ユニークネス尺度を用いたメタサーチエンジン

山下　陽子† 　森下　真一†† 　増永　良文†††

† お茶の水女子大学人間文化研究科　〒 1112–8610 東京都文京区大塚 2–1–1
†† 東京大学大学院新領域創成科学研究科　〒 113–0033 東京都文京区本郷 7–3–1
††† お茶の水女子大学理学部　〒 1112–8610 東京都文京区大塚 2–1–1
E-mail: †yoko@dblab.is.ocha.ac.jp, ††moris@k.u-tokyo.ac.jp, †††masunaga@is.ocha.ac.jp

あらまし　近年, WWW 上の情報量の増加に伴い, ユーザが情報を検索するための, さまざまな戦略を持ったサーチエンジンが提供されている. これらサーチエンジンの検索結果を統合することで, 総合的なランキング結果を返すメタサーチエンジンがある. しかし, 既存の手法では, あるサーチエンジンでは上位に, 他のサーチエンジンでは下位にランキングされている特徴のあるページが, 最終的に下位にランキングされ見過ごされる場合が多い. 本研究では, さまざまなサーチエンジンの差異を活かして, 特徴のあるページを発見する, メタサーチエンジンを提案する. 特徴のあるページを見つける評価基準として, ユニークネス尺度を定義し, システムを実装した. また, WWW における実験を通して, ユニーク度の高いページは Yahoo! に載っていない傾向にあり, 新しい情報を掲載したページが多いことが確認された.
キーワード　Web とインターネット, 情報統合, 情報検索

A Metasearch Engine Using the Measure of Uniquenes

Yoko YAMASHITA†, Shinichi MORISHITA††, and Yoshifumi MASUNAGA†††

† Graduate School of Humanities and Sciences, Ochanomizu University,
2–1–1 0tsuka, Bunkyo-Ku, Tokyo, 112–8610 Japan
†† Graduate School of Frontier Sciences, University of Tokyo,
7–3–1 Hongo, Bunkyo-ku, Tokyo, 113–0033 Japan
††† Faculty of Science, Ochanomizu University,
2–1–1 0tsuka, Bunkyo-Ku, Tokyo, 112–8610 Japan
E-mail: †yoko@dblab.is.ocha.ac.jp, ††moris@k.u-tokyo.ac.jp, †††masunaga@is.ocha.ac.jp

Abstract　The rapid growth of data available in the WWW has been demanding a wide variety of search engines that facilitate the task of mining informative web pages. However, divergence among ranking strategies of various search engines may yield serious disagreement among ranks even against the same query, motivating the development of meta-search engines that aggregate multiple ranks into a single list. Although there have been a lot of studies on aggregation procedures, most of them are likely to make little of niche and unique web pages that occupy high positions in a couple of ranks but low places in the others. In this paper, we propose three different kind of measures to define such unique and niche pages. Tests uncover that web pages ranked high according to these measures are typically fresh pages that are not listed in Yahoo!.
Key words　Web and Internet, information fusion, information retrieval

1. 背景と目的

近年, WWW(World Wide Web) 上のデータ量は爆発的な勢いで増加している. この膨大なデータの中から自分の必要とする情報を, リンクだけを辿って見つけ出すのは困難になってきており, 自分の必要とする情報のキーワードのみから関係するページを検索するサーチエンジンの重要性がますます高まっている.

現在, 多くのユーザに使用されているサーチエンジン数は 10 を超えている. これらのサーチエンジンはそれぞれ, データ収集の方式, 索引付けの方式, ランキング方式の戦略が異なっている. そのため, 同一検索キーワードを問い合わせても, サーチエンジン毎に検索結果には違いが生じる. それぞれのサーチエンジンは, 他との差別化を図り独自性のある検索結果を提供している.

そこで複数の検索結果を統合することで, 偏りの少ない総合的なランキング結果を返すメタサーチエンジン [3] [6] の研究・開発が盛んになっている. また, WWW 上のデータ量が膨大なた

め, 1 つのサーチエンジンでは全体をカバーすることが不可能になっている. メタサーチエンジンは, 複数のサーチエンジンの検索結果を用いるため, WWW 上を広範囲カバーすることができる利点がある.

