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

2008－DBS－145 (6)
2008－ＦＩ－ 91 (6)
2008／6／19

自然言語文からの構造化クエリの自動生成による Web情報検索

柴田 鉄也† 江口 浩二†

†神戸大学大学院工学研究科情報知能学専攻
〒657-8501 神戸市灘区六甲台町1-1

あらまし 本稿では，日本語の自然言語文クエリを解析し，その結果をもとにクエリを構造化する手法を提案する．自然言語文クエリの解析には，形態素解析と係り受け解析を用いる．そして，得られる文節や係り受け関係などの情報をもとに文を構造化し，その構造をもとにクエリを生成する．このように構造化されたクエリを用いて検索を行うことで自然言語文に適用し，高効率を維持しつつ高精度な検索が可能となる．提案手法を評価するため，主に日本語で記述された100GBのwebテキストコレクションを用いて，文節や係り受け関係などの情報を用いた場合とそうでない場合を比較した．その結果，検索精度が約8.5%向上した．さらに提案手法を擬似適合フィードバックと組み合わせたところ，約21.3%向上した．

Automatic Query Structuring from Sentences for Japanese Web Retrieval

TETSUYA SHIBATA† KOJI EGUCHI†

†Kobe University
Department of Computer Science and Systems Engineering
1-1 Rokkoudai, Nada-ku, Kobe 657-8501 , Japan

Abstract This paper proposes a query structuring method by analyzing Japanese natural language sentences that users input. We use linguistic structure information obtained by morphological analysis and dependency parsing on the natural language inputs. Our proposed method converts a sentence into a structured query on the basis of linguistic structures such as phrases and modification relations. Such structured queries enable effective retrieval with reasonable efficiency. To evaluate our proposed method, we compare it with the method using no structures at all, using a 100-gibabyte web collection mostly written in Japanese. We demonstrate through the experiments that mean average precision was improved about 8.5% using our query structuring method alone, and about 21.3% by combining our query structuring method with pseudo-relevance feedback.

1 はじめに

現在，情報検索クエリについて多くの研究が行われている．この中でも様々な研究課題があり，ひとつはクエリを拡張するような研究である [1, 2, 3]．また，クエリを自動的に構造化する研究 [4, 5] があり，本研究ではこれに位置する．さらに，クエリの性能を評価する手法 [6, 7] なども研究されている．一般的に情報検索に用いられるクエリは，複数の語の集合で表わされる．その一方で，自然言語クエリによる検索の需要が存在する．しかし日本語の自然言語文をクエリとして情報検索を行うには大きな問題点がある．日本語の自然言語文がクエリとして与えられた場合，英語などのように空白で区切られていないので，文を単語単位に分解してクエリを生成する必要がある．その際，単に分解された単語の集合をクエリとして用いるのでは，検索に有効であるとはいえない．これはクエリとして与えられた文の中にある単語すべてがユーザーの情報ニーズを的確に表しているとは限らないからである．さらに単語の数が大きくなることによって計算量も増大してしまう．これらの問題点を解決するために，自然言語文を解析し，その結果を用いてクエリを構造化する必要がある．本研究では，形態素解析器と係り受け解析器を用いて文を解析し，得られた文節や係り受け関係などの情報をもとに文を構造化する．そして，その構造に基づいてクエリを構造化する．そのようにクエリを構造化することで，語間の関係性を考慮したクエリを生成し，さらに必要のない語を除くことができる．このように構造化されたクエリを用いて検索を行うことで，自然言語クエリによる情報検索の高効率を維持しつつ高精度な検索が可能となる．提案手法を評価するために，文節や係り受け関係などのような解析情報を用いた場合とそうでない場合と比較する．それによって文の解析情報が，日本語の自然言語クエリを用いた検索に有効であることを示す．また，提案手法によって生成されたクエリで検索された文書集合を用いて，擬似適合フィードバックを行った場合の有効性についても評価する．

