Vol.2011-DBS-153 No.24
2011/11/3

類似性を考慮したマイナースポーツ検索手法の提案

服部祐基[†1]　灘本明代[†2]

近年のインターネットの普及により，インターネット上には様々な情報が存在している．その結果認知度や知名度が低い情報が見つけにくいという問題がある．そこで本研究では認知度や知名度が低く見つけにくい情報をマイナー情報とし，このマイナー情報を検索し提示する手法の提案を行う．本論文ではマイナー情報の検索の第一歩として検索対象ドメインをスポーツに絞り，Wikipedia からマイナースポーツを検索する手法の提案を行う．具体的には，ユーザの興味や関心のあるスポーツを入力とし，それと類似し且つマイナーであるスポーツをたとえ表現に基づく記事検索手法と関連性に基づく記事検索手法の２つの手法を用いて検索を行う．ユーザの興味や関心のあることから検索を行うことによってマイナー情報を見つけやすくすることを本研究の目的としている．

The Minor Sports Search System based on Similar Major Sports

Yuki Hattori[†1] and Akiyo Nadamoto[†2]

In this research, we propose the method of searching for the minor information which has low acknowledged-degree and low popularity. In this paper, the first step in searching for the minor information from the Internet, we propose the system which searches for the minor sports from Wikipedia. In our proposed system, user first inputs the name of the sports which he/she wants to know and the system searches for it which is similar to the user's input and minor sports. We propose two types of methods to extract minor sports from Wikipedia. One is the method of calculating relevance of content, the other is the method of based on example sentence.

1. はじめに

現在，インターネット上には大量かつ様々な情報が溢れている．その中で現在の検索システムでは，認知度，知名度が高く，良く知られている情報は比較的簡単に見つけることができるのに対し，認知度，知名度が低くあまり知られていない情報は見つけにくいという問題がある．その原因として認知度，知名度の低い情報は，個々の情報が少なく，どのように探せば良いかわからないという点が上げられる．しかしそのような情報の中にも有益な情報が多く含まれていると考えられる．例えば，スポーツであればメジャーなスポーツとマイナーなスポーツがある．我々が良く知っているスポーツというのはメディアなどによく取り上げられているメジャーなスポーツであることが多いが，それは多数あるスポーツのごく一部にすぎない．メジャーなスポーツに対し，メディアなどに取り上げられることが少ないマイナーなスポーツというのは，知るきっかけがほとんどないという問題がある．しかしそのようなマイナーなスポーツの中にも知らないだけで興味深いスポーツが多数存在すると考えられる．我々はこのようなマイナーな情報を知るためのきっかけを人々に与えることが必要であると考えた．そこで本研究では，認知度，知名度が低い情報をマイナー情報とし，このマイナー情報を検索し提示するシステムの提案を行う．具体的には，マイナー情報を検索する手法としてユーザがある程度興味のあるマイナー情報を提示する事が必要であると考え，ユーザの興味や関心のあることを入力とし，その入力されたキーワードと類似し且つマイナーである情報の検索を行う．ユーザの興味や関心のあることを入力とすることで，見つけにくいマイナー情報を探しやすくできる．これにより，自分の知らなかった情報や気づいていない情報を取得することができ，新たな知識や興味の発見が可能になる．本論文ではマイナー情報の検索の第一歩として，検索対象ドメインをスポーツに絞り，ユーザが入力した興味や関心のあるスポーツを用いて，そのスポーツに類似しているが，認知度，知名度が低いマイナーなスポーツを Wikipedia から検索し提示するマイナースポーツ検索の提案を行う．具体的には，我々の提案する「たとえ表現に基づく検索」と「関連性に基づく検索」の 2 つの検索手法とフィルタリング手法を用いてマイナースポーツを検索する．マイナースポーツ検索の大まかな流れを以下に，システムの流れを図 1 に示す．

†1 甲南大学大学院 自然科学研究科
Konan University
†2 甲南大学 知能情報学部
Konan University

(1) ユーザは興味や関心のあるスポーツ名を入力する．
(2) 「たとえ表現に基づく検索」として，ユーザの入力したスポーツ名をたとえている Wikipedia の記事を取得する．
(3) 「関連性に基づく検索」として，ユーザの入力したスポーツ名の記事と関連の高い Wikipedia の記事を取得する．この時，関連の高い Wikipedia の記事はリンク解析と類似度計算を行うことにより決定する．
(4) （2）と（3）で取得した記事に対して編集回数，編集人数が少ない物をマイナー情報と仮定し，抽出する．これをマイナー情報の候補とする．
(5) （4）で取得したマイナー情報の候補からスポーツ名のフィルタリングを行う事により，スポーツの記事のみを抽出し，それらをマイナースポーツとして提示する．

