Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

Wikipediaのリンク共起とカテゴリに基づくリランキング手法

倉門浩二[†1] 大石哲也[†1] 長谷川隆三[†1]
藤田博[†1] 越村三幸[†1]

近年，インターネットの普及に伴って，膨大な文書を閲覧することが可能となり，適切な文書を探すために検索エンジンを利用する機会が多くなっている．しかし，検索エンジンを利用しても，求める情報を得ることが難しい場合も多い．本研究は，Wikipediaのリンク情報やカテゴリ構造を解析することで，検索クエリの関連語を抽出し，検索結果の適切なリランキングを行うことを目的としている．

Reranking Methods Based on Wikipedia Link Co-occurrence and Category

Koji Kurakado,[†1] Tetsuya Oishi,[†1]
Ryuzo Hasegawa,[†1] Hiroshi Fujita[†1]
and Miyuki Koshimura [†1]

In recent years, we can access a vast document with the spread of Internet. And we often use the search engine in order to find an appropriate document. However, even if we use the search engine, it is often the case that we cannot find desired information easily. In this paper, we extract related words for the search query by analyzing link information and category structure. And, we aim to assist user in retrieving Web pages by reranking search results.

†1 九州大学
Kyushu University

1. はじめに

近年，インターネットの進歩により，一般家庭からでも容易にWeb(World Wide Web)にアクセスすることができる環境になっている．また，目的のページを見つけるための手段として検索エンジンを利用することが普及した．

大手検索エンジンにGoogleやYahooがある．これらはユーザが自分の興味ある事柄について，単一，或いは複数のキーワードを入力するだけで，膨大なデータベースから最適なページを取捨選択してくれるものである．例えば，Googleはページの評価に各ページのリンクに基づいたPageRankアルゴリズム[12]を用いて，ユーザの必要とするページを検索上位に提示することに成功している．

しかし，Web上に存在するデータ量は莫大で，かつ常に増大しているため，その中からユーザの意図に沿ったページを，短時間で見つけ出せないことがしばしばある．

我々はこの問題を解決するために，Wikipediaに基づいたリランキング手法を提案する．Web百科事典Wikipediaは有用な独自のデータ構造を持つことから，近年，自然言語処理やデータマイニングの分野で注目されている．

我々は，Wikipediaのリンク情報やカテゴリ構造を元に検索語に関連する語群を抽出し，リランキングを行うシステムの実装を行った．

2. 関連研究

「Wikipedia」とはWiki[9]をベースとしたWeb上で最大の百科事典である．Wikipediaはテキスト解析を行うにあたって，いくつか有用な特徴を持つ．そして，その特徴を利用した様々な研究が行われている．

その中でも活発に研究されている分野として，関連度計算がある．Strubeら[14]はWikipediaのカテゴリ構造のみを利用し，2つの語の関連度を計算した．Gabrilovichら[6]は，ESAというWikipediaの記事に現れる全ての単語を利用した手法を提案した．単語をWikipediaに基づく高次元ベクトルに写像し，2つの単語のベクトルの距離を比較することで，関連度を算出している．「WordSimilarity-353 Test Collection」[4][5]によって評価を行われ，ESAはWikipediaの関連度計算手法の中で最も良い結果を残している．Milneら[16]はWikipediaの内部リンクの共起情報を利用した手法を提案した．結果は，ESAよりやや劣るが，計算量ははるかに小さい手法である．Chernovら[1]は，カテゴリに含まれるページのリンク情報を利用して，あるカテゴリと関連するカテゴリ群を抽出した．実験結果は，

Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

inlink 情報が outlink 情報より良い成果を残している．

国内では，Wikipedia を利用したデータマイニングの手法の総称として「Wikipedia マイニング」と名付けられている．中山ら[11] は Wikipedia のリンク情報を pfibf というモデルで解析し，大規模な連想シソーラスを構築する手法を提案している．同じく，伊藤ら[8] はシソーラスの構築手法を提案しているが，こちらではリンクの共起性を用いて関連度を計算している．実験結果は，pfibf と同程度の精度を持ち，計算量ははるかに小さい手法であると言及されている．また，リンクの共起を利用しているという点で，Milne らの手法と類似している．実験結果を比較すると，Milne らの手法がやや良い結果を残しているが，完全に同条件での比較ではないと考えられるので，優劣は決め難い．

中谷ら[10] は Wikipedia のリンクとカテゴリ構造を解析することで，検索結果の評価を行う手法を提案している．検索語の属するカテゴリ領域を検出し，そのカテゴリ領域に含まれる単語を利用して検索結果の評価を行っている．堀ら[7] はクエリ拡張のための情報源として，Wikipedia のページを利用している．