2 関連研究

現在では，クエリ尤度モデル [8, 9, 10] や適合モデル [1] などの確率的言語モデル [11] を使用した検索手法が多く提案されている．確率的言語モデルの基本的な考え方は，文書などのような言語表現に含まれる語の分布を用いて，クエリが与えられた下で，与えられた語が生成される確率がどれくらいかを推定するものである．しかしこのような手法には欠点が存在する．これらの手法は，基本的に bag of words と呼ばれる語の独立性の仮定の下に成り立っている．これはすべての語がそれぞれ他の語に依存しないという仮定である．よってほとんどの手法が，単語の役割や語の依存性などの，重要な問題を無視していることになる．とはいえ，大規模な文書集合に対してコストの高い言語解析を行うのは現実的とはいえない．そこで，マルコフ確率場と呼ばれるグラフィカルモデルを用いて語の依存性をモデリング [4, 12] することで，上に述べた問題を解決することが考えられる．しかしながら，自然言語クエリに対してはそのままでは特徴量を計算する語の組み合わせが爆発的に増えるなどの問題があり，何らかの拡張が必要である．. 本研究は，上に述べたマルコフ確率場に基づく語間依存性モデルの考え方に基づきつつ，自然言語クエリの分析によって得られる文節や係り受け関係などの構造情報を利用して特徴量の選択を行うものである．なお，当該手法を日本語情報検索に拡張した研究として，本研究とは別に [5] がある．本研究が日本語の自然言語クエリを想定しているのに対して [5] はそうではない．

2.1 語間依存性モデル

この節では，マルコフ確率場（以下，MRF と称する）モデルを検索に用いた，最初の研究である語間依存性モデル [4] について説明する．

MRF モデルは，語の依存性をモデリングする手法であり，統計的機械学習で広く使われている．Metzler と Croft は MRF モデルをクエリの構造化とそれによる文書ランキングに適用した語間依存性モデルを提案した [4]．このモデルではクエリと文書の間に，同時分布 $P_\Lambda(Q, D)$ が存在すると仮定する．ユーザーが適合だとしたクエリ・文書ペアが与えられるとする．このリストは適合分布からのサンプルだとみなすことができる．それによって，なんらかの基準によって Λ が推定できる．さらに $P_\Lambda(D|Q)$ によって文書をランク付けする．

MRF はグラフ G で構成される．ノードは確率変数を表し，辺は確率変数間での依存関係の有無を表している．特に各確率変数は隣接するもののみに依存する．このモデルでは，G は，クエリノード q_i と文書ノード D で構成されるとする．そして G での確率変数の同時確率分布は次のように定義される．

$$P_\Lambda(Q,D) \quad = \quad \frac{1}{Z_\Lambda}\prod_{c\in C(G)}\psi(c;\Lambda) \qquad (1)$$

ここで $Q=\{q_1,...,q_n\}$ で $C(G)$ は G 中のクリーク集合である．ここでクリークとは完全部分グラフのことである．ψ はクリーク構造における，非負のポテンシャル関数である．そして $Z_\Lambda=\sum_{Q,D}\prod_{c\in C(G)}\psi(c;\Lambda)$ は正規化定数である．同時確率分布はグラフ G，ポテンシャル関数 ψ，パラメータ Λ によって一意に定義される．

文書をランク付けするために，クエリが与えられた下での文書が得られる確率を計算する．この際，相対評価によって文書をランク付けするため，両辺の対数をとっても問題ない．さらに $P_\Lambda(Q)$ はクエリ Q が生成される確率であるが，文書をランク付けする際には定数と考えてよいため，次のようになる．

$$\begin{aligned} P_\Lambda(D|Q) \quad &= \quad \frac{P_\Lambda(Q,D)}{P_\Lambda(Q)} \\ &\stackrel{\text{rank}}{=} \quad \log P_\Lambda(Q,D)-\log P_\Lambda(Q) \qquad (2) \\ &\stackrel{\text{rank}}{=} \quad \sum_{c\in C(G)}\log\psi(c;\Lambda) \end{aligned}$$