以下，第 2 章に関連研究を，第 3 章にマイナースポーツの定義をし，第 4 章にマイナースポーツの検索手法，第 5 章にスポーツ名のフィルタリング手法について提案し，第 6 章に提案手法を用いて作成したプロトタイプシステムについて，第 7 章に評価実験，第 8 章でまとめと今後の課題を述べる．

2. 関 連 研 究

大島ら[1] は文書群をクエリとし，それと似ているが異なる文書の検索を提案している．我々の提案はクエリから，類似し異なるものを検索する点においては似ているが，本研究では一つの文書をクエリとし，それと類似しているが異なり，かつ認知度，知名度が低いものを見つけてくる点においては異なる．

情報検索の際に Google や Yahoo!などを用いると検索結果が下位に出てくるページというのは見つけるのが難しくマイナーな記事となるが，そのようなページにおいても有益なページは含まれている．そこで近年そのようなページを上位に上げるためのリランキング手法[2],[3] が多く提案されている．これらリランキング手法を用いた多くの情報検索の研究ではユーザの嗜好に合った情報を検索できるように，Web 閲覧履歴や検索結果に対するユーザの操作などを用いている場合が多い．ランキングが下位のページを上位に上げることにより，マイナーな記事を取得できる．それに対してマイナーな記事を取得するところでは類似しているが，本研究では対象ページを Wikipedia に絞り，またパーソナライズを目的としていない点が異なる．また竹原ら[4] は Web 検索エンジンでページのランク付けのためによく用いられているリンク構造を解析した手法では，マイナーなページが表示されにくいという点を問題にしている．リンク構造を解析する手法では，より多くの他のページからリンクされているページがより上位にランク付けされるという点が挙げられる．より多くの他のサイトや人に知られている有名なページが検索結果の上位に表示されるということになる．ユーザは有名なページを欲しているとは限らず，マイナーだがそのユーザにとって有用であるページが提示されないという問題を解決するために，Blog サイトの解析に基づいた検索システムの構築手法について提案している．マイナーで有益なページを検索する点においては非常に似ているが本研究では Blog ではなく Wikipedia を用い，ドメインをスポーツに絞っている点において異なる．

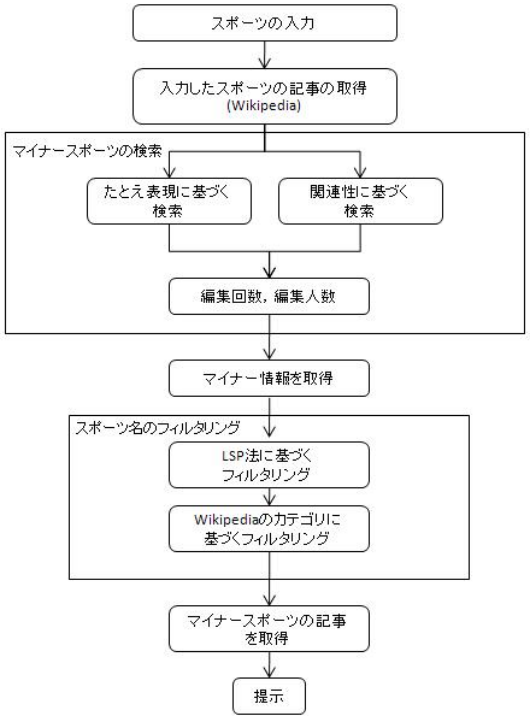

図 1 システムのフロー
Fig. 1 System Flow

Vol.2011-DBS-153 No.24
2011/11/3

3. マイナースポーツの定義

本章では Wikipedia を用いてマイナースポーツを検索するにあたり，まずマイナースポーツの定義を行う.「マイナー」という言葉の意味は大辞林では「[1] 規模や重要度が小さいさま．[2] あまり知られていないさま．有名ではないさま.」となっている．マイナースポーツの場合，あまり知られていないスポーツや競技人口の少ないスポーツを示す場合が多いが，本研究では前者のあまり知られていないスポーツをマイナースポーツとして，そのマイナースポーツの検索手法を提案する．この時，あるスポーツがマイナーであるかメジャーであるかはユーザによって異なるという問題がある．そこで本研究ではこのマイナースポーツの定義を行うために実験を行った．被験者 13 名に対し，ある 150 個のスポーツ名を提示し，各スポーツに対して以下 1〜3 の 3 段評価をしてもらった．