既存のメタサーチエンジンでは, 複数のサーチエンジンで上位にランキングされているページを, 重要度の高いページと評価してランキングする, いわゆる統合手法 [2] [4] [7] を工夫している. しかし, 既存のメタサーチエンジンでは, あるサーチエンジンでは上位に, 他のサーチエンジンでは下位にランキングされている特徴のあるページが, 統合された際に下位にランキングされ見過ごされる場合が多い.

本研究では, 複数のサーチエンジンの検索結果の差異を活かして, 特徴のあるページを発見する, メタサーチエンジンを提案する. つまり, それぞれサーチエンジンが独自の戦略によりランキングした結果の独自性を, 既存のメタサーチエンジンのように消すのではなく, 逆に着目して, 他のサーチエンジンと異なる特徴のあるページを活かす. この特徴のあるページを見つける評価基準として, 一部のサーチエンジンが高い評価を与えるページに高い重要度を与えるユニークネス尺度を定義する. また, WWW における実験を通して, 提案したシステムの有効性を検証する. 特徴のあるページがスパム (spam) の結果生じているのか, 新しいページのため一部の検索エンジンでしか認知されていないのか, もしくは特定のランキング戦略でしか高く評価されないのか, すなわち, 真に重要な情報を提供してくれるか否かを検証する.

2. 各サーチエンジンの異種性

各サーチエンジンが返す検索結果に違いが生じることは, すでに述べたが, 実際にどれだけ異なるかを表 1 に示す. 国内の代表的なサーチエンジン, Google [1] [8], Lycos Japan [9], goo [10], Fresheye [11], infoseek [12], Naver Japan [13], 計 6 つのサーチエンジン間の検索結果上位 20 件の URL の重複数を調べた. 検索キーワード 50 個をそれぞれ問合せた結果の平均値である.

特に, goo-Fresheye 間 (58%), Google-goo 間 (49%), Google-Fresheye 間 (40%) は, 重複率が他より高いが, 6 割以下の重複率である. infoseek と他の 5 つのサーチエンジンは 15%以下の重複率であり, infoseek は他と比較して異なった URL を上位にランキングしていることがわかる.

このように, 各サーチエンジンの異種性は非常に大きく, サーチエンジンは, 他と差別化を図り独自性のある検索結果を提供していることが読み取れる. このように異種性が生じているのは, 各サーチエンジンのデータ収集方式や, 索引付けの方式, 検索方式が異なっていることに起因するものである. このためユーザはしばしば複数のサーチエンジンを利用しなければならない場合がある.

表 1　検索キーワード 50 個の検索結果上位 20 件の重複率

	Google	Lycos Japan	goo	Fresheye	infoseek	Naver Japan
Google		39%	49%	40%	15%	33%
Lycos Japan	39%		36%	31%	15%	27%
goo	49%	36%		58%	15%	34%
Fresheye	40%	31%	58%		15%	30%
infoseek	15%	15%	15%	15%		12%
Naver Japan	33%	27%	34%	30%	12%	

このような各サーチエンジンの異種性に着目して, ユニークネス尺度を定義する. 詳しくは, 次章から説明する.

3. ユニークネス尺度を用いたメタサーチエンジンシステムの概要

本システムは, ユーザインタフェース, クエリディスパッチャ, サーチエンジン, テキスト処理, 検索結果の統合のモジュールで構成されている. 本システムの全体図を図 1 に示す. 次節から順に各モジュールについて詳しく説明する.

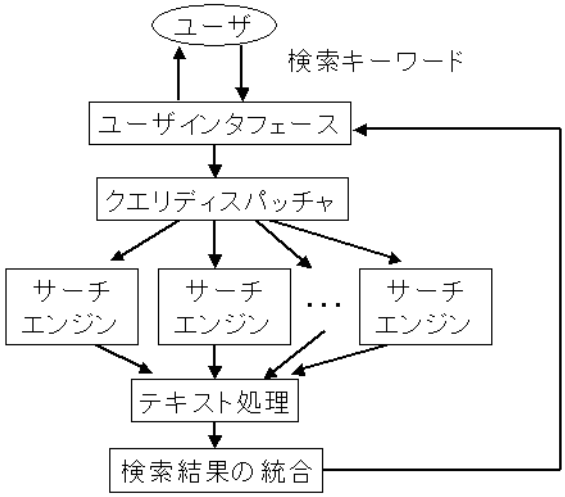

図 1　システムの概要

3.1 ユーザインタフェース

WWW 上で, ユーザが検索キーワードを入力するブラウザ画面を, 図 2 に示す. ユーザが, 検索フォームに検索キーワードを入力し, 検索ボタンを押すと, クエリディスパッチャモジュールが起動する. また, 検索結果を出力したブラウザ画面を図 3 に示す.