すべてのポテンシャル関数は非負であり，一般的に次のようにパラメータ化される

$$\psi(c;\Lambda) \quad = \quad \exp[\lambda_c f(c)] \qquad (3)$$

ここで $f(c)$ はクリーク値のもとでの実数の特徴関数であり，λ_c は特定の特徴関数に与えられる重みである．これをランキング関数に戻して適用すると次のようになる．

$$P_\Lambda(D|Q) \quad \stackrel{\text{rank}}{=} \quad \sum_{c\in C(G)}\lambda_c f(c) \qquad (4)$$

ここで例として，3つの依存性仮定を述べる．

- クエリ語が互いに独立である仮定 (full independence)
- 隣り合うクエリ語だけに依存がある仮定 (sequential dependence)
- 任意のクエリ語の間に依存がある仮定 (full dependence)

これら3つの仮定を表すために，次のようなランキング関数が導かれる．

$$P_\Lambda(D|Q)\stackrel{\text{rank}}{=}\sum_{c\in T}\lambda_T f_T(c)+\sum_{c\in O}\lambda_O f_O(c)+\sum_{c\in O\cup U}\lambda_U f_U(c) \qquad (5)$$

ここで T はクエリ語と文書の2つのノードを持つクリーク集合である．O は文書ノードとクエリにおいて連続して現れるような，2つ以上のクエリ語ノードを持つクリーク集合である．U は文書ノードとクエリにおいて任意の組み合わせで得られる，2つ以上のクエリ語ノードを持つクリーク集合である．図1はこれらを表現したグラフィカルモデルである．

2.2 日本語係り受け解析

本研究では，クエリとして日本語の自然言語文を想定する．自然言語文からクエリを生成するために，まず文を語単位に分ける必要がある．そのために，形態素解析と呼ばれる処理を行い，文を形態素（おおまかにいえば，言語で意味を持つ最小単位）の系列に分割し，それぞれの品詞を判別する．表1に形態素解析の例を示す．なお，この例では形態素解析器として，本研究でも使用した茶筌 (ChaSen) を用いた．

表 1: 形態素解析の例

文字列	読み	原形	品詞の種類
お待ち	オマチ	お待ち	名詞
し	シ	する	動詞
て	テ	て	助詞
おり	オリ	おる	動詞
ます	マス	ます	助動詞
.	.	.	記号

また日本語において，文はいくつかの形態素を持つ文節という単位に分けられる．そして各文節には修飾する側と修飾される側が存在する．このような関係のことを係り受けという．この係り受け構造を発見するために，係り受け解析を行う．係り受け構造がわかれば，文節間の依存関係がわかることになる．本研究では，この文節間での係り受け関係を用いて，クエリを構造化する．図2に係り受け構造の例を示す．この例では，係り受け関係を「→」で示し，係り受け関係の特殊な場合である同格・並列関係を「=」で示している．提案手法において，この同格・並列関係は係り受け関係とは区別して扱う．また本研究では，係り受け解析器として南瓜 (CaboCha)[13] を用いた．

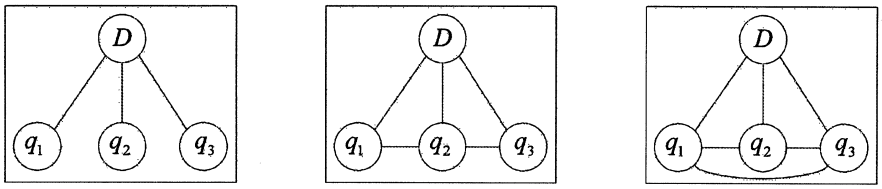

図 1: 3 つの依存性仮定 ((左) full independence (中央) sequential dependence (右) full dependence) の下での, 3 つのクエリ語が与えられたときの MRF モデル例 [4]

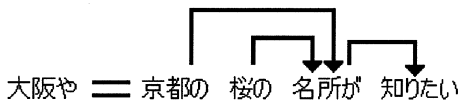

図 2: 係り受け構造の例

3 構造化クエリの自動生成

本節では, 2.1 節で説明した手法を拡張した提案手法を紹介する. 2.1 節の手法は, 入力として複数のクエリ語を想定していた. この手法では, 自然言語クエリにはうまく適用できない. そのため, 本研究では, この手法を自然言語文に適用し, さらに文の解析結果を利用してクエリを構造化する. 提案手法では, 文節, 係り受け関係, 品詞情報の 3 つの情報を用いる.

