あいまいなエピソードからのオブジェクト検索とクエリの対話的修正

落合　修也† 　加藤　誠†† 　田中　克己††

† 京都大学工学部情報学科計算機科学コース　〒606–8501 京都府京都市左京区吉田本町
†† 京都大学大学院情報学研究科社会情報学専攻　〒606–8501 京都府京都市左京区吉田本町
E-mail: †ochiai@dl.kuis.kyoto-u.ac.jp, ††{kato,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし　本論文では，あいまい性を含む検索クエリに対して，検索結果を効率よく絞り込むような質問をユーザに提示し，それに対してユーザが回答することでクエリを対話的に修正する手法を提案する．システムは，検索結果を絞り込むために必要な情報をユーザのクエリから推定し，その情報がユーザの回答によって得られるような質問をユーザに提示する．また，ユーザが回答しうる情報を網羅し，必要な情報の特徴をよく反映したオブジェクトを選択肢としてユーザに提示する．また，このシステムを利用することによるオブジェクト検索の有効性について検証する．

キーワード　クエリ修正，質問提示，対話的情報検索，オブジェクト検索，適合フィードバック

1. はじめに

1.1 背　景

インターネットと Web 検索エンジンの普及により，自分の忘れてしまった知識について，コンピュータを用いて手早く簡単に入手することが可能になった．その代表的な例として，人物や歴史的な事象などのオブジェクトの名前を覚えているが，その人物や事象に関する詳細な情報を忘れてしまった場合などが挙げられる．この場合，自分が覚えているオブジェクト名を検索クエリとして検索することにより，そのオブジェクトについて書かれた Web ページから詳細な情報を獲得することができる．例えば，メジャーリーグベースボールで活躍したバリー・ボンズ選手がどのような選手であったか忘れてしまった場合,「バリー・ボンズ」というクエリで検索すれば，ユーザはバリー・ボンズに関するエピソードやプロフィールを容易に取得できる．

しかし，その逆を行う検索，すなわち人物や事象に関する詳細な情報を覚えているが，その人物や事象などのオブジェクトの名前を忘れてしまった場合，その名前を Web 検索で取得するのは困難である．これには 2 つの要因がある．1 つはユーザの記憶している情報がエピソードを含んだ自然文で構成されている点である．現状の Web 検索エンジンでは，記憶から呼び出した自然文をそのまま入力しても，ユーザが探している情報を発見できないことが多い．したがって，システムへ検索意図を正確に伝えるための適切なクエリをユーザ自身で模索しなければならない．もう 1 つはユーザの入力する情報の根拠が自身の記憶による点である，脳内にある記憶は，時が経つとともに互いに干渉し合う．ユーザは干渉し合っている記憶から情報を部分的に引き出すため，情報が不十分であったり，誤った情報，すなわちノイズが元の情報に含まれてしまうことがある．

例えば,「10 年程前に MLB で数々の記録を打ち立てるなどして，日本のメディアでもその活躍が大々的に報じられていた，元ニューヨーク・ジャイアンツ所属の巨漢選手」というクエリの元となる情報には，選手が所属していたチーム名が誤っており(注1)，選手が達成した具体的な記録が表記されていない．このように，ユーザの記憶による情報は不完全であり，情報の至る所にあいまいな表現が含まれている．本稿では，以後この情報を**エピソード**と呼び，エピソードに含まれる情報のあいまいな表現のことを**あいまい性**と呼ぶことにする．

1.2 あいまい性の種類

エピソードを元にした情報検索において，出現するあいまいな表現には 4 つの種類があり，それぞれ「情報の欠損」,「単語の汎化」,「情報の誤り」,「印象に基づいた表現」と定義する．

「情報の欠損」は，実際に見たり体験したエピソードの情報を，忘却や記憶同士の干渉により部分的に喪失してしまうことを指す．1.1 節の例で考えると，ユーザは巨漢選手がどのような活躍をしたか忘れてしまっており,「数々の記録を打ち立てる」という表現で置き換えているが，この部分が「情報の欠損」にあたる．

「単語の汎化」は，エピソードに登場する正しい情報（単語）を正確には覚えておらず，その情報の上位概念にあたる情報を代わりにクエリとして含めることを指す．例えば，ある野球選手が所属するチームを正確に覚えていないため，少なくともナショナルリーグに在籍する球団に所属していたという意味合いで「ナショナルリーグ」という単語をクエリに追加した場合などが「単語の汎化」に該当する．

「情報の誤り」は，エピソードが持つ情報そのものに誤りがあることを指す．1.1 節での例では，先述の通り，所属球団の表記誤りが「情報の誤り」の一例となるが，別の例として正しい所属チームと実在する別のチーム（ニューヨーク・ヤンキースなど）を取り違えてしまう場合も考えられる．この場合も「情報の誤り」として扱うことにする．

(注1)：「元ニューヨーク・ジャイアンツ所属」の情報が誤っている．正しくは「元サンフランシスコ・ジャイアンツ所属」.