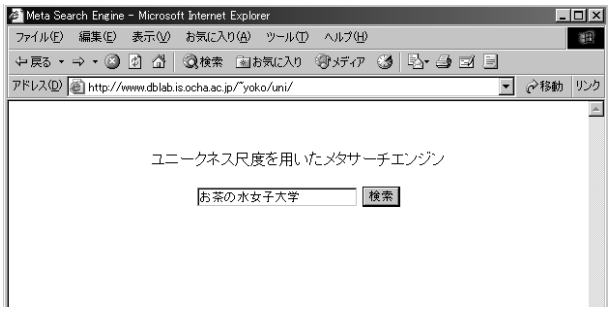

図 2　ユーザインタフェース 入力画面

3.2 クエリディスパッチャ

本モジュールでは, それぞれのサーチエンジンに対して, TCP/IP のコネクションを作成し, 入力された検索キーワードを GET 方式によって発行する.

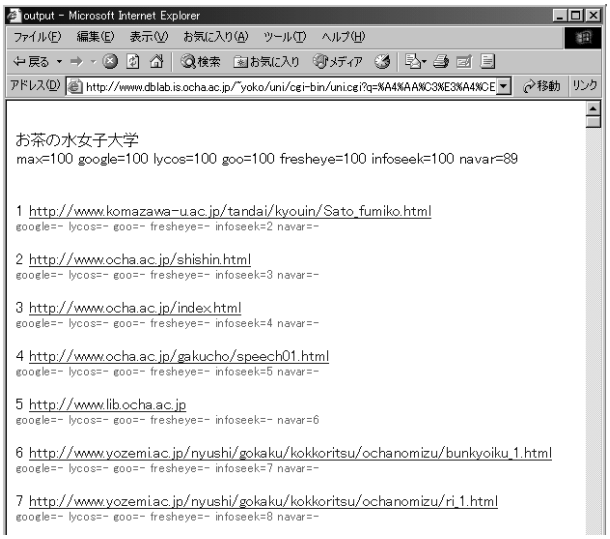

図 3　ユーザインタフェース 出力画面

3.3　サーチエンジン

本システムでは, Google, Lycos Japan, goo, Fresheye, infoseek, Naver Japan, 計 6 つの国内の主要なロボット型サーチエンジンを統合している. なお, 実験では Google は, Google のエンジンを用いている Yahoo! JAPAN [14] のページ検索機能を用いた. 各サーチエンジンの検索条件を等しくするため, 検索に対するオプションを表 2 のように設定した. なお, 空欄は, 検索機能のオプションがない場合を示す.

表 2　各サーチエンジンのオプション設定

サーチエンジン	言語	対象	ソート
Google	日本語	ウェブ	
Lycos Japan		ページ	
goo	日本語	ページ	
Fresheye		ページ	一致順
infoseek		ページ	
Naver Japan		ウェブ	

3.4　テキスト 処理

利用する 6 つのサーチエンジンの検索結果から, 上位 100 件内の URL 情報を抽出するためのモジュールである. 抽出された URL 情報は, 検索結果の統合モジュールに渡される. この URL 情報として, URL とランク情報とどのサーチエンジンでランキングされているかを保持する.

3.5　検索結果の統合

各サーチエンジンの検索結果上位 100 件を統合して, 検索結果を返すモジュールである. テキスト処理モジュールで, 抽出した URL の重複を除去し, URL を表 3 のようにリストアップする. 次に各 URL 毎にランキングを昇順に並べる. 結果は表 4 のようになる. そのランキングを URL 毎に $rank_{url}$ で表すこととし, 特に k 番目を $rank_{url,k}$ と表すことにする.