１：　知っている，だいたい知っている．
０：　名前は知っている（名前は知っているが何をするのかわからない）
-１：　知らない（名前すら聞いたことがない）

この実験の結果を**表 1** に示す．ここでの平均とは，そのスポーツにおける評価の結果の合計を人数で割ったものである．よって’１’の値に近ければ知っている人が多くなり，’-１’に近ければ知っている人が少ない，つまりはマイナスの評価値のスポーツはマイナースポーツとなる．本研究では，ここで決定したマイナースポーツ及びメジャースポーツを教師データとして，マイナースポーツの検索手法の提案を行う．

4. マイナースポーツの検索手法

本論文ではマイナー情報検索のはじめの一歩として，Wikipedia からマイナースポーツを抽出する事を行う．図 1 に示すようにマイナースポーツを抽出する手法として,「たとえ表現に基づく検索」と「関連性に基づく検索」の２つの検索手法により入力したスポーツの類似した記事を取得し,「編集回数，編集人数」として Wikipedia の記事の編集回数，編集人数が少ない記事をマイナースポーツの候補とし取得する．以下それぞれの手法について説明していく．

4.1 たとえ表現に基づく検索

マイナースポーツの Wikipedia の記事の特徴として，マイナースポーツを説明するために他の有名なスポーツに例えて表現している場合が多い．この特徴を我々は「たとえ表現」と呼び，この手法を用いて，ユーザの入力したスポーツ名からマイナースポーツを抽出する．実際に 3 章におけるマイナースポーツの教師データの 30 % がたとえ表現を用いている． 例えばマイナースポーツの教師データの一つである,「セパタクロー」というマイナースポーツの Wikipedia の記事では,「足のバレーボール」や「ルールはバレーボールと似て」というようにバレーボールに例えて表現がされている．また「ローンボウルズ」では「ボウリングの前身」と表記されており，このようにたとえ表現を用いた文を読むだけでどのようなスポーツであるかおおよそのイメージをすることができる．そこで「〜と似て」や「〜の前身」のようなたとえ表現を用いてマイナースポーツの取得を行う．**表 2** にたとえ表現の例と表記例を示す．たとえ表現を表す語は,「〜に類似」や「〜に似て」などの類似を表現している場合や,「〜の原型」,「〜の前身」,「〜の派生」などの起源を用いて表現している場合,「〜と…を組み合わせた」や「〜と…を融合した」などの複数のスポーツを組み合わせて表現している場合,「〜で行う…」などの特定の場所や特別な道具を用いて表現されている場合などがある．よってユーザが入力したスポーツに対して，たとえ表現である表 2 の 8 パターンを拡張クエリとして付加し検索を行う．例えば，入力が「バレーボール」であった場合には,「バレーボールに類似」や「バレーボールと似て」,「バレーボールの原型」のように拡張クエリを付加し検索を行う．さらに「〜と良く似て」などのように強調語を用いて表記されている場合もあるため,「〜と似て」と「〜と良く似て」のように強調語がある場合

表 1　実験の結果
Table 1　Result Of Experiment

No	スポーツ名	平均	評価値 1(人)	評価値 0(人)	評価値-1(人)
1	フットサル	1	13	0	0
2	ハンドボール	1	13	0	0
3	水球	0.8462	11	2	0
4	ポロ	-0.6290	2	0	11
5	ロデオ	0.6154	8	5	0
6	カポエイラ	0.2308	5	6	2
7	フェンシング	1	13	0	0
8	バイアスロン	-0.3850	3	2	8
9	アーチェリー	1	13	0	0
10	ダーツ	1	13	0	0
11	ラクロス	0.8462	11	2	0
12	スカッシュ	0.7692	10	3	0
13	リュージュ	-0.5380	2	2	9
14	スノーカイト	-0.9230	0	1	12
15	カヌーポロ	-0.9230	0	1	12
：	：	：	：	：	：