前述したように, 提案手法は MRF モデルに基づいている. 2.1 節で説明した語間依存性モデルと同じように, 提案手法においてもグラフは 1 つの文書ノードといくつかのクエリノードからなる. このときクエリノードは形態素単位とする. このため, 任意の組み合わせを考えると計算量が爆発的に増大する. 本研究では, この問題を解決するために係り受け関係と品詞情報を用いる. これらの情報を用いることで, 依存性を仮定するクリークを限定する. まず, 係り受け関係として依存性が認められるクリークのみを用い, それ以外のクリークは依存性がないものとみなして無視する. さらに, 品詞情報によって, 有効でないと思われる品詞の語を省く. これらによって, 計算量を減少させることができる. また, 文節情報については, クエリの中で文節内の語は特に依存性が高いと思われるので, それを 1 つのフレーズとして扱うことを考える. なお, すでに述べた係り受け関係も文節を単位とする.

以上の考えに基づいて, 新たに提案する 3 つの依存性仮定について述べる.

- クエリ語が互いに独立である仮定
  (FI : full independence)
- 文節内で隣接した語の間にだけ依存性がある仮定
  (PD : phrase dependence)
- 係り受け関係にある文節を構成するすべての語が互いに依存している仮定
  (MD : modification dependence)

これら 3 つの仮定を表すために, 式 (6) のようなランキング関数が導かれる.

$$P_\Lambda(D|Q) \stackrel{\text{rank}}{=} \sum_{c\in T} \lambda_T f_T(c) + \sum_{c\in P} \lambda_P f_P(c) + \sum_{c\in M} \lambda_M f_M(c) \tag{6}$$

ここで T は, 品詞が名詞・英語・数字・未知語であるクエリ語それぞれと文書の 2 つのノードを持つクリーク集合である. P は, 文書ノードと自然言語クエリ内の各文節内で連続する複数の名詞からなるフレーズもしくは文節全体のフレーズのみを対象としたクリーク集合である. このとき, フレーズが動詞・形容詞・名詞・英語・数字・未知語以外の語のみで構成される場合は, 意味のないフレーズであると考えて対応するクリークを無視する. また, 特殊な例として, 接尾辞・接頭詞が含まれる場合, これらの被修飾語との依存性がより強いため, その部分を 1 つのフレーズとして扱う. M は, 文書ノードと自然言語クエリ内の各文節における連続する 1 つ以上の語からなるフレーズと, 係り受け関係にある文節内の連続する 1 つ以上の語からなるフレーズを対象としたクリーク集合である. このとき, 係り受け関係にある文節に対応するクリークにおいて, どちらか一方の文節においても, そのフレーズが動詞・形容詞・名詞・英語・数字・未知語以外の語のみである時は, 意味のないフレーズであると考え

て対応するクリークを無視する．また，同格・並列の場合のみ，同格・並列関係にある相手が係り受け関係にある文節とのクリークについても考慮する．提案手法では，このように特徴関数に用いるクリークを限定する．また，$\lambda_T + \lambda_P + \lambda_M = 1$ とする．

4 適合モデルによるクエリ拡張

適合モデルは [1] によって提案され，さらに擬似適合フィードバックで通常行われるようにユーザーが入力した元々のクエリと組み合わせて用いることで有効性が向上することが報告されている [14, 3]．