ここで, $rank_{url,k}$ について詳しく説明する. 表 3 の URL_No の 1 に着目する. この URL は 6 つのサーチエンジンの検索ランキングが (-, 5, 70, -, 10, -) であった. ここで, "-"は上位 100 件に現れなかったことを意味する. 次に, このランキングを昇順に並び替える. 結果は表 4 にあるように (5, 10, 70, -, -, -) となる. これが $rank_{url}$ である. ここで $rank_{url,1} = 5, rank_{url,2} = 10$ である. また, se_num_{url} は, url がサーチエンジンの上位 100 件に現れた回数とする.

表 4 の情報を用いて, ユニークネス尺度で評価し 検索結果を統合する. 統合方法は, 次章で詳しく説明する.

表 3　URL 毎の各サーチエンジンのランキング結果

URL_No	URL	Google	Lycos	goo	Fresheye	infoseek	Naver
1	http://www.gow··	-	5	70	-	10	-
2	http://www.agg··	12	-	15	78	23	45
3	http://www.std··	-	-	5	-	-	-

表 4　URL 毎のランキングソート結果

URL_No	URL	k=1	k=2	k=3	k=4	k=5	k=6
1	http://www.gow··	5	10	70	-	-	-
2	http://www.agg··	12	15	23	45	78	-
3	http://www.std··	5	-	-	-	-	-

4.　ユニークネス尺度の定義

検索結果を統合する基準として, ページのユニーク度を測る, ユニークネス尺度を定義して, 導入する. 本システムでは, 3 つの方法を提案した. ユニーク度は値が高いほど, 良いと評価する.

4.1　方法 1: ランクと重複数を用いる

方法 1 では, ランクが高い URL ほど, ユニーク度が高いと考え, また, ランキングされているサーチエンジン数が少ないほど, ユニーク度が高いと考える. この 2 つの基準を別々に考慮した. ランキングの平均 $(\sum_{k=1}^{se\_num_{url}} rank_{url,k})/se\_num_{url}$ を計算し, ランキングされているサーチエンジンの数 $(se\_num_{url}-1) \times 10$ を重みとして加える. そして全体の逆数をとり, 各 URL のユニーク度 $U1_{url}$ を定義することにする.

$$U1_{url} = 1/\left(\frac{\sum_{k=1}^{se\_num_{url}} rank_{url,k}}{se\_num_{url}} + ((se\_num_{url} - 1) \times 10)\right).$$

方法 1 では, ランキングされているサーチエンジン数を重みとして加算しているため, 統合結果の上位には, 1 つのサーチエンジンのみにランキングされいる URL が集まる傾向にある. ただし, 上位 100 件のランキングを絶対的に扱っているため, 次のような場合に問題が生じる. 例えば, ある URL が (1, -, -, -, -, -) のようにランキングされているとする. この URL が上位 100 件以内にランクキングされているのは, 1 つのサーチエンジンのみである. しかし"-"が, 100 位にはランキングされていないが, 仮に 101 位にランキングされているとする. そうすると, 実際には, この URL に対してもっと低いユニーク度を与えることが適切となる. 方法 2 は, この場合を考慮して定義を行った.

4.2　方法 2: 傾きを用いる

方法 2 では, ユニーク度を視覚的に捉える. 表 4 の $rank_{url,k}$ をグラフ化したのが, 図 4 である.

グラフの傾きが急なほど, ユニーク度が高いと考える. 図 4 の場合を例に挙げると, 直感的に URL_No=3 つまり (5, -, -, -, -, -) の線が, ユニーク度が一番高いと読み取れる. 次に, URL_No=2 (5,

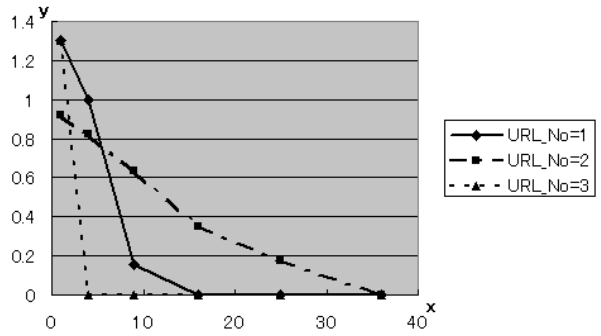

図 4　URL 毎のランキング状況

10, 70, - ,- ,-) がユニーク度が高い. なだらかな傾きの URL_No=1 は, 一番ユニーク度が低い.