Vol.2011-DBS-153 No.24
2011/11/3

表 2 たとえ表現と表記例
Table 2 The Expression Of Example and The Notation Example

たとえ表現の例	表記例（記事）
～に類似	サッカーに類似して（バンディ）
～と似て	バレーボールと似て（セパタクロー）
～と…を組み合わせた	クロスカントリーとライフル射撃を組み合わせた（バイアスロン）
～と…を融合	チェスとボクシングを融合したスポーツ（チェスボクシング）
～の原型	ゲートボールの原型（クロッケー）
～の前身	ボウリングの前身（ローンボウルズ）
～の派生	テニス等の派生（スカッシュ）
～で行う…	自転車で行うサッカー（サイクルサッカー）

も拡張クエリとして用いる．また，複数のスポーツを組み合わせている場合，例として「A と B を融合」の場合には,「バレーボールとサッカーを融合」と「サッカーとバレーボールを融合」のように 2 通りのクエリが考えられるため，これも考慮した拡張クエリとする．

4.2 関連性に基づく検索

すべてのマイナースポーツが Wikipedia 上でたとえ表現を用いているとは限らないため，たとえ表現に基づく検索だけではマイナー情報を取得しきれない．実際にマイナースポーツの教師データでは 30%がたとえ表現を用いており，残り 70%はたとえ表現を用いていない．そこで我々はさらにユーザの入力したスポーツと関連性の高いスポーツの記事を記事間のリンクグラフと類似度計算を用いて抽出する．以下の手順により関連性に基づくマイナースポーツの候補を抽出する．

(1) ユーザの入力したスポーツを基準としてリンクグラフを作成する（図 2 参照）．ここでユーザの入力したスポーツの記事を基準ノードと呼ぶ．

(2) 基準ノードにリンクしている記事や双方向にリンクしている記事（以下，2 つあわせてインリンクノードと呼ぶ）は基準ノードに対して何らかの関係を持っていると考えられる．よってインリンクノード以外の記事つまりは基準ノードからのリンクのみの記事を削除する．基準ノードからのリンクのみの記事は，マイナーな記事である可能性が低く，また地域名や組織名等の様々なスポーツ以外の記事が抽出されたため，今回は考慮しない．

(3) インリンクノード内で，基準ノードを示すアンカー文字列が 1 回のみ出現の記事はあまり関連が深くないと考え削除する．

(4) インリンクノードのタイトルに地域名，人名，組織名が含まれている場合はスポーツである可能性が低いため削除する．

(5) 関連の高い記事同士は類似度が高くなると考えられるため，基準ノードとインリンクノード間で類似度計算を行い，類似度がある閾値 α 以上のノードをマイナースポーツの候補とする．ここで本研究では以下のコサイン類似度を用いて，類似度 $cos(x, y)$ を求める．

$$\cos(x, y) = \frac{\sum x_i \cdot y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}} \quad (1)$$

ここで，x を基準ノード，y をインリンクノード，そして x_i は x における名詞 i の出現頻度，y_i は y における名詞 i の出現頻度である．ここで実験により閾値 α を 0.35 とし閾値 α 以上の記事を入力したスポーツの記事と関連の高い記事とし取得する．

例えば図 2 ではノード C，G は基準ノードからのリンクのみのため削除する．またノード D はインリンクノードであるが，基準ノードを示すアンカー文字列がノード D 内で 1 つだけのため削除する．ノード F はタイトルに人名が含まれているため削除する．ノード H は類似度が 0.2 であり閾値 0.35 以下であるため削除する．これにより図 2 の場合マイナースポーツの候補となる記事はノード A,B,E となる．

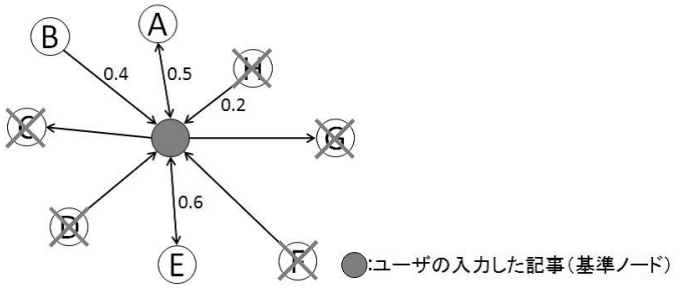

図 2 リンクグラフ
Fig. 2 Link Graph

4.3 編集回数，編集人数

ここではたとえ表現に基づく検索と関連性に基づく検索から得られた記事からマイナースポーツの記事の取得を行う．マイナースポーツの記事を抽出するために，本研究では Wikipedia の編集回数と編集人数に着目する．メジャーなスポーツは知っている人も多く，