「印象に基づいた表現」は，ユーザの記憶に存在するエピソードをクエリとして入力する際に，自らの印象や主観による表現が含まれてしまうことを指す．1.1 節での例では，エピソード内に「日本のメディアでもその活躍が大々的に報じられていた」と記述されているが，この表現はあくまでも自分の考えや経験に依ったものであり，報道自体がそれほど大きなものではなかったかもしれない．他にも，エピソードに対して「びっくりした」「感激した」「興味深かった」などの形容詞・動詞を含んだ表現も「印象に基づいた表現」である．

ユーザの記憶の中にあるエピソードからオブジェクトを検索するには，単純に入力情報をクエリにするだけではなく，システムがユーザの入力した情報のあいまい性について吟味する必要がある．

1.3 提案手法

本研究での目的は，ユーザが入力するあいまいなクエリから，ユーザの意図に適合するオブジェクトを特定することである．我々は，1 つのアプローチとして，システムが調べたいオブジェクトの属性や性質についてユーザに質問と選択肢を提示し，ユーザがそれに回答することによってクエリを対話的に修正していく手法を提案する．システムはクエリのあいまい性を解消できる質問や，ユーザが知識として持っていると予想される質問，すなわちユーザにとって答えやすい質問を優先的に提示する．同時に選択肢を提示するが，選択肢は特徴がはっきりとしており，かつ知名度の高いオブジェクトを提示する．そして，選択したオブジェクトが持つ特徴語の中で，質問に関連するものを抽出し，それを元にしてクエリの修正をする．

本提案の特筆すべき点として，選択肢を単語などの言葉で提示するのではなく，具体的なオブジェクトで提示している点が挙げられる．これには 2 つの理由があり，1 つはユーザが単語をクエリとして直接的に入力して検索する手法には限界があるからである．この問題の典型例が「顔の記憶」である．過去の出会った人物の顔の情報を「再生」(注2) しても，その視覚的表象が間接的で質の低い言語的説明に変換されるため，ターゲットとなる人物の同定には至らないことが多い [2]．本手法では，オブジェクトが持つ特徴語を複数の Web ページから取得し，これらの特徴語が質問の回答としてふさわしいか判定する．そして，ふさわしいとされる特徴語をすべてクエリとして追加するため，ユーザがうまく言葉で表現できないような特徴を，システムによって提示されたオブジェクトの選択という簡単な操作でクエリ入力できる．

もう 1 つは，記憶を「再生」させるより「再認」させる方が正しい記憶を引き出すことができるという心理学の見解があるからである．記憶の測定方法には先述の「再生」の他にも「手がかり再生」(注3) や「再認」(注4) があるが，一般に再生よりも手がかり再生や再認の方が同じ記憶であっても記憶の内容の正答率が高いとされている [6]．本手法では，記憶想起のきっかけとなる「手がかり」として質問を提示し，同時に回答をする上での選択肢を提示しているという点で，心理学における「手がかり再生」と「再認」の二面性を保有していると考えられる．また，言語的説明の断定的な選択肢を提示するよりも，ユーザにとってなじみがあって単峰性の強いオブジェクトを選択肢として提示する方が，選択肢の適合・不適合の境界がよりあいまいになり，厳密に選択肢を選択しなければならないと考えるユーザの心理的な負担を軽減させることができると考えた．

以上の手法により，ユーザはクエリの初期入力と質問に対する多肢選択形式の回答のみによって，探しているオブジェクトを特定することができる．

本稿の構成は以下の通りである．2 章で関連研究について触れ，3 章で入力語からオブジェクトを検索する手法について述べる．4 章でユーザの検索意図に基づいたクエリの修正方法について俯瞰し，5 章で提示する質問の優先度と選択肢が満たすべき性質について言及する．6 章でシステムの実装について簡潔に述べる．7 章で実装したシステムの評価結果を報告し，8 章では，本研究のまとめと今後の課題について述べる．

2. 関連研究

2.1 オブジェクトの検索

名前のわからないオブジェクトをの検索方法として，稲川ら [3] はある語集合がオブジェクトを特定するという関係を，Web から抽出する手法を提案している．これは，クエリとなる語集合が特定するオブジェクトの候補を求め，各候補についてクエリから特定される関係としての適合度を評価するというものである．語集合からオブジェクトを検索するという過程は本研究と同一であるが，本研究ではユーザの入力する語集合（エピソード）の情報があいまい性を含むことを想定している．すなわち，入力した情報が不適切であったり不十分であったりする場合について，ユーザからのキーワード入力による直接的なクエリ修正なしで，調べたいオブジェクトを発見する手法を我々は提案する．また，オブジェクトが持つ属性や情報を直接クエリにしない方法もある．大島ら [8] は，クエリを 1 つのオブジェクト名とし，ユーザにとって既知のオブジェクトから，そのオブジェクトと同位関係にあるオブジェクトをたどっていく手法を提案している．

2.2 あいまい性の解消

検索クエリのあいまい性を解消する研究は広く行われている．Brill and Moore [1] はクエリに含まれる語の部分文字列を考慮した誤り訂正モデルを考案し，スペルミスを含んだクエリの高精度な訂正を可能にした．このスペルミスという概念は，1.2 節で述べたあいまい性の区分「情報の誤り」の一部にあたる．本研究では，質問に対するユーザの回答から得られる新しい情報を元のクエリに追加することにより，クエリそのものに対して適切な検索結果が得られるような修正を行う．

栗田ら [4] は，ユーザに学習用の絵画に対して印象語を付けさせ，その結果から正準相関分析により印象語と画像特徴との相関関係の学習を行っている．これにより，「ロマンチックで暖

(注2)：心理学における記憶の測定方法の 1 つ．学習した刺激を思い出して，書き出したり口頭で説明したりする．

(注3)：刺激項目に対する手がかりを与え，それをもとにして再生する方法．

(注4)：学習した項目と学習しなかった項目を提示し，学習時に提示された項目を指摘させる方法．

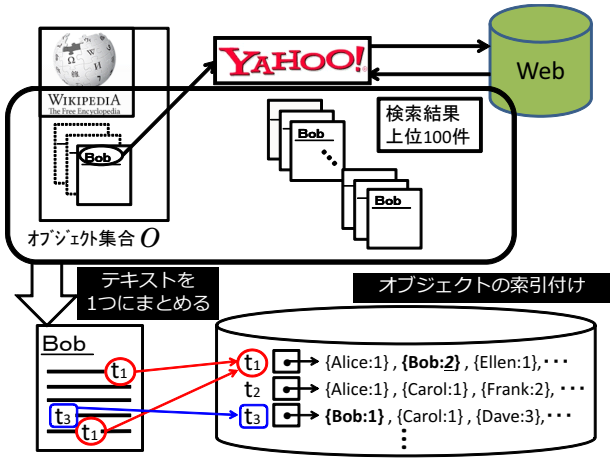

図 1　オブジェクトに関する文書生成とオブジェクトの索引付け

かい」と言った印象語から絵画データベースを検索することが可能になる．このように，あいまい性「印象に基づいた表現」をユーザの好みや文化的背景を反映させて検索に利用する試みもなされている．