3. Wikipedia に基づくリランキング

本章では，我々が提案する Wikipedia に基づくリランキング手法について述べる．我々は，検索エンジンが返す検索結果の各サイトに対して評価値を求め，評価値の高い順にリランキングを行った．

我々は，リランキングのために Wikipedia から利用できる素性として,「inlink」,「outlink」,「リンク共起」,「カテゴリ」の 4 つがあると考えた．そして，それぞれの素性を利用したモデルによる Web サイト評価手法を提案する．

ここで,「inlink」とは，ある Wikipedia のページへの他のページからの内部リンクのことで,「outlink」とは，ある Wikipedia のページから他のページへの内部リンクのことである．例えば，ページ A の文章中に B というリンクが現れた場合，A は B への outlink を持ち，B は A からの inlink を持つと言える．

以下では，初めに，検索クエリが属するカテゴリを元々クエリが属しているカテゴリだけではなく，その他の関連性の高いカテゴリもクエリの属するカテゴリとして拡張する手法について述べる．

次に，検索クエリが複数の Wikipedia の見出し語を含む場合に，各検索語の重要度を設定する手法について説明する．最後に，各素性を利用した Web サイト評価モデルに付いてそれぞれ説明を行う．

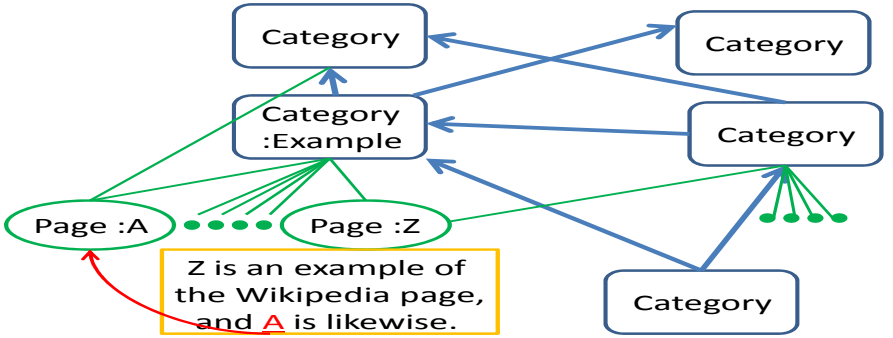

図 1 Wikipedia カテゴリ構造の概観

3.1 クエリの属するカテゴリ集合の抽出

Wikipedia では，1 つのページが 1 つ以上のカテゴリに属している．さらに，wordnet[3] の様なシソーラスと異なり，Wikipedia のカテゴリ構造は単なる階層構造ではなく，図 1 のような 重複を許すツリー構造となっている．

Nakayama ら[11] はクエリの元々所属しているカテゴリだけでは十分な情報を得られないと考えた．そこで，クエリへの inlink を持つページを多く含むようなカテゴリもクエリが所属するカテゴリとみなした．

ここで，$size(c)$ をカテゴリ c に属するページの総数とし，$in(c)$ を c に属するページからクエリ q への inlink の数とするとき，以下の $CScore_{in}(c)$ の値がしきい値以上のカテゴリをクエリの所属しているカテゴリとしている．

$$CScore_{in}(c) = \frac{in(c)}{size(c)} \tag{1}$$

我々も，これに習いクエリの属するカテゴリを拡張する手法を取り入れた．さらに，inlink に加えて，outlink，リンク共起，カテゴリツリーをそれぞれ利用する手法を提案する．

3.1.1 outlink を利用した手法

この手法は非常に単純で，クエリからの outlink を持つページを多く含むカテゴリをクエリの所属するカテゴリとみなす．すなわち，$out(c)$ を c に属するページへのクエリ q からの outlink 数とすると，カテゴリの評価値を以下の $CScore_{out}(c)$ で表す．

$$CScore_{out}(c) = out(c) \tag{2}$$

Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

3.1.2 リンク共起を利用した手法

リンクの共起とは，最も単純に考えるとあるページ中にリンク A とリンク B が同時に現れるということである．しかし，8) で述べられているように一般的にはウインドウを設けて，一定距離以内に現れる単語のみを共起していると考える．また，Wikipedia は独自のデータ構造として Level2 から Level4 までの階層構造になった段落を持つ．例えば，コンピュータのページは「コンピュータの仕組み」という level2 の段落を持ち，その下に「記憶装置（メモリ）」と「入出力」という level3 の段落を持つ．

ここで，我々は以下の 3 つの種類のリンク共起の方法を試した．

(1) 同じページに現れる全てのリンク語
(2) ある語からウインドウサイズ K 以内のセンテンスに現れるリンク語
(3) ある語が含まれる段落で，最もレベルが大きくかつウインドウサイズ K 以上のセンテンスを含む段落に現れるリンク語