まず適合モデルについて説明する．検索対象の文書集合が仮に適合クラスと不適合クラスに分割できるとする．このとき，適合モデルは適合クラスに属する文書群を用いて推定された，検索要求に関する言語モデルである．これをクエリとして用いて文書集合から適合文書を検索する．現実には適合クラスを完全に把握することは困難なため，ユーザーから与えられるクエリから適合文書モデルを推定し，検索対象の文書を順序付ける．クエリを $Q = \{q_1, ..., q_n\}$，適合クラスを R とするとき，語 w を適合モデル R から観測する確率 $P(w|R)$ は次の式のように表すことができる．

$$P(w|R) = \sum_{D \in C} P(w|D)P(D|q_1, ..., q_n) \qquad (7)$$

現実に適合モデルを用いる場合は，$P(w|R)$ は次のような近似によって求められることが多い．式 (7) における $P(D|q_1, ..., q_n)$ を用いて降順に順序付けされた文書の上位 N 個を利用して，式 (7) により混合分布を構築する．この混合分布から，確率値が大きい語 w のうち M 語を用いることで，$P(w|R)$ を近似する．

本研究では，この適合モデルを次に示すように，前節で述べた構造化クエリと組み合わせて用いる [14]．提案手法によって生成された構造化クエリ Q_{st} が与えられると，文書集合が検索結果として得られる．この検索結果の上位 N 件により，式 (7) に基づいて適合モデルを推定する．推定された適合モデルから確率値の大きい語から M 語をクエリ Q_{rm} として近似する．そして構造化クエリ Q_{st} と適合モデルによるクエリ Q_{rm} を組み合わせる．最終的に拡張クエリは ν をパラメータとして，次のような式で表せる．

$$Q_{new} = \nu Q_{st} + (1.0 - \nu)Q_{rm} \qquad (8)$$

5 実験

5.1 データと実験準備

提案手法の有効性を検証するため，評価実験を行う．提案手法では，自然言語クエリにおける係り受け関係などの情報を用いて，クエリを再構築する．そこで比較対象として，すべての語の独立性を仮定したモデル（FI モデル）を用いる．

評価実験の説明をする前に，検索システム Indri[15] について説明する．本研究では，これをインデクシングなどの検索プラットフォームとして利用する．本提案手法は Indri に限らず，近接演算の可能な検索エンジンで適用可能である．

今回用いたクエリ表現は次の 2 つである．

- #N(...)：各語が並び順に出現し，各語の間隔は $N-1$ 語以内であることを表す
- #uwN(...)：語の順序に関係なく，N 語の間にすべての語が出現することを表す

提案手法において，PD モデルに用いた#N(...) の N はクエリの語数，MD モデルに用いた#uwN(...) の N はクエリの 語数 × ウィンドウ幅 とした．ここでウィンドウ幅にパラメータとして経験的に定めることにする．4 節で述べた擬似適合フィードバックは Indri のクエリ言語で次のようにして実現した．

$$Q_{new} = \#\text{weight}(\nu Q_{st} \ \ (1.0 - \nu)Q_{rm}) \qquad (9)$$

ここで ν はパラメータ，#weight は Indri の演算子で，それぞれのクエリに重みを付けるために用いる．Q_{st} は本研究で提案する自然言語クエリから生成した構造化クエリを示す．上式で得られる拡張クエリによって検索を行う．

評価実験では，NTCIR-3 WEB と呼ばれる Web 検索評価用テストコレクション [16] を用いて，評価を行った．NTCIR-3 WEB コレクションは，100GB の文書データ（主として．jp ドメインから収集した HTML もしくはプレーンテキストファイルであり，言語は主に日本語と英語，ごく一部にその他の言語が用いられている），検索課題 47 件（日本語），適合判定データからなる．本論文では，上述のテストコレクションを用いて各種パラメータを調整し，検索有効性の評価を行った．また検索課題の DESC フィールドに記述された自然言語文を，クエリとして使用した．このような検索課題が 47 件あり，それぞれから得られる自然言語クエリに対して提案した手法でクエリを再構築