また，軽部 [5] は，ユーザが検索結果ページ内の文章片を自分の検索意図に応じて取捨選択することにより，検索結果を動的にリランキングしている．この手法により，クエリに多義語が含まれている場合について，検索結果を検索意図と同義のページのみに絞り込むことを可能にした．多義語を 1 つの意味に特定する情報を，クエリ修正で補完したという点でクエリが持つあいまい性の 1 つ「情報の欠損」を解消している．システムとの対話的なクエリ修正によってあいまい性を解消する点は軽部の提案手法と類似している．しかし，本研究ではユーザの行動をシステム側で選択肢として限定することにより，ユーザの検索意図とマッチするオブジェクトを簡単な動作で発見できるシステムを提案する．

3. エピソードからのオブジェクト検索

エピソードからのオブジェクト検索は以下の手順で行う：

(1) 各オブジェクトについて書かれている Web ページを取得し，それらのページを結合して 1 つの文書にする．

(2) 結合した文書に含まれる単語を用いてオブジェクトの索引付けを行う．

(3) クエリが入力されたときに，各オブジェクトとの適合度を評価し順位付けを行う．そして，上位のオブジェクトから順に検索結果として提示する．

なお，ユーザが入力するクエリはエピソードを表現する単語集合とする．以下では各手順について詳しく述べる（図 1）．

索引付けの対象となるオブジェクト集合 O はあらかじめ与える．本研究では，「Wikipedia」に項目として存在するオブジェクトに限定することにした．また，オブジェクトを索引付ける語には Wikipedia の記事およびそのオブジェクトの名前をクエリとしたときの Yahoo!検索(注5)の結果上位 100 ページに出現する語を利用する．これらのページから本文を抽出して，すべてのテキストを結合し 1 つの文書を生成した．次に，各オブジェクトとそれについて書かれた文書を関連づけてデータベースに格納し，文書から語を抽出してオブジェクトの索引付けを行った．ユーザから語集合 T がクエリとして入力されたときには，ベクトル空間モデル [9] を用いたランキングアルゴリズムにより，クエリと各オブジェクトの適合度を計算する．本研究では，オブジェクト検索のアルゴリズムについては深く議論せず，これ以降はクエリの対話的な修正方法について述べる．

(注5)：http://www.yahoo.co.jp

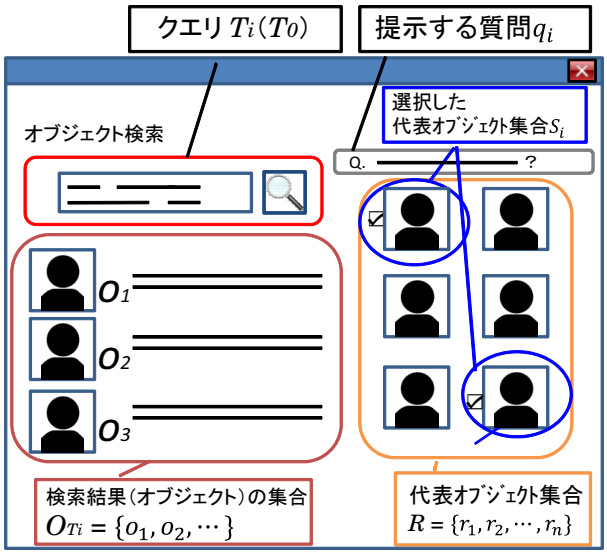

図 2　各変数の定義

4. クエリの対話的な修正方法

本章では，ユーザの入力したクエリ T_0 がどのような過程を経て修正されるかについて述べる．

本手法で用いる変数の定義を図 2 に示す．3. で述べた方法により，オブジェクト検索の結果 $O_{T_0} = \{o_1, o_2, \cdots\}$ を画面の左側に表示する．同時に，右側にはユーザに与える質問 q_0 とその選択肢 $R = \{r_1, r_2, \cdots, r_n\}$ を提示する．ここで提示される選択肢は，通常のアンケートのような語や自然文ではなく，ユーザの検索したいオブジェクトが所属するカテゴリと同一のカテゴリから選出されるオブジェクトである．本研究では，選択肢として提示するオブジェクトを**代表オブジェクト**と呼ぶ．さらに，それ以外の選択肢として「わからない」のボタンがある．なお，質問と選択肢の具体的な選択方法は 5. で詳しく述べる．

ユーザは質問 q_0 を読み，その質問が答えられるものであれば選択肢にチェックを入れることにより回答する．このとき，代表オブジェクトが持つ特徴語の中で，質問の回答となりうる特徴語のみを抽出し，クエリ修正を行う選択された代表オブジェクトと選択されなかった代表オブジェクトそれぞれが持つ特徴語集合 $S_0, \bar{S}_0$ について，新しいクエリ T_1 を導出する．そして，T_1 によって，再度オブジェクト検索を行う．