各サーチエンジンの検索結果の上位 100 件に現れる回数が少ないページほど, また, 各サーチエンジンで上位にランキングされているページほど, よりユニーク度が高いと評価するため, 二乗, 常用対数を用いて, 傾きの緩急を付加した. 各点 (x, y) は, 次の計算式により求められる.

$$x = k^2,$$

$$y = \log_{10}(100/rank_{url,k}).$$

上記の式について詳しく説明する. x 値を, URL がランキングされているサーチエンジンの数の二乗としたのは, 1 つのみもしくは 2 つのサーチエンジンでランキングされた URL のユニーク度を高くしたいが, ほとんど全てのサーチエンジンにランキングされた URL のユニーク度を非常に低くしたいためである. 次に y 値を $\log_{10}100 - \log_{10} rank_{url,k}$ としたのは, 1 つでもあるサーチエンジンで上位にランキングされれば, その上位のランキングを強調して, ユニーク度を高くしたいためである. また, 方法 1 で問題となった"-"が 101 位と仮定した場合のユニーク度の格差についてだが, "-"は 100 位だと考えているため, 解消される. また仮に"-"が 200 位だとしても, 下位にいくほど, ランクの重要さを減らしているため, ユニーク度の与える影響は少ない.

各 URL のユニーク度 $U2_{url}$ は, 各区分の傾きの総和とし, 傾きは負であるので, 式全体に-1 を掛けて定義することにする.

$$\begin{aligned} U2_{url} &= -\sum_{k=2}^{se\_num_{url}} \frac{\log_{10}\frac{100}{rank_{url,k}} - \log_{10}\frac{100}{rank_{url,k-1}}}{k^2-(k-1)^2} \\ &\quad + \frac{\log_{10}\frac{100}{rank_{url,se\_num_{url}}}}{(se\_num_{url}+1)^2 - se\_num_{url}^2} \\ &= -\sum_{k=2}^{se\_num_{url}} \frac{\log_{10}\frac{rank_{url,k-1}}{rank_{url,k}}}{2k-1} + \frac{2-\log_{10} rank_{url,se\_num_{url}}}{2se\_num_{url}+1}. \end{aligned}$$

方法 2 の統合した検索結果は, 方法 1 と比較すると, 2 つ以上のサーチエンジンにランキングされている URL が上位にランキングされる傾向がある.

4.3　方法 3: ランク差を用いる

方法 3 は URL のランク差に着目する. また, パラメタ値を導入して, 方法 1 や方法 2 のように式を固定するのではなく, パラメタ値を動かし最適な結果を得られようにした.

それぞれのサーチエンジンが, ある URL に対して, それぞれの評価つまりランクを与えているが, このランク差が大きければ大きいほど, ユニーク度が高いと考える. また, 同じランク差でも, 高くランクされている URL の方が, 価値が高いと考えたい.

そのような要請を満たすために, 適当なパラメタを導入しながらランク差 $rank_{k+1}^{\beta} - rank_{k}^{\beta} + \gamma$ に, ランク差の価値 $1/(k \times rank^{\alpha})$ を重みとした量の総和をユニークネス尺度として構成する. 具体的には, 各 URL のユニーク度 $U3_{url}$ は, 次のように表される.

$$U3_{url} = \sum_{k=1}^{k=6} \frac{1}{k \times rank_{url,k}^{\alpha}} (rank_{url,k+1}^{\beta} - rank_{url,k}^{\beta} + \gamma).$$

パラメタ α は, 値を大きくすればするほど, 高いランクを持つ URL のユニーク度を高くし, パラメタ β は, 値を大きくすればするほど, ランク差の大きい URL のユニーク度を高くする働きを持つ. パラメタ γ は, URL のランキングが 6 つのサーチエンジンで全て同じ場合, つまり, (24, 24, 24, 24, 24, 24) のような場合に, ユニーク度が 0 となってしまわないように, 導入した.

パラメタの最適化は, 5.2 節で説明する実験方法に基づいて行った. その結果 $\alpha = 1.2, \beta = 1.0, \gamma = -20$ が最適となった.