表 2: ウィンドウ幅を変化させたときの，提案手法 (FI+PD+MD) の実験結果

window size	MAP
2	0.1641
5	0.1659
10	0.1678
20	0.1661
50	0.1657

する．4 節で説明した擬似適合フィードバックを行った場合についても評価を行う．評価指標には平均精度 (MAP: Mean Average Precision —non-interpolated) を用いた．

またクエリ生成の時点で次のような 2 種類のストップワードを指定した．

- 「知りたい」や「探したい」などの自然言語のクエリ文に特有の，検索に有効でないと思われるフレーズ
- 1 バイトのアルファベット・数字，1 バイト・2 バイトの記号，単一のひらがな・カタカナ，任意のひらがな対

5.2 実験結果

まず 1 つ目の実験として，提案手法での，MD モデルに用いている語の出現順序を考慮せずに依存性を仮定するウィンドウ幅について最適値を求める実験を行った．この実験の結果を表 2 に示す．その際，MAP を評価基準として評価を行った．

表 2 によると，ウィンドウ幅 10 まで MAP の値が上昇し，そのあとウィンドウ幅が大きくなるにつれて減少している．よってウィンドウ幅の最適値を 10 として固定し，このあとの実験を行う．

提案手法では，文節間での依存を仮定したモデルと係り受け関係のある語間での依存を仮定したモデルを提案した．そこで次に，文節間での依存を仮定したモデルを用いた場合 (FI+PD) と，係り受け関係のある語間での依存を仮定したモデルを用いた場合 (FI+MD)，さらにこれらを結合した場合 (FI+PD+MD) について実験を行った．この際，ウィンドウ幅を最適値に固定し，各々のモデルに対し，ポテンシャル関数の重みパラメータの最適値を求める実験を行った．その結果，文節間での依存を仮定したモデル (FI+PD) の最適パラメータは $(\lambda_T, \lambda_P) = (0.95, 0.05)$，係り受け関係のある語間での依存を仮定したモデル (FI+MD) の最適パラメータは $(\lambda_T, \lambda_M) = (0.9, 0.1)$，これらを結合したモデル (FI+PD+MD) の最適パラメータは $(\lambda_T, \lambda_P, \lambda_M) = (0.9, 0.05, 0.05)$ となった．パラメータ調整に関し，各パラメータを 0.1 刻みで変えていき，最適値付近では 0.05 刻みで実験を行った．これらのパラメータのもとで，FI モデルと提案手法を比較した結果を表 3 に示す．表 3 によると，FI モデルと比べ，文節間だけに依存を仮定したモデル (FI+PD) では 4.40%，係り受け関係のある語だけに依存を仮定したモデル (FI+MD) では 8.08%，これら 2 つのモデルを結合したモデル (FI+PD+MD) では 8.47% 上昇した．クエリ中のすべての語の独立性を仮定した FI モデルと比べ，文節間や係り受け関係のある語の間にだけ依存関係を仮定した提案手法で大きな改善が見られた．これによって自然言語文のクエリが与えられたとき，文節や係り受け関係の情報を検索に使用することは有効であるということがわかった．

さらに 4 節の擬似適合フィードバックを組み合わせた場合についても実験を行った．この際，擬似適合フィードバックに用いる文書数 N とクエリの数 M をそれぞれ 5, 10, 20 と変化させ最適値を求め，さらに式 (8) の構造化クエリに与える重み ν を 0.5, 0.7, 0.9 と変化させ最適値を求めた．その結果，FI モデルにおいて，フィードバックに用いる，文書数，単語数，および式 (8) のパラメータは $(N, M, \nu) = (5, 20, 0.7)$，文節間での依存を仮定したモデル (FI+PD) の最適パラメータは $(N, M, \nu) = (5, 20, 0.5)$，係り受け関係のある語間での依存を仮定したモデル (FI+MD) の最適パラメータは $(N, M, \nu) = (10, 5, 0.7)$，これらを結合したモデル (FI+PD+MD) の最適パラメータは $(N, M, \nu) = (10, 10, 0.5)$ となった．これらのパラメータのもとで，FI モデルと提案手法を比較した結果を表 4 に示す．また各モデルにおける，$\nu = 0.7$ に固定したときの，パラメータの組み合わせごとの結果についても表 5 に示す．ここで，$\%\text{increase}_{fd}$ は各モデルにおいて，擬似適合フィードバックによる改善率，$\%\text{increase}_{total}$ は擬似適合フィードバックを行わない場合の FI モデルと比べたときの改善率を示す．