以後，ユーザが探しているオブジェクトを発見できるまで再帰的に繰り返すことにより，オブジェクトを発見するためのク

エリをより適切なものに修正する．クエリ T_i によって構成される質問ベクトルを $\boldsymbol{t_i}$，特徴語集合 $S_i, \bar{S}_i$ によって構成される特徴ベクトルをそれぞれ $\boldsymbol{s_i}, \boldsymbol{\bar{s}_i}$ とすると，一連のクエリ修正は次のように定式化できる．

$$\boldsymbol{t_{i+1}} = \alpha \boldsymbol{t_i} + \beta \boldsymbol{s_i} - \gamma \boldsymbol{\bar{s}_i} \tag{1}$$

本手法は適合フィードバックの概念に類似している [9]．しかし，適合フィードバックでは検索結果そのものに適合・不適合を指定するのに対し，本研究ではシステムが別で指定したオブジェクトにユーザが適合・不適合を指定する．

なお，もし質問に答えられなければ,「わからない」のボタンをクリックすると他の質問に移る．このとき，クエリ修正は行われず，システムは他の質問を試みる．

5. 質問選択と選択肢生成

本章では，クエリの対話的修正を行うにあたり，状況に応じた質問の選択方法と，ユーザが回答しやすい選択肢の生成方法についてその詳細を述べる．

5.1 質問集合の定義

本手法において提示される質問 q は，すべて予め用意されている質問集合 Q から選択される．質問集合にある質問は，大きく分けて次の 3 種類に分類される．

(1) オブジェクトが持つ性質を問う質問
(2) オブジェクトの見た目を問う質問
(3) エピソードの周辺情報を問う質問

(1) はユーザの検索対象としているオブジェクトが持つ属性を問う質問である．もし，検索対象がスポーツ選手ならば，その選手のしているスポーツや，所属チーム，ポジションに関する質問がこの種類に該当する．正解が明確に定まっているため，代表オブジェクトよりも属性を示す単語を選択肢として提示した方がユーザにとって回答しやすい場合がある．

(2) はオブジェクトの外見や，オブジェクトがユーザに与える印象について問う質問である．人の顔や体つきの特徴や，スポーツ選手のプレイスタイルなど，そのオブジェクトから連想されるイメージについてユーザに質問する．正解が明確に定まっていないため，直接的な文字入力や単語による選択肢では回答しづらいと考えられる．そこで選択肢として代表オブジェクトを提示し，それをユーザが選択することで，クエリとして入力できないようなあいまいな表現を追加の文字入力なしでクエリに追加できる．

(3) はオブジェクトそのものに関する質問ではなく，ユーザの持つエピソードに関連した情報（周辺情報）を問う質問である．エピソードの周辺情報には,「エピソードを経験した時間」,「エピソードを経験した場所」,「エピソードの情報源」,「エピソードのターゲット」,「エピソードのリアリティ」(注6) の 5 つがある [7]．本研究ではこの 5 つのうち，エピソードの時間に着目して，そのエピソードを経験した時代についてシステムが情報を取得できるような質問を数問考案した．例えば，検索対象のエピソードがあった時代と同時代に起こったエピソードを，選択肢の中から選ばせる質問などがある．

(注6)：ここでは，エピソードが本当にあったことなのかどうかを指す．

5.2 質問とオブジェクトの特徴語

次に，本手法で用いる質問とオブジェクトが持つ特徴語について触れる．質問 q_l には，質問文だけでなく質問から得る情報を端的に表すキーワードと回答となり得る特徴語の集合が関連づけられている．例えば,「その動物の色は次のどの動物に似ていますか.」という質問には,「色」というキーワードと「白」「黒」「緑」などの特徴語が関連づけられる．また，同様にオブジェクト o_j にもそれが持つ特徴を反映する特徴語の集合が関連づけられている．例えば,「カラス」というオブジェクトには「鳥類」「黒」「害鳥」などの特徴語が関連づけられる．これら特徴語は Wikipedia などの文書集合からキーワードに関連する単語を検索して事前に取得したものである．

5.3 質問の提示優先度

ユーザの検索対象となるオブジェクトを短時間で特定するために，システムはどの質問を提示すれば良いか考える必要がある．したがって，ユーザに質問を与える際，以下の 3 点を評価し，質問に優先順位をつけた．

5.3.1 クエリのあいまい性解消度

本手法におけるクエリ修正では，初期クエリだけでは取得できなかった情報を質問により取得し，検索対象のオブジェクトを再ランキングする．したがって，現段階で不足している情報，すなわち検索結果を絞り込む上で必要な情報をユーザから聞き出さなければならない．例えば，図 3 のように，検索結果の上位 10 件のオブジェクトがすべてお笑い芸人であった場合を考える．このとき，探している人物の職業について質問しても，ユーザにお笑い芸人と回答されるとクエリ修正後と修正前の情報量の差が小さい．なぜならば，検索結果の上位をお笑い芸人が独占しているということは，その時点におけるユーザのクエリに，お笑い芸人を探していることが容易に推測できるような単語が含まれている可能性が高いからである．結果として，検索結果は変化せず，質問の効果は薄い．この場合，もし検索結果の上位について，お笑い芸人の芸風が漫才，コント，落語などのように分散していれば，質問としてそのお笑い芸人の芸風を質問した方が，ユーザの回答によって得られる情報量が多いと考えられる．このように，検索結果のオブジェクトから，より情報量の多い回答を得られる質問を推定する．

そのため，質問がクエリ T_i のあいまい性を解消する度合を測るための指標を導入する必要がある．まず，質問 q_l の特徴語を n 個のクラスタ $C_k(k = 1, \cdots, n)$ に分類する．次に，オブジェクト o_j が持つ特徴語の中で，質問 q_l の回答になり得る特徴語のみを抽出し，クラスタ C_k によって分類する．抽出した特徴語の数（重複含む）を N_{jq_l} 個とし，そのうち，C_k に分類された特徴語の数を N_{jk} 個とすると，o_j が C_k に所属する確率は次の通りになる．

$$P(C_k|o_j) = \frac{N_{jk} + \alpha}{N_{jq_l} + \alpha n} \tag{2}$$

次に，検索結果の上位 m 件のオブジェクト $O_{T_i} = \{o_1, o_2, \cdots, o_m\}$ が，C_k に分類される確率 $P_m(C_k|O_{T_i})$ を，