5.　実験と評価

提案システムを評価するため WWW において実験を行った. 実験は 3 つの視点から行った. 最初に, 客観的評価をするため, ニッチ (niche) な情報を含んでいるページの割合を調べた. 次にニッチな情報を含んでおり, かつ有益なページを被験者に選んでもらい, その割合を調べ, 主観的な評価を行った. また, 本システムが具体的にどのような内容のページを上位にランキングするかを, 内容を分類して評価した.

5.1　実験データ

実験では, 50 の検索キーワードを用いた. 分野による偏りをなくすために, 10 の分野からそれぞれ, 5 キーワードずつ選んだ. 用いた実験データを表 5 に示す.

表 5　実験で用いた検索キーワード

分野	検索キーワード
芸能人	浜崎あゆみ, B 'z, おニャン子, 広末涼子, SMAP
政治と行政	国土交通省, 自民党, リストラ, 行政改革, ムネオハウス
スポーツ	ゴルフ, 相撲, マラソン, 野球, 柔道
大学	お茶の水女子大学, 東京大学, 慶應義塾大学, 一橋大学, 関西大学
国	アフガニスタン, 中国, 日本, フランス, イギリス
企業	三菱商事, トヨタ自動車, 島津製作所, NEC, ソニー
一般的話題	タマちゃん, プレイステーション, だんご 3 兄弟, 占い, アイボ
病気	エイズ, 白血病, ぜんそく, 癌, 糖尿病
趣味	キャンプ, 茶道, フィッシング, フィギュア, 競馬
コンピュータ	perl, コンピュータウイルス, フリーウェア, HTML, ADSL

5.2　ニッチな情報に関する実験と評価

特徴のあるページを検出できたどうかを調べるために, 検索

キーワードを主題としたページではなく, ページの話題の 1 つとして, 検索キーワードを取り上げているようなページ, つまりニッチな情報を含んでいるページかどうかのみに着目した.

ニッチな情報に関して提案した 3 つの手法を評価する. また, 比較対象として, サージエンジンの代表として, Google とメタサーチエンジンを用いる. メタサーチエンジンの統合方法は [5] の Agreement を用いた. これは, 各ランクの逆数を加算していく方法である. 使用するサーチエンジンは, 本システムと同じである.

ページがニッチな情報を含んでいるかどうかの評価をするために, 登録型エンジンの代表の Yahoo! Japan 登録サイトを用いる. URL が Yahoo! Japan に登録されているページは, 公式ページや有名なページが多いため, ユニーク度が低いと考え, 逆に登録されていなければ, ユニーク度が高いと考える. 対象としたサーチエンジンの検索結果上位 20 件で, 以下の計算方式で Precision(適合率）を計算した. なお上位 20 件としたのは, サーチエンジンが, 1 ページに検索結果を表示する URL 数の標準であり, ユーザが必要とする情報を探す手掛かりとなる範囲と考えられるためである.

$$Precision = \frac{\text{検索された適合 } URL \text{ 数}}{\text{検索された } URL \text{ 数}} \times 100.$$

ここで, 適合 URL とは, Yahoo! Japan に登録されていない URL を示す. つまり以下の式のようになる.

$$Precision\_1 = \frac{\text{検索された } Yahoo!Japan \text{ に登録されていない } URL \text{ 数}}{\text{検索された } URL \text{ 数}} \times 100.$$

3.2.3 節で説明した方法 3 のパラメタ値の最適化を, この Precision_1 をもとに行った. 結果, $\alpha = 1.2, \beta = 1.0, \gamma = -20$ が最適値となった. 実験では, このパラメタ値を用いる.

実験結果を表 6 に示す. 検索キーワード 50 件の Precision_1 の平均と標準偏差である.

表 6　Precision_1 の平均と標準偏差

手法	平均	標準偏差
方法 1	90.0%	15.3%
方法 2	84.0%	16.7%
方法 3	89.0%	15.6%
Agreement	60.8%	20.3%
Google	59.8%	24.7%

ユニークネス尺度を用いた方法 1〜3 のサーチエンジンは, 既存メタサーチエンジン Agreement やサーチエンジン Google よりも Precision_1 が高く, ユニーク度が高いと評価できる. なかでも, 方法 1, 3 が Precision_1 が高いことがわかる. 理由として, 方法 1 や 3 では, 1 つのサーチエンジンにのみランキングされた URL が上位に集まっている, ということが考えられる. つまり, 1 つのサーチエンジンのみにランキングされた URL は, ニッチかどうかという評価基準において, 高く評価されている.