Vol.2011-DBS-153 No.24
2011/11/3

Wikipedia の記事では編集する回数や人数が多くなるが，マイナーなスポーツは知っている人が少ないため Wikipedia の記事では編集する回数や人数が共に少なくなる場合が多いと考えられる．そこで，3 章で決定した 150 の教師データを用いて，メジャースポーツとマイナースポーツの Wikipedia の記事の編集回数と編集人数を調べた．その結果を**表 3** と図 3，図 4 に示す．表 3 の No1 から No4 はメジャースポーツであるが，これらは編集回数・編集人数が多くなっている．それに対し No5 から No8 はマイナースポーツであるが，編集回数・編集人数が少なくなっている．また，図 3，図 4 では X 軸にそれぞれ編集回数と編集人数を，Y 軸には 3 章の実験で行った評価の平均としている．ここでは，編集回数，編集人数共に回数，人数が増えると，評価の平均の’ 1 ’ である知っている人が多くなり，回数，人数が減ると共に，評価の平均の’- 1 ’ である知らない人が多くなっていることがわかった．これにより，メジャーなスポーツは編集回数と編集人数が多く，マイナーなスポーツは編集回数，編集人数が共に少なくなることが判明した．この結果より，たとえ表現に基づく検索と関連性に基づく検索により得られた記事の編集回数，編集人数を調べ，共に閾値 β 以下の記事はマイナースポーツであるとする．ここで実験より閾値 β は編集回数を 180，編集人数を 80 とする．この手法より得られた記事の中にはスポーツ以外の記事も含まれているため，ここではマイナースポーツの候補とする．

表 3 編集回数と編集人数
Table 3 Editing Number of Times And People

No	スポーツ名	編集回数	編集人数
1	野球	1087	507
2	サッカー	880	411
3	卓球	874	389
4	柔道	555	267
5	居合道	153	70
6	ペタンク	95	62
7	セパタクロー	76	46
8	ペサパッロ	28	24

5. スポーツ名のフィルタリング

4 章で得られたマイナースポーツの候補の中には，スポーツ選手やスポーツ用具などのスポーツ名でない記事も取得されている．よってフィルタリングを行いマイナースポーツの候補からスポーツの記事のみを抽出することを提案する．ここで (1)Wikipedia のカテゴリに基づくフィルタリング，(2)LSP 法に基づくフィルタリングの 2 つの方法を用いてスポーツ記事のフィルタリングを行う．

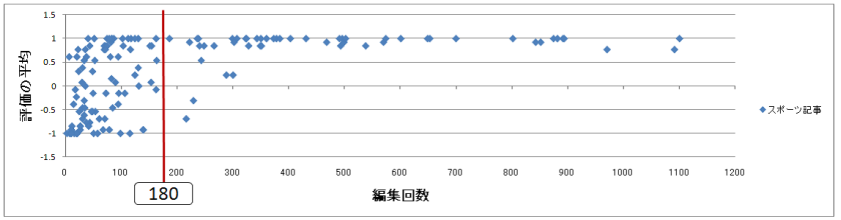

図 3 編集回数と評価平均の関係
Fig. 3 The Relations With The Editing Number of Times and The Average Of The Evaluation

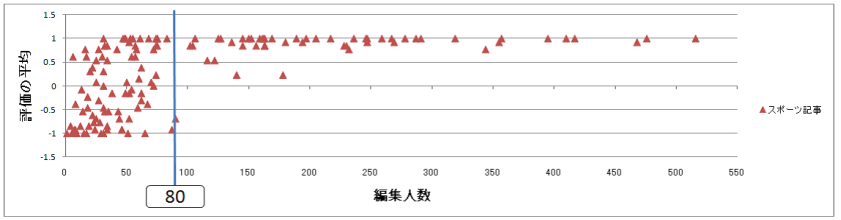

図 4 編集人数と評価平均の関係
Fig. 4 The Relations With The Editing Number of People and The Average Of The Evaluation