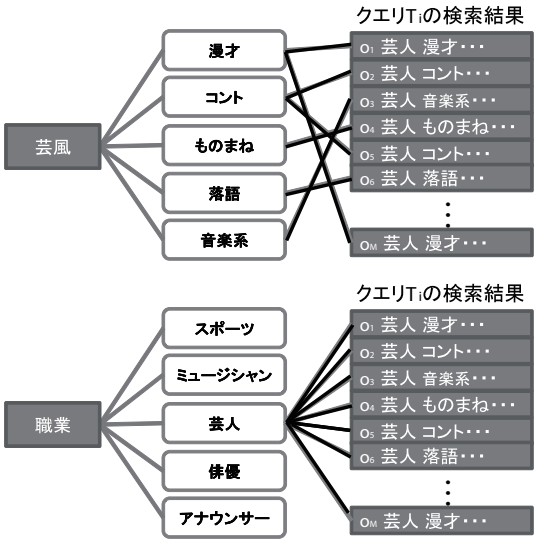

図 3　あいまい性を解消する質問の例

式 (2) を用いて求める．

$$\begin{aligned} P_m(C_k|O_{T_i}) &= \sum_{o_j} P(C_k, o_j|O_{T_i}) \\ &= \sum_{o_j \in O_{T_i}} P(C_k|o_j)P(o_j|O_{T_i}) \\ &= \frac{1}{|O_{T_i}|}\sum_{o_j} P(C_k|o_j) \end{aligned} \quad (3)$$

このとき，クエリ T_i の検索結果上位 m 件における，q_l のあいまい性解消度 $\mathrm{QDM_m}$[注7] は情報エントロピーを用いて次のように計算される．

$$\mathrm{QDM}_m(q_l, O_{T_i}) = -\sum_{C_k} P(C_k|O_{T_i})\log_2 P(C_k|O_{T_i}) \quad (4)$$

各 m について $\mathrm{QDM_m}$ を算出して QDM を計算する．ユーザは上位の検索結果を優先的に調べようとするため，検索結果上位のオブジェクト集合における QDM がより重要である．したがって，$\mathrm{QDM_m}$ に m が小さくなるほど，値が大きくなるような重み w_m を $\mathrm{QDM_m}$ に与える．

$$\mathrm{QDM}(q_l, O_{T_i}) = \sum_{m=1}^{M} w_m \mathrm{QDM}_m(q_l, O_{T_i}) \quad (5)$$

QDM の値が大きい質問 q_l は，クエリ T_i のあいまい性を解消できる有用な質問である．

5.3.2　質問自体の難易度

Web 検索をするにあたって，一般にユーザは欲しい情報をなるべく手間をかけずに入手したいと考える．そこで，本手法では各質問ごとに質問自体の難易度 D を 10 段階で設定し，ユーザになるべく簡単な質問 q_l を提示する．

5.3.3　ユーザの記憶に基づいた回答容易性

5.3.2 節で質問自体の難易度を設定したが，質問に対するユーザの回答容易性をこの数値のみで評価するのは不十分である．なぜならば，本研究ではユーザの記憶を想起させる情報検索を対象としており，その質問に答えられる知識を持っているかどうかはエピソードの記憶範囲によって大きく変動するからである．例えば，NBA 選手の「コービー・ブライアント」を検索で発見したいとする．もし，普段 NBA の試合を観戦しているユーザならば，その選手のプレイスタイルを問われたときに正確に回答を選択できる可能性が高い．逆に，NBA はおろかバスケットボールに興味が無いユーザからすると，バスケットボール選手のプレイスタイルに関する質問には答えられない．この例に見られるように，システムはユーザの持っている知識の範囲を考慮した上で最も適切な回答が得られる質問を提示すべきである．本手法では，直前までの質問 $q_1, q_2, \cdots, q_{l-1}$ におけるユーザの回答の有無，回答にかかった時間を考慮し，ユーザにとって回答が容易な質問を推定し，次の質問 q_l を提示する．

本研究では，コミュニティQ&A サイト「Yahoo!知恵袋」[注8] 内の文書について，質問の回答に対する共起度を調べた．コミュニティQ&A サイトでは，本研究で取り上げたようなあいまい性を含んだ記憶・知識の検索を，他のユーザとの対話により行うことができる．今回は，その中でも質問 q に回答できる情報を持ったエピソードを手がかりとして，オブジェクトの名前を問うている文書集合について考える．

この文書集合が，別の質問 q' にも答えられる確率 $P(q'|q)$ を計算する．$P(q'|q)$ はエピソードからオブジェクトを尋ねる質問において，質問 q と q' に対する答えがどれくらい共起しているかによって推定する．$P(q'|q)$ は質問 q', q に回答できる情報を含む文書数 $N(q, q')$，質問 q に回答できる情報を含む文書数 $N(q)$ を用いて下記のように定義できる．

$$P(q'|q) = \frac{N(q, q')}{N(q)} \quad (6)$$

$P(q'|q)$ が高いほど，質問 q に対する質問 q' の関連度が強い．すなわち，ユーザが持つエピソード記憶について，q に回答できる情報を持つならば，q' にも回答できるような情報を持っている可能性が高いと考えられる．

一方，ユーザが質問 q に答えられなかった場合については，q に関連する質問をするよりも質問 q とは無関係の質問をした方が，ユーザが持っている知識・記憶の範囲を詳しく知ることができるという点で有用と考えた．したがって，次に出題する質問は質問 q との関連性が薄いものを優先的に提示する．