表 4 によると，提案手法に擬似適合フィードバックを行った場合とそうでない場合を比較すると，行った場合 11.9% の向上が見られた．また，擬似適合フィードバック環境において，すべてのクエリが独立であると仮定したモデル (FI) と提案手法を同じ条件で比較すると，提案手法は，14.0% の向上が見られた．また，

表 3: 各モデルにおける最適パラメータでの実験結果

Model	MAP	%increase
FI	0.1547	0.000
FI + PD	0.1615	4.396
FI + MD	0.1672	8.080
FI + PD + MD	0.1678	8.468

すべてのクエリが独立であると仮定したモデル (FI) でフィードバックを行わない場合と提案手法に擬似適合フィードバックを行った場合を比較したところ，最大 21.3% の向上が見られた．擬似適合フィードバックを行った場合，FI モデルでは最大 6.46%，提案手法では最大 11.9% の改善率であったことから，提案手法の方がフィードバックによる改善率が高くなっている．そのような理由からも，提案手法で生成されるクエリによって検索された文書集合の方が，より洗練されているということがわかる．

6 おわりに

本稿では，日本語の自然言語クエリを用いた文書検索において，係り受け関係などの構文情報を用いてクエリを再構築する手法を提案した．自然言語クエリには，ユーザーの情報ニーズとは異なる語が現れるという問題があった．これに対して従来の語間依存性モデルでは，多くの有効でないクエリを生成してしまう可能性があった．提案手法を用いてクエリを生成することで，無駄なく，かつより精度の高い検索ができるようになった．すべてのクエリが独立であると仮定したモデルと文節内や係り受け関係にある語の間に依存を仮定した提案した手法を比較したところ，提案手法は，8.47% の向上が見られた．また，提案手法で得られた検索結果に，擬似適合フィードバックを行った場合についても評価した．それによると，提案手法に適合フィードバックを行った場合とそうでない場合を比較すると，行った場合 11.9% の向上が見られた．また，それぞれ最適パラメータでの擬似適合フィードバック環境において，すべてのクエリが独立であると仮定したモデルと提案手法を比較すると，提案手法は，14.0% の向上が見られた．また，すべてのクエリが独立であると仮定したモデルと提案手法に擬似適合フィードバックを行った場合を比較したところ，最大 21.3% の向上が見られた．クエリ中のすべての語の独立性を仮定したモデルと比べ，文の解析情報を用いた提案手法で大きな改善が見られた．これによって自然言語文のクエリが与えられたとき，文節や係り受け関係などの構文情報を検索に使用することは有効であるということがわかった．

謝辞

本研究の一部は，科学研究費補助金特定領域研究「情報爆発 IT 基盤」(19024055)，基盤研究 (B) (20300038)，萌芽研究 (18650057) の援助による．

参考文献

[1]Lavrenko, V. and Croft, W. B.: Relevance Based Language Models, pp. 120–127 (2001).

[2]Zhai, C. and Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval, *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, New York, NY, USA, ACM, pp. 403–410 (2001).

[3]Diaz, F. and Metzler, D.: Improving the Estimation of Relevance Models Using Large External Corpora, *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, pp. 154–161 (2006).

[4]Metzler, D. and Croft, W. B.: A Markov Random Field Model for Term Dependencies, *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479 (2005).

[5]Eguchi, K. and Croft, W. B.: Query Structuring with Two-stage Term Dependence in the Japanese Language, *Information Retrieval Technology: Third Asia Information Retrieval Symposium*, pp. 522–529 (2006).