5.1 Wikipedia のカテゴリに基づくフィルタリング

Wikipedia の記事はすべてカテゴリ分けがされている．我々はこのカテゴリに注目し，カテゴリを用いてマイナースポーツの候補からスポーツ記事の抽出を行う．本研究ではスポーツと判断するカテゴリを「スポーツ」,「スポーツ競技」,「オリンピック競技」,「団体競技」とする．カテゴリ構造は階層構造になっており，一つの記事に対し複数のカテゴリがある場合もある．例えば,「サッカー」の記事の 1 階層上のカテゴリは「サッカー」,「オリンピック競技」である.「オリンピック競技」からサッカーはスポーツであると判断することができる．しかし,「セパタクロー」のカテゴリは「セパタクロー」となっている．この場合セパタクローは本来スポーツであるがスポーツと判断することができない．そのためこのような場合は記事のカテゴリのもう一つ上の上位 2 階層のカテゴリを見ることとする.「セパタ

クロー」の上位１階層のカテゴリは「セパタクロー」だが，その上位２階層目のカテゴリを見ると「球技」,「タイのスポーツ」,「団体競技」となり,「団体競技」からセパタクローはスポーツと判断することができる．多くのスポーツの記事を調べた結果上位２階層のカテゴリまででスポーツかどうかを判断できる事がわかったため，我々は２階層上のカテゴリまでを対象としてカテゴリによるフィルタリングを行う．ここで，スポーツ名に対しては，この方法によりスポーツと判断することができたが，スポーツの道具や用語などスポーツに深く関係のある記事にも上位２階層のカテゴリにスポーツと判断するカテゴリが含まれてしまう問題があり，これらを削除する必要がある．例えば「ボール」のカテゴリには「球技」が含まれており，球技の上位のカテゴリには「スポーツ競技」とスポーツと判断するカテゴリが含まれていたため，ボールはスポーツと判断されて取得してしまう問題がある．そこでスポーツの道具や用語などを削除するために，LSP 法に基づくフィルタリング手法を用いる．

5.2 LSP 法に基づくフィルタリング

Wikipedia のカテゴリを用いて取得されたスポーツの道具や用語などのスポーツ以外の記事を削除するために，本研究では中山ら[5),6)] の提案している LSP 法を用いる．中山らは Wikipedia 内の記事における文法に着目し，記事のリード部分を重要文と見なして解析する手法である LSP 法を提案している．Wikipedia 内の文章は多くの場合冒頭文が is-a 関係にある．この関係を利用し単語の上位概念を抽出することを行う．具体的には，記事の１文目の最後に出てきた名詞をその記事の上位概念とする．この手法を用いると例えば「ボール」の Wikipedia の記事の１文目は「ボール（ball、玉）はゲーム（球技や遊戯）などに使う球形の用具」となっており，最後に出てきた名詞である「用具」がボールの上位概念となる．この「用具」はスポーツ名ではない．そこで最後に出てきた名詞が「スポーツ」や「スポーツ競技」,「オリンピック競技」,「球技」,「武道」,「国技」などのスポーツと判断できる単語が上位概念として取得された場合には，その記事がスポーツの記事であるとし取得する．ここで「スポーツのこと」などのように表記されていた場合，最後の名詞は「こと」になる．これではスポーツである記事がスポーツと判断できず，削除されてしまう問題がある．そのため最後に出てきた名詞が「こと」,「事」,「一つ」,「ひとつ」,「名称」などであった場合は，一つ前の名詞を上位概念とする．先ほどの「スポーツのこと」という場合には，最後の名詞が「こと」になっているため，一つ前の「スポーツ」が上位概念となり，スポーツと判断できるためスポーツの記事とする．

以上２つのフィルタリング手法を用いて得られた記事をマイナースポーツの記事としてユーザに提示する．

6. プロトタイプシステム

我々の提案手法を用いてプロトタイプシステムを作成した．プログラミング言語には Ruby を，ユーザインターフェースには CGI を用いた．システムの起動画面と出力例を図 5 と図 6 に示す．まずユーザは好きなスポーツを入力し，決定をクリックするとシステムはユーザの入力したスポーツと類似するマイナースポーツを検索する．そして得られたマイナースポーツのリストを表示する．また Yahoo!検索により得られたサマリを同時に載せることでそのスポーツの大体の意味がわかるようになっている．ここでユーザがリストから一つのマイナースポーツを選択するとその Wikipedia の記事を見ることができる．これによりそのスポーツについてより詳しく知ることができるようになる．