以上の考えに基づいて，ユーザの記憶に基づいた回答容易性 AF を求め，l 回目に出題する質問 q_l を決定する．各質問について，ユーザから返答を受けた質問 $Q_{end} = q_1, q_2, \cdots, q_{l-1}$ を回答容易な質問集合 Q_f と回答困難な質問集合 $\bar{Q}_f$ に分離する．ここで，ユーザが選択肢「わからない」を選択した質問，および回答に一定時間以上かかった質問は $\bar{Q}_f$ に配置される．

$$\mathrm{AF}(q', Q_f, \bar{Q}_f) = \alpha \sum_{q \in Q_f} P(q'|q) - \beta \sum_{q \in \bar{Q}_f} P(q'|q) \quad (7)$$

AF の値が大きい質問 q_l は，ユーザの記憶に基づいて推測された，回答が容易な質問である．

(注7)：QDM は”Question Disambiguation Measure”の略である．

(注8)：http://chiebukuro.yahoo.co.jp/

5.4 代表オブジェクト集合の生成

1.3 節で述べたように，本手法ではユーザの記憶の想起をより促進させるために，具体的なオブジェクトを代表オブジェクトとして提示することによって，記憶を思い起こすための手がかりを与える．代表オブジェクトはランダムに提示するのではなく，ユーザにとってわかりやすくて選ぶのが容易な代表オブジェクトを提示すべきである．

まず，5.3.1 節で述べた方法で生成したクラスタ C_k と，オブジェクトが持つ特徴語に基づいて代表オブジェクトの候補をクラスタリングする．そして，各クラスタから選択肢としての代表オブジェクトを 1 つ選択し，選択肢の 1 つとして提示する．このとき，代表オブジェクトとしての適性を以下の指標に基づいて評価し，決定する．

(1) 代表オブジェクトの知名度
(2) 代表オブジェクトが持つ特徴の単峰性

(1) は代表オブジェクトがどれくらい有名であるかを評価する指標である．代表オブジェクトを選択肢として提示したとき，ユーザがその代表オブジェクトを知らなければ，選択肢としての意味を理解するのは難しく，正しい回答をシステムに与えることができない．したがって，ある程度有名なオブジェクトを代表オブジェクトとして選択する．

この問題は文書頻度 (df 値) を使うことにより近似できる．「現実世界で有名なオブジェクトは Web 空間においても有名である」という仮定のもと，オブジェクトの文書頻度に閾値を設け，閾値より低い文書頻度を持つオブジェクトを代表オブジェクトとして選択しないことにした．

一方，(2) は質問の選択肢となる代表オブジェクトが，選択肢としての提示の意図をユーザに正確に伝えられるかどうかを評価する．例えば，職業を問う質問に対して，選択肢として「アーノルド・シュワルツェネッガー氏」を提示すると，その選択肢が「俳優」としての側面を持っているのか「政治家」としての側面を持っているのか分からない．結果，ユーザがその選択肢を選択することで，システムに誤った情報を与えてしまうかもしれない．したがって，選択肢には，質問の回答として当てはまる特徴が多次元に及ばないようなオブジェクトを提示しなければならない．この性質を本研究ではオブジェクトの**単峰性**と呼んでいる．

次に，適切な代表オブジェクト選択のための計算手法について述べる．

代表オブジェクトの候補となるオブジェクト o_j が所属するクラスタは，o_j が持つ特徴語の中で最も多くの特徴語が所属するクラスタとし，C_k に所属するオブジェクトの集合を O'_k とする．

$$O'_k = \{\, o_j \mid \arg\max_k N_{jk} = k \,\} \tag{8}$$

C_k の代表オブジェクトは，O'_k に所属するオブジェクトの中から決定される．C_k における $o_j (\in O'_k)$ の代表オブジェクト r_k を，5.3.1 節で求めた $P(C_k|o_j)$ を用いて次のように導出する．

$$\mathrm{fm}(q_l, o_j) = -\sum_{C_k} P(C_k|o_j)\log_2 P(C_k|o_j) \tag{9}$$

図 4 システムの実行例

$$r_k = \arg\max_{o_j \in O'_k} \frac{1}{\mathrm{fm}(q_l, o_j)} \tag{10}$$

ここで，$\mathrm{fm}(q_l, o_j)$ は，質問 q_l に対するオブジェクト o_j の単峰性を求める関数で，値が小さいほど単峰性が強い．

この計算により，各クラスタから選択肢として適切なオブジェクトを代表オブジェクトとして 1 つずつ決定し，ユーザに提示する．

6. システムの実装

本研究では，Python で CGI を実装し，Web 上から検索機能を利用することができるシステムを構築した．図 4 はシステムの実行例である．オブジェクトの名前，Wikipedia の関連ページへのリンク，オブジェクトに関連するスニペット，オブジェクトの画像をクエリに対する検索結果として表示させた．

7. システムの評価

本章では，本手法によって構築したシステムによる実験とその結果について述べ，システムの評価を行う．

実験対象のオブジェクトとして，日本のプロ野球選手を用いた．人物が持つエピソードや情報からのオブジェクト検索は，正しいオブジェクトの発見が一層難しい．例えば，「ある日本人投手が，ルーキーイヤーから白星を重ねて新人王を獲得した」というエピソードが記憶に残っていたとする．このエピソードを持つ選手は複数存在するため，仮にユーザ自身によってクエリを正確に作成できたとしても，一選手を特定するのは通常の検索エンジンでは困難である．

