句および名詞からその意味が推測される動詞（例えば，「影響」からみた「与える」）などである．「必須」と判断される表現には，固有表現およびつながりの強い複合名詞がある．語句の重要度の判定は，おもに大規模ウェブコーパスから得られる語と語の共起情報および固有表現解析結果をもとに行う．

評価実験は NTCIR-3,4 で構築されたテストセットを利用した．その結果，固有表現およびつながりの強い複合名詞内の語句については，必須として扱うことで検索性能が改善されることがわかった．また，名詞から意味が推測される動詞を「任意」として扱うことで，全動詞を検索に用いたり，削除したりする方法よりも高い性能が達成できることがわかった．

今後の課題として，今回は名詞から意味が推測される動詞を検出する手法について述べたが，このような表現は名詞にも存在する．そのため，動詞以外の表現について検出する方法に取り組む予定である．

参考文献

[1] James Allan, Jamie Callan, W. Bruce Croft, Lisa Ballesteros, John Broglio, Jinxi Xu, and Hongmin Shu. Inquery at trec-5. In *NIST*, pp. 119–132, 1997.

[2] Michael Bendersky and W. Bruce Croft. Discovering key concepts in verbose queries query. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2008*, pp. 491–498, 2008.

[3] James P. Callan, W. Bruce Croft, and John Broglio. Trec and tipster experiments with inquery. *Inf. Process. Manage.*, Vol. 31, No. 3, pp. 327–343, 1995.

[4] Koji Eguchi, Keizo Oyama, Akiko Aizawa, and Haruko Ishikawa. Overview of web task at the fourth ntcir workshop. In *Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization*, 2004.

[5] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama. The web retrieval task and its evaluation in the third ntcir workshop. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[6] Sadao Kurohashi and Makoto Nagao. A syntactic analysis method of long japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, Vol. 20, No. 4, pp. 507–534, 1994.

[7] Sadao Kurohashi, Toshihisa Nakamura, Yuji Matsumoto, and Makoto Nagao. Improvements of japanese morphological analyzer juman. In *The International Workshop on Sharable Natural Language*, pp. 22 – 28, 1994.

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

[9] 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*, pp. 787–792, 2008.

[10] 笹野遼平, 黒橋禎夫. 大域的情報を用いた日本語固有表現認識. 情報処理学会論文誌, Vol. 49, No. 11, pp. 3765–3776, 2008.