図 5 プロトタイプシステムの画面
Fig. 5 Display of Prototype System

7. 実　　験

7.1 編集回数，編集人数によるたとえ表現の評価実験

たとえ表現で得られた記事がマイナー情報であるかを検証するために Wikipedia の編集回数と編集人数を用いて実験を行った．ここでは編集回数と編集人数を用いて，たとえ表現の評価を行う．図 3 と図 4 よりマイナー情報と判定する編集回数と編集人数の閾値をそれ

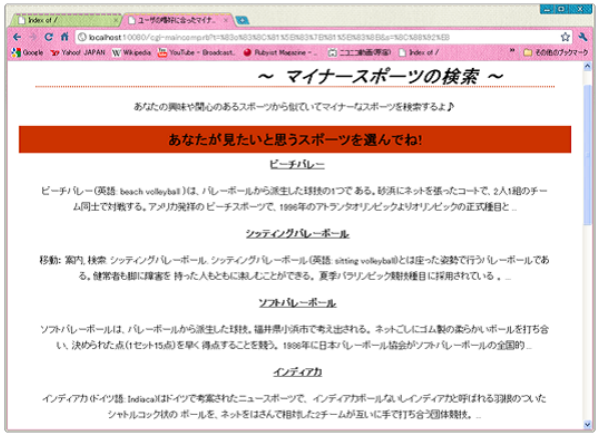

図 6 出力画面
Fig. 6 Output

表 4 たとえ表現の検証
Table 4 The Inspection Of The Expression Of Example

入力	取得されたマイナースポーツ	編集回数	編集人数
サッカー	バンディ	38	28
	フットサル	299	152
	サイクルサッカー	26	12
	電動車いすサッカー	33	21
	フリースタイルフットボール	5	3
バレーボール	セパタクロー	77	47
	インディアカ	43	25
	シッティングバレーボール	41	24
	ソフトバレーボール	20	14
	ビーチバレー	135	80
バスケットボール	ストリートバスケットボール	44	24
	ビーチバスケットボール	6	5
	ウォーターバスケットボール	6	6
	カヌーポロ	79	47

ぞれ 180，80 とし，提案手法であるたとえ表現から取得された記事のマイナーの度合いの評価を行う．任意の 3 種類のメジャースポーツを入力とした．

結果と考察

たとえ表現より取得されたスポーツ名とその編集回数と編集人数との関係を**表 4** に示す．結果よりたとえ表現から得られたスポーツの記事の多くがそれぞれの閾値を下回る結果となった．たとえ表現により取得されたマイナースポーツ記事のうち，フットサルやビーチバレーなどある程度有名であるスポーツもたとえ表現が用いられていることがわかった．しかしながら，フットサルは編集回数，編集人数とも閾値以上となっている．その為，編集回数，編集人数によりマイナースポーツの候補から削除されるため，提案手法は有用である事がわかった．

7.2 システムの評価実験

本節ではプロトタイプシステムを用いて，本システムの有用性を示す．本システムを評価する尺度として，再現率と適合率，F 値を用いる．再現率を求める際の正解データとして，3 章の実験から得られたマイナースポーツの教師データを用いる．ここで，マイナースポーツの教師データのみで分類しただけでは正解データが少ないため，これに合わせて人手で集めたマイナースポーツ 35 個を正解データとして加えた．実験データは，メジャースポーツの中から 7 つのメジャースポーツを入力データとして本提案手法を用いてマイナースポーツ検索を行った．