7.1 評価概要

次の 2 つの項目について，評価を行う．

7.1.1 質問回答によるクエリ修正の妥当性

システムが提示する質問に回答していくことにより，検索対象の人物が検索結果の上位に登場するようになるのが望ましい．今回は，質問に対して常に正しい回答を返すと仮定して，ユーザが質問に答えていくことにより検索対象の人物のランキングがどのように推移するかを観察する．

表 1　検索 1,2 のタスク一覧

検索対象選手	初期クエリ T_0
中田翔	日本ハム，ドラフト 1 位，4 番
野村克也	ヤクルト，監督
福留孝介	メジャーリーガー，阪神，移籍
岩瀬仁紀	中日，守護神，スライダー
熊代聖人	西武 スイッチヒッター

初期クエリ	初期クエリ T_0 の情報	検索対象選手
捕手	ポジション	谷繁元信
代打の切り札	プレイスタイル	前田智徳
巨体	体格	清原和博
スキンヘッド	髪型	森本稀哲
穏やか	性格	湯舟敏郎

7.1.2　検索結果のあいまい性解消スコアの重要性

提示する質問の優先度を決定するにあたり，5.3.1 節で述べたあいまい性解消スコアを考慮することが有用であるかどうかを検証する．同一のタスク実行中の質問提示優先度について，あいまい性解消スコアを考慮した場合と考慮しなかった場合を試し，検索対象の人物のランキング推移を比較する．

7.2　評 価 実 験

上の 2 点を検証するため，以下の実験を行い，結果を記録した．

実験 1:

タスクとして，検索対象の野球選手名 A と単純な初期クエリ T_0 を与える．まず，初期クエリを入力し，初期クエリにおける検索結果から A を探し出し，A の順位を記録する．次に，提示された質問 q_0 の正しい回答をシステムに与えて再検索を行い，A の検索結果の順位を記録する．この動作を 3 回繰り返し，検索結果 A のランキングの推移を調べる．なお，顔の特徴や髪型など厳密な正解のない質問に対しては，選択肢の比較を行い，A の特徴を最も反映していると思われる選択肢を選択した．

実験 2:

ある質問 q の特徴語に含まれる単語 1 つを初期クエリ T_0 として与え，実験 1 と同様のタスクをあいまい性解消度を考慮した場合，考慮しない場合について実行する．このとき，質問 q が登場するまでに回答した質問の数を調査し，比較を行う．もし，あいまい性解消度設定における仮定が正しければ，質問 q の回答は初期クエリで表現されているため，質問 q をユーザに提示しても検索結果のあいまい性を解消させる効果は薄いはずである．したがって，質問 q は検索過程の序盤には登場しない方が検索効率が良いと予想される．

最後に，実験 1,2 のタスクの一覧を表 1 に示す．

7.3　実 験 結 果

7.3.1　実　験　1

タスクを実行した結果を表 2 に示す．T_0 の項目にある数字が，各タスクにおける対象選手の，初期クエリによる検索結果順位である．その後，3 問の質問解答による検索結果順位の推移を，修正 1，修正 2，修正 3 に示した．「解答に利用した情報」の項目では，各質問に対する回答として，ユーザがシステムに伝えようとした情報の概要を順に記した．「--」となっているものは，検索対象の野球選手が検索結果上位 100 件に現れなかったものである．また，実験結果をグラフ化したものを図 5 に示す．

図 5 から推察できるように，タスクによって多少の変動はあるものの，提示された質問に対してユーザが適切な回答を選択すれば，検索対象の野球選手が上位にランキングされる．初期クエリで入力されなかった情報が，追加の文字入力なしでシステムに伝えられたことがわかる．

特に，選手の性格，プレイスタイルに関する質問によるクエリ修正の効果が大きく，対象選手の検索結果の登場順位が大幅に向上した．選手の性格やプレイスタイルの情報は，5.1 節で挙げた質問集合の定義 (2) に当てはまり，クエリにするのが難しい，人によって表現が異なる，などの事情からあいまい性を含みやすい．したがって，本手法はクエリのあいまい性の解消に貢献していると考えられる．

一方，ポジションや活躍した時代に関する質問によるクエリ修正は効果が薄く，かえってクエリの質を低下させてしまったケースも見られた．これについては次のような原因が考えられる．

ポジションの場合，オブジェクトに関連付けられた文書内にオブジェクトとは無関係の特徴語が含まれており，それをシステムがオブジェクトそのものの属性と勘違いして抽出してしまったと思われる．例えば，文書内に「投手のクセを盗む」「一塁手のグラブをはじく内野安打を放つ」というフレーズが含まれていた場合，その選手のポジションに関する特徴語集合に本来のポジションとは無関係に「投手」や「一塁手」といった特徴語が含まれてしまう．このノイズが大きく，初期クエリ「捕手」の検索結果について，各選手が持つポジションの情報が分散されてしまったと考えられる．この問題を解消するには，選手とその選手のポジションを正確に関連付けする必要がある．

活躍した時代の場合も同様である．さらに，今回の実験において，活躍した時代に関する質問は，特徴語として「50 年代」「60 年代」「70 年代」といった，10 年ごとのざっくりとした単語のみを使用していた．そのため，各選手が活躍した時代をシステムが一意に特定できなかったと思われる．何らかの手法で質問の特徴語を充実させることができれば，クエリ修正の質が向上する可能性はある．

7.3.2　実　験　2

設定したタスクを実行した結果を 3 に示す．表中の数字が，質問 q がシステムによって提示されるまでの，ユーザの質問の回答数である．

髪型や体格，プレイスタイルについて初期クエリが入力されたとき，検索結果のあいまい性解消度を質問決定の際に考慮して，必要性の薄い質問を後回しにしている．したがって，あらかじめユーザから情報を与えられている事項についてユーザへ二重の問い合わせをするリスクを軽減させることができたと考えられる．