5.3　ニッチな情報とユーザの評価に関する実験と評価

4.2 節の評価方法では, ページがニッチ情報を含んでいるかのみに着目して, ページの有益性を考慮していない. ニッチな情報を含んでいても, そのページが無益な内容であったら意味がない. そこでニッチな情報と有益性 (usefulness), 両方を考慮した評価を行う. ここで言う有益性とは, ページがユーザに対してどれだけ必要な情報を提供できるかである.

各サーチエンジンから抽出した URL の中から, Yahoo! Japan に登録されていない URL を抜き出し, 被験者にその中から, 有益と思われるページを, 15 件を選んでもらった. 方法 1〜3 の各検索結果, 上位 20 件で Precision, Recall(再現率）を計算した. Recall は次の計算式で表せる.

$$Recall = \frac{\text{検索された適合 } URL \text{ 数}}{\text{適合 } URL \text{ 数}} \times 100.$$

ここで, Precision, Recall の適合 URL とは, Yahoo! Japan に未登録で有益であると判断した URL を示す. つまり次式のようになる.

$$Precision\_2 = \frac{\text{検索された } Yahoo!Japan \text{ に未登録かつ有益である } URL \text{ 数}}{\text{検索された } URL \text{ 数}} \times 100.$$

$$Recall\_2 = \frac{\text{検索された } Yahoo!Japan \text{ に未登録かつ有益である } URL \text{ 数}}{Yahoo!Japan \text{ に未登録かつ有益である } URL \text{ 数}} \times 100.$$

実験結果を表 7 に示す. 検索キーワード 50 件の Precision_2 と Recall_2 の平均と標準偏差である.

表 7　Precision_2, Recall_2 の平均と標準偏差

	Precision_2		Recall_2	
手法	平均	標準偏差	平均	標準偏差
方法 1	27.6%	10.8%	36.8%	14.4%
方法 2	37.5%	11.3%	50.0%	15.1%
方法 3	30.6%	10.2%	40.8%	13.6%

Precision_1 の評価が低かった, 方法 2 がこの実験では, 一番良い結果である. 方法 2 は, 方法 1, 3 と比較して, 2 つ以上のサーチエンジンにランキングされている URL も上位に多く集まっていることが理由として挙げられる. つまり, 1 つのサーチエンジンのみにランキングされた URL は, 複数のサーチエンジンにランキングされている URL と比較し, 有益性という点では, 劣っていると考えられる. サーチエンジンとして実際に利用する上で, 有益性を考慮することは重要であり, 方法 2 が, ニッチと有益性という両方において優れているといえる.

5.4　内容分類に関する実験と評価

5.2 節, 5.3 節で, Precision, Recall を用いて数値的にシステムの評価を行った. しかし, この実験だけでは, 本システムが具体的にどのような内容のページを上位に集めるかが検証されていない. そこで次に, 各サーチエンジンによって上位に集められる具体的な URL の内容を調べる. ユニークネス尺度を用いる代表として, 方法 2 を, 比較対照として, メタサーチエンジンと Google を用いる.

実験では, 50 の検索キーワード中,「お茶の水女子大学」「島津製作所」を用いた. 対象としたサーチエンジンの検索結果上位 20 件を, 検索キーワード毎に, URL に着目して分類する. また, 実際にページにアクセスして, 内容が間違っていないないか, スパムではないか, リンク切れしていないかを確かめる.

表 8 は,「お茶の水女子大学」の結果である.「公式」はお茶の水女子大学のサイト内,「関連の公式」は, お茶の水女子大学の付属小学校や中学校のサイト内のページであることを意味する.「その他」は, 上記 2 つに分類されなかったページである. 本システムでは, 公式ではないページを多く検出することができた.

表 8 「お茶の水女子大学」の上位 20 件の URL 分類

	方法 2	メタサーチ	Google
公式	8	12	11
関連の公式	1	3	2
その他	10	4	6
リンク切れ	1	1	1

表 9 は,「島津製作所」の結果である.「ニュース」は新聞局のサイト内のページであり,「公式」は島津製作所のサイト内であることを意味する.「その他」は, 上記 2 つに分類されなかったページである. 本システムは, 公式ではないページを多く検出することができた. 特に, 話題となっている, 島津製作所の社員のノーベル賞受賞ニュースを, 多く検出することができた.