結果と考察

結果を**表 5** に示す．表 5 よりスポーツによって値に差が出ているが，再現率，適合率，F 値の平均は良い結果となった．しかしゴルフとサッカーの再現率が他のスポーツに比べとても低い値になっている．この原因として 2 点考えられる．1 点目として，4.2 節の「関連性に基づく検索」においてタイトルに人名や組織名を含んでいる記事を削除したが,「スナッグゴルフ」のように組織名と同じスポーツ名の場合，形態素解析器が「スナッグゴルフ」を組織名と判断したためマイナースポーツとして抽出されなかった．そして 2 点目として，5.2 節の「LSP 法に基づくフィルタリング」において，記事の 1 文目の最後に出てきた名詞をその記事の上位概念とし，最後に出てきた名詞が「スポーツ」や「スポーツ競技」,「オリンピック競技」などであれば，その記事がスポーツの記事であるとし取得している．しかし 1 文目が複文になっていた場合には正確に上位概念を取得することができないという問題がある．例えばサッカーのマイナースポーツの正解データである「フリースタイルフットボール」において，記事の 1 文目は「フリースタイルフットボール（Freestyle football）は、サッカーから派生したスポーツでサッカーボールを用いてパフォーマンスを行なうものである。」と複文で構成されている．この記事ではパフォーマンスが上位概念となりスポーツと判断されずに取得できなかった．このように LSP 法に基づくフィルタリングの問題点が明らかに

なった．

表 5 実験結果
Table 5 Experimental result

スポーツ	再現率	適合率	F 値
サッカー	50%	71%	59%
ホッケー	57%	75%	65%
野球	67%	100%	80%
バレーボール	78%	88%	82%
テニス	57%	100%	73%
ゴルフ	40%	100%	57%
バスケットボール	100%	100%	100%
平均	64%	91%	75%

8. まとめと今後の課題

本論文では見つけにくいマイナー情報をユーザの興味や関心のあるスポーツからマイナースポーツを検索する手法の提案を行った．マイナーなスポーツを検索し提示することを行うことで，ユーザの新たな知識の取得が可能である．また本システムを用いることで少しでもマイナースポーツの発展に貢献できることを期待している．今後の課題は以下の通りである．

- スポーツ名のフィルタリングの改善
  LSP 法に基づくフィルタリング手法には 1 文目の最後の名詞が上位概念となるが，記事により 1 文目が複文になっている場合がある．そのような記事の場合，本来スポーツの記事であってもスポーツと判断されないため取得できないという問題がる．複文などにも対処していく必要がある．
- 出力結果の改善
  本システムでの出力では，マイナースポーツのリストと Yahoo!検索により得られたサマリのみを提示しているが，マイナースポーツの画像を一緒に載せることでそのスポーツをよりイメージしやすかったのではないかと考えた．しかし画像検索によりマイナースポーツを検索した結果，マイナーであるために画像が取得できないということがわかった．そのため Wikipedia の記事に画像が使われていた場合にはその画像も一緒に提示することも考えている．またマイナーなスポーツにもそのスポーツをメジャーとして考えている人もいる．そのメジャーとしている人が集まっている SNS などのコミュニティでの話題などを載せることで，そのスポーツのことがより理解し易くなるのではないかと考えている．
- その他の分野への応用
  本研究ではスポーツのみを対象として行っているため，今後は他の分野にも応用していくことを考えている．また Wikipedia のみを用いてマイナー検索をしていたので今後は Wikipedia だけでなく，SNS なども用いていきたい．
- 評判情報の付加
  出力されたスポーツに対してより興味を持ってもらうために，評判情報を付け加えることを考えている． SNS のコミュニティを用いることで，そのコミュニティ内でのそのスポーツに対しての評判を抽出し，ユーザに提示する．これによりユーザがそのスポーツに対してより興味や関心が増すと考えられる．

参 考 文 献

1) 大島裕明，小山 聡，田中克己：文書群をクエリとした ”似て非なる ”文書の検索，日本データベース学会論文誌（DBSJ Letters）， Vol.5, No.1, pp.121–124 (2006).
2) 山本岳洋，中村聡史，田中克己：編集操作を用いたウェブ検索結果の最適化，データ工学ワークショップ (DEWS2007)， Vol.17 (2007).
3) 松岡研二，范 薇，小柳佑介，渡邉豊英：Web ページに対するユーザの適・不適の評価を用いた検索結果のリランキング，*DEIM Forum 2010* (2010).
4) 竹原幹人，中島伸介，角谷和俊，田中克己：Web 情報検索のための Blog 情報に基づくトラスト値の算出方式，*DBSJ Letters*, Vol.3, No.1.
5) 中山浩太郎，原 隆治，西尾章治郎：自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジー自動構築，日本データベース学会論文誌， Vol.7, No.1, pp.67–72 (2008).
6) Nakayama, K., Hara, T. and Nishio, S.: Wikipedia Mining - Wikipedia as a Corpus for Knowledge Extraction, *Proceedings of annual Wikipedia Conference(Wikimania)* (2008).