逆に，ポジション，性格についての情報を持つ初期クエリで

表 2　実験 1 の結果

検索対象	T_0	修正 1	修正 2	修正 3	解答に利用した情報
中田翔	26	24	24	23	パワーヒッター，左翼手，現役選手（2012 年現在）
野村克也	21	23	10	10	タフ，毒舌，1960 年代
福留孝介	49	36	38	26	走攻守そろった選手，外野手，自尊感情が強い
岩瀬仁紀	82	82	76	--	長期間の安定した実績，クローザー，投手
熊代聖人	29	29	34	29	顔の特徴，小柄な体格，坊主頭

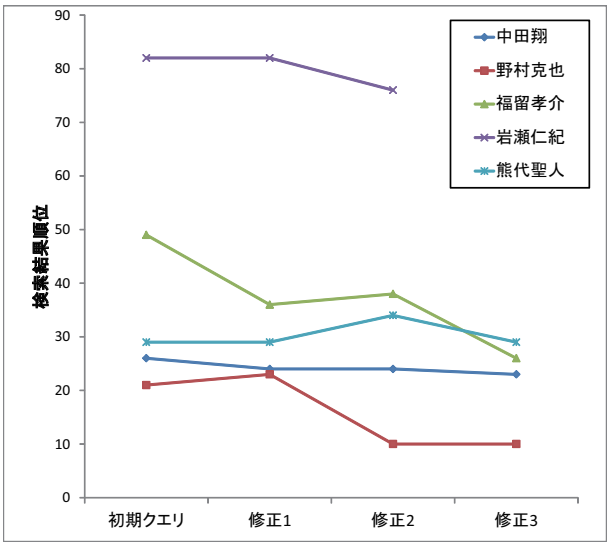

図 5　質問回答による検索結果順位の推移

表 3　実験 2 の結果

T_0	質問 q	QDM	q 提示までの回答数
捕手	ポジション	考慮	2
		無視	4
代打の切り札	プレイスタイル	考慮	16
		無視	11
巨体	体格	考慮	11
		無視	5
スキンヘッド	髪型	考慮	4
		無視	2
穏やか	性格	考慮	3
		無視	4

は逆の結果が得られた．すなわち，あいまい性解消度を考慮する場合の方が，初期クエリで入力した情報を直接問う質問を先に登場させている．このような結果が得られた原因として，それぞれ次の点が考えられる．

ポジションの場合は実験 1 での考察と同様で，オブジェクトに関連付けられた文書内にオブジェクトとは無関係の特徴語が含まれていることが原因と考えられる．

一方，性格については，初期クエリでその選手の性格を完全に把握できなかったことが原因だと思われる．ありとあらゆる側面を持つ人間の性格を 1 語で表現するのは困難であり，かつ性格のとらえ方は体格や髪型以上に人それぞれ異なるため，初期クエリの一致で単純に検索結果のあいまい性を評価する点に問題があったかもしれない．性格などの単語表現に対し，いかに検索精度を高めるかが今後の研究の焦点となる．

8.　まとめと今後の課題

本稿では，記憶があいまいになってしまったエピソードに関係するオブジェクトの名前を想起するために，システムが適切な質問および選択肢を提示して，ユーザがそれに回答することにより，候補を絞り込む方式で検索する手法を提案した．また，適切なクエリの修正を行うために，情報検索において生じるエピソードのあいまい性について 4 つに分類した．すなわち，「情報の欠損」，「単語の汎化」，「情報の誤り」，「印象に基づいた表現」の 4 つである．本稿ではこの中でも，情報の「欠損」，「汎化」に着目したクエリの修正を，選択肢である代表オブジェクトが内包するキーワード集合を用いて行った．さらに，質問と同時にユーザに提示する代表オブジェクトを，その代表オブジェクトの「知名度」や特徴の「単峰性」に基づいて選択することにより，ユーザにとって正しいと思われる回答を適切に選択しやすいインタフェースを設計した．

今後の課題としては，まず情報の「誤り」や「印象に基づいた表現」といったあいまい性を解消する方法の検討が挙げられる．今回提案した手法では，クエリ内に含まれる「誤り」を持った入力語を直接的には修正することができておらず，また，形容詞など「印象に基づいた表現」に該当する単語も，固有名詞や名詞などと同一視してクエリを修正しているため，検索の精度の面に疑問が残る．今後はこのような問題に対処できるような有効性・新規性のある手法の考案を目指す．

謝辞　本研究の一部は，文科省科研費基盤 (A)「ウエブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究」(24240013，研究代表者：田中克己）によるものです．ここに記して謝意を表します．

文　　献

[1] E. Brill and R. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pp. 286–293. Association for Computational Linguistics, 2000.
[2] G. コーエン（編）．「日常記憶の心理学」．サイエンス社.
[3] 稲川雅之, 大島裕明, 小山聡, 田中克己. オブジェクトを特定する語集合の web からの抽出. データベースと Web 情報システムに関するシンポジウム (DBWeb2007), 2007.
[4] 栗田多喜夫, 加藤俊一, 福田郁美, 坂倉あゆみ. 印象語による絵画データベースの検索. 情報処理学会論文誌, 33(11):1373–1383, 1992.
[5] 軽部孝典. 検索結果の対話評価に基づくリランキングインタフェース. 2008.
[6] 高野陽太郎（編）．「認知心理学 2　記憶」．東京大学出版会.
[7] 杉森絵里子. 「「記憶違い」と心のメカニズム」．京都大学学術出版会.
[8] 大島裕明, 小山聡, 田中克己. Web 検索エンジンのインデックスを用いた同位語とそのコンテキストの発見. 情報処理学会論文誌. データベース, 47(19):98–112, 2006.
[9] 北研二, 津田和彦, 獅々堀正幹. 「情報検索アルゴリズム」．共立出版.