[6]Cronen-Townsend, S., Zhou, Y. and Croft, W. B.: Predicting Query Performance, *Proceedings of SIGIR 2002*, pp. 299–306 (2002).

[7]Zhou, Y. and Croft, W. B.: Query Performance Prediction in Web Search Environments, *Proceedings of the 30th Annual International ACM SIGIR Conference*, New York, NY, ACM, ACM, pp. 543–550 (2007).

表 4: 擬似適合フィードバックを用いた場合の各モデルの最適パラメータでの実験結果

Model	MAP	MAP(after feedback)	%increase$_{fd}$	%increase$_{total}$
FI	0.1547	0.1647	6.46	6.46
FI + PD	0.1615	0.1778	10.10	14.93
FI + MD	0.1672	0.1831	9.51	18.36
FI + PD + MD	0.1678	0.1877	11.86	21.33

表 5: 擬似適合フィードバックを用いた場合の各パラメータにおける実験結果（$\nu = 0.7$ に固定）．%increase$_{fd}$ は各モデルにおいて，擬似適合フィードバックによる改善率，%increase$_{total}$ は擬似適合フィードバックを行わない場合の FI モデルと比べたときの改善率を示す．

Model	(N, M, ν)	MAP	%increase$_{fd}$	%increase$_{total}$
FI	FI only	0.1547	0.00	0.00
	(5,5,0.7)	0.1586	2.52	2.52
	(5,10,0.7)	0.1626	5.11	5.11
	(5,20,0.7)	0.1647	6.46	6.46
	(10,5,0.7)	0.1459	-5.69	-5.69
	(10,10,0.7)	0.1521	-1.68	-1.68
	(10,20,0.7)	0.1531	-1.03	-1.03
	(20,5,0.7)	0.1577	1.94	1.94
	(20,10,0.7)	0.1572	1.62	1.62
	(20,20,0.7)	0.1598	3.30	3.30
FI + PD + MD	FI + PD + MD only	0.1678	0.00	8.47
	(5,5,0.7)	0.1744	3.93	12.73
	(5,10,0.7)	0.1758	4.77	13.64
	(5,20,0.7)	0.1764	5.13	14.03
	(10,5,0.7)	0.1833	9.24	18.49
	(10,10,0.7)	0.1863	11.03	20.43
	(10,20,0.7)	0.1795	6.97	16.03
	(20,5,0.7)	0.1636	-2.50	5.75
	(20,10,0.7)	0.1697	1.13	9.70
	(20,20,0.7)	0.1753	4.47	13.32

[8]Ponte, J. and Croft, W. B.: A Language Modeling Approach to Information Retrieval, New York, NY, ACM, ACM, pp. 275–281 (1998).

[9]Hiemstra, D.: A Linguistically Motivated Probabilistic Model of Information Retrieval, *European Conference on Digital Libraries*, pp. 569–584 (1998).

[10]Song, F. and Croft, W. B.: A General Language Model for Information Retrieval, *Proceedings of Eighth International Conference on Information and Knowledge Management* (1999).

[11]Manning, C. D., Raghavan, P. and Schütze, H.: *Introduction to Information Retrieval*, Cambridge University Press (2008).

[12]Metzler, D. and Croft, W. B.: Latent Concept Expansion Using Markov Random Fields, *Proceedings of the 30th Annual International ACM SIGIR Conference*, New York, NY, ACM, ACM, pp. 311–318 (2007).

[13]工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, Vol. 43, No. 6, pp. 1834–1842 (2002).

[14]Metzler, D., Strohman, T., Turtle, H. and Croft, W.: Indri at TREC 2004: Terabyte Track (2004).

[15]Metzler, D. and Croft, W.: Combining the Language Model and Inference Network Approaches to Retrieval, *Information Processing and Management Special Issue on Bayesian Networks and Information Retrieval*, pp. 735–750 (2004).

[16]Eguchi, K., Oyama, K., Ishida, E., Kando, N., Kuriyama, K.: Overview of the Web Retrieval Task at the Third NTCIR Workshop (2003).