表 9 「島津製作所」の上位 20 件の URL 分類

	方法 2	メタサーチ	Google
ニュース	8	4	4
公式	5	8	8
その他	7	8	7
リンク切れ	0	0	0

また, この 2 つの検索キーワードでは, スパムは検出されなかった. リンク切れしている URL 数は, Google, メタサーチエンジンと比較して, 同等程度であった. ユニーク度が高い, つまり特徴のあるページは, スパムの結果によって生じるのではなく, 鮮度の高いページであるため, 一部の検索エンジンでしか認知されていないか, あるいは, 特定のランキング方式でしか, 高く評価されないといえる.

6. まとめと今後の課題

本研究では, 複数のサーチエンジンの検索結果を統合することで, 偏りの少ない総合的なランキング結果を返すメタサーチエンジンではなく, 差別化を図っているサーチエンジンの独自性ある検索結果の差異を活かして, 特徴のあるページを発見するシステムを提案した.

特徴の度合をユニーク度で計るユニークネス尺度を 3 つの視点から定義した. 方法 1 は, 単純にランクと重複数に着目した. 方法 2 は, ランキング結果をグラフ化し, 傾きを用いて視覚的に定義した. 方法 3 は, ランク差に着目し, パラメタ値を導入して, 式の最適化を行った.

これらの 3 つの方法と既存のメタサーチエンジン, サーチエンジンのそれぞれの検索結果を実験により比較させた. 評価を, ページのニッチさ, ニッチさと有益性, 具体的内容の 3 つに焦点を当てて行った. その結果, 方法 2 に基づくユニークネス尺度が, ニッチさと有益性において優れていることがわかった. また, 本システムは, 特徴のあるページを検出できることが確かめられた. つまり, ユニーク度の高いページは, Yahoo! JAPAN では登録されていない傾向が強く, スパムの結果生じている場合は少なく, 新鮮なページである場合や, ニッチな情報を多く含んでいる場合であることが確認された.

今後の課題としては, まず, システムの機能として, ブーリアン検索に対応できるように拡張することが挙げられる. また, 統合に使用するサーチエンジンを現在は固定しているが, ユーザに自由に選択する機能も付加したい.

ユニークネス尺度の定義については, 本システムでは, それぞれ統合するサーチエンジンのランキングの価値, つまり正当性をどれも等しく扱っているが, 検索キーワード毎に, また, ユーザが探したいページの内容により, 適切なサーチエンジンを自動的に選び, それぞれに適した重みを付加することも考えられる.

また, 方法 3 のパラメタの最適化を 5.2 節の実験で行ったが, 5.3 節の実験でも最適化を行い, 再度, 有効性について検証することが考えられる.

文　献

[1] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” WWW7 / Computer Networks 30(1-7), 1998, pp.107-117

[2] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, “Rank Aggregation Methods for the Web,” WWW10 ACM, 2001

[3] W. Meng, C. Yu, and K. Liu, “Building Efficient and Effective Metasearch Engines,” ACM Computing Surveys, Vol.34, No.1, March 2002, pp.48-89

[4] R. Kosala and H. Blockeel: “Web Mining Research: A Survey,” SIGKDD Explorations, July 2002

[5] B. U. Oztekin, G. Karypis, and V Kumar “Expert Agreement and Content Based Reranking in a Meta Search Environment using Mearf,” WWW2002, May 2002

[6] M. C. McCabe, A. Chowdhury, and D. A. Grossman “A Unified Environment for Fusion of Information Retrieval Approaches,” CIKM, November 1999

[7] M. Montague and J. A. Aslam “Relevance Socre Normalization for Metasearch,” CIKM, November 2001

[8] Google: http://www.google.co.jp/

[9] Lycos: http://www.lycos.co.jp/

[10] goo: http://www.goo.ne.jp/

[11] Fresheye: http://www.fresheye.com/

[12] infoseek: http://www.infoseek.co.jp/

[13] Naver: http://www.naver.co.jp/

[14] Yahoo! Japan: http://www.yahoo.co.jp/