(3) について，ウインドウサイズ 10 で上記のコンピュータの例で考える．コンピュータのページが 100 センテンスあり,「コンピュータの仕組み」という段落が 15 センテンス含み,「記憶領域（メモリ）」が 5 センテンス含むとする．さらに,「DRAM」というリンク語が「記憶領域（メモリ）」の中で現れたとき,「DRAM」のリンク共起を考える．

初めに，最大のレベルの段落「記憶領域（メモリ）」のセンテンス数を見る．ウインドウサイズ以下であるので，次に 1 つ上の段落である「コンピュータの仕組み」のセンテンス数を見る．ここで，センテンス数がウインドウサイズ以上であるので,「コンピュータの仕組み」に現れる全てのリンク語を共起していると考える．(もちろん，その中の段落である「記憶領域（メモリ）」中のリンク語も共起しているとカウントする．)

リンク共起でのカテゴリの評価値は c に属するページがクエリ q と共起する回数を $co(c)$ とする時，以下の $CScore_{co}(c)$ で表す．

$$CScore_{co}(c) = \frac{co(c)}{size(c)} \tag{3}$$

3.1.3 カテゴリを利用した手法

Wikipedia のカテゴリツリー上で，一定距離以内にあるカテゴリ同士は関連性が高いと考えられる．そこで，我々はクエリが属するカテゴリの親カテゴリと子カテゴリ，そして共通の親カテゴリを持つカテゴリ（兄弟カテゴリ）にそれぞれ以下の式で与えられる $CScore_{cat}$ を設定した．ただし，$length(c)$ はクエリの属するカテゴリまでの距離である．

$$CScore_{cat}(c) = \frac{1}{2^{length(c)}} \tag{4}$$

3.2 検索語の重要度設定手法

Nakayama らは検索クエリに複数の Wikipedia の見出し語が現れていた場合に，各語の検索語としての重要度を設定する手法を提案した．例えば,「IPhone 日本」という検索クエリがあった場合に「IPhone」は特徴的な語であり，検索語として有用であるが,「日本」は検索語としてあまり有用ではないと考えられる．

Nakayama らは，検索クエリ q 中の各検索語 $q_i \in q$ の重要度 $w(q_i)$ を以下の式で算出した．但し，$inlink(q_i)$ とは q_i への inlink 数とし，$Outllink(q_i)$ とは q_i からの outlink 数とする．

$$w(q_i) = \frac{\log\left(1 + outlink(q_i)\right)}{\log\left(1 + inlink(q_i)\right)} \tag{5}$$

上記の重要度は，ある検索語が多くの inlink を受けるときは一般的な語である可能性が高く，多くの outlink を持つときは詳細な説明を必要とする特徴的な語であるという仮定のもとで建てられた．

一方で，我々はエントロピーを利用した検索語の重要度手法を提案する．エントロピーはある語が等確率で文書に現れたときに最大値をとる．よって，ある語のエントロピーが小さいほど偏った文書に現れていることを意味する．そこで，我々は特定のカテゴリーに偏って出現する検索語ほど特徴的な語であると考え，ある検索語のカテゴリに対する出現頻度のエントロピーを利用した評価値 $w_e(q_i)$ を以下の式で求めた．

$$w_e(q_i) = 1 - H_i = 1 + \frac{1}{\log\left(|C|\right)} \sum_{c \in C} \frac{in_{q_i}(c)}{IN_{q_i}} \log \frac{in_{q_i}(c)}{IN_{q_i}} \tag{6}$$

但し，$|C|$ は Wikipedia のカテゴリ総数で，$in_{q_i}(c)$ は c に属するページからクエリ q_i への inlink の数で，IN_{q_i} はクエリ q_i への inlink の総数である．

3.3 Wikipedia を用いた Web サイト評価

我々は，Web サイトに含まれる Wikipedia の見出し語を利用して，その web サイトの評価を行った．クエリとの関連度を Wikipedia を利用したモデルで算出し，その関連度の高い見出し語を多く含むサイトほど良いサイトと評価する．これは，Wikipedia は信頼性の高いコーパスであり，その中でクエリと関連性が高い語は重要であるという仮定に基づいた評価方法である．

ある web サイト s 中に，Wikipedia の見出し語群 $w(s) = \{t_1, t_2, \ldots, t_n\}$ が現れたとき，s の評価値 $SiteScore(s)$ は，以下の式で求められる．但し，$Score(t)$ は後述の各モデルで

Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

算出する Wikipedia の見出し語 t のクエリ q との関連度である．

$$SiteScore(s) = \sum_{t \in w(s)} Score(t) \tag{7}$$

3.3.1 inlink を利用したモデル

inlink を利用した関連度は，ある Wikipedia のページ p からクエリ q への inlink 数を $inlink(p)$ とし，p に含まれるリンクの数を $linknum(p)$ とすると，以下の $Score_{in}(p)$ で算出する．

$$Score_{in}(p) = \frac{inlink(p)}{linknum(p)} \tag{8}$$

3.3.2 outlink を利用したモデル

outlink を利用した関連度は以下の 2 つの手法を用いて求めた．

(1) tfidf に基づく手法

(2) tfidf ベクトルに基づく手法

(1) は，クエリ q からのあるページ p への outlink 数を $outlink(p)$ とし，Wikipedia の全ページ数を $|W|$，リンク語 p の文書頻度を $|P|$ とすると，以下の $Score_{outtfidf}(p)$ で算出する．

$$Score_{outtfidf}(p) = \frac{outlink(p)}{linknum(q)} \cdot \log \frac{|W|}{|P|} \tag{9}$$

(2) は 11)，16) で用いられている．(1) の tfidf 値を全てのページと算出し W 次元のベクトルを生成する．そして，関連度は各ページ p と q の関連度は tfidf ベクトルの cosine 類似度によって求められる．

よって，ページ p の tfidf ベクトルを $v_p = \{l_{p1}, l_{p2}, \ldots, l_{pn}\}$ とすると，$Score_{outtfidfvec}(p)$ は以下の式で算出する．

$$Score_{outtfidfvec}(p) = \frac{\sum_{k=1}^{n} l_{pk} l_{qk}}{\sqrt{\sum_{k=1}^{n} l_{pk}^2} \sqrt{\sum_{k=1}^{n} l_{qk}^2}} \tag{10}$$

3.3.3 リンク共起を利用したモデル

リンク共起を用いた手法は以下の 3 つの手法を用いて求めた．

(1) cosine 類似度を用いた手法

(2) 2 次共起[13] を用いた手法

(3) Normalized Google Distance[2] モデルに基づく手法

(1) では，Wikipedia 全記事中でのページ p の出現回数を $f(p)$ とし，クエリ q と共起するあるページ（リンク語）p の数を $cooOccur(p)$ とすると，以下の $Score_{cocos}(p)$ で算出できる．

$$Score_{cocos}(c) = \frac{cooOccur(p)}{\sqrt{f(p) \cdot f(q)}} \tag{11}$$

(2) は 8) で用いられている．(1) の cosine 類似度を全てのページと算出し W 次元の共起リンクベクトルを生成する．そして，関連度は各ページ p と q の関連度は共起リンクベクトルの cosine 類似度によって求められる．

よって，ページ p の共起リンクベクトルを $v_p = \{c_{p1}, c_{p2}, \ldots, c_{pn}\}$ とすると，$Score_{cocosvec}(p)$ は以下の式で算出する．

$$Score_{cocosvec}(p) = \frac{\sum_{k=1}^{n} c_{pk} c_{qk}}{\sqrt{\sum_{k=1}^{n} c_{pk}^2} \sqrt{\sum_{k=1}^{n} c_{qk}^2}} \tag{12}$$

(3) は 16) で用いられている．ページ p への inlink を持つページの数を $|P|$ とし，p と q の両方への inlink を持つページの数 (p と q の共起リンクの文書頻度) を $|P \cap Q|$ とするとき，関連度 $Score_{congd}(p)$ は以下の式で求められる．

$$Score_{congd}(p) = \frac{\log(max(|P|, |Q|)) - \log(|P \cap Q|)}{\log(|W|) - \log(min(|P|, |Q|))} \tag{13}$$

3.3.4 カテゴリを利用したモデル

カテゴリを用いた手法は以下の 3 つの手法を用いて求めた．

(1) クエリのカテゴリのみを用いる手法

(2) 3.1 節で拡張したカテゴリを利用する手法

(1) はクエリが属するカテゴリ集合を $C_{set}(q) = \{c_1, c_2, \ldots, c_n\}$ とし，あるページ p がカテゴリ c に属している時に 1 を，属していない時には 0 を表す bool 値を $b(p, c)$ とすると，関連度 $Score_{cat}(p)$ は以下の式で求められる．

$$Score_{cat}(p) = \sum_{c \in C_{set}(q)} \frac{b(p, c)}{size(c)} \tag{14}$$

(2) は，3.1 節で算出した各カテゴリのスコア $CScore(c)$ の高い上位 K カテゴリを $C_{setex}(q)$ とする時，関連度 $Score_{catex}(p)$ は以下の式で求められる．

$$Score_{catex}(p) = \sum_{c \in C_{setex}(q)} \frac{b(p, c) \cdot CScore(c)}{size(c)} \tag{15}$$

4. 実験・評価

本実験は，提案した手法と初期検索結果の比較，そして，複数の提案手法同士の比較のために行う．

Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

4.1 実験概要

我々は，日本語 Wikipedia の 2010 年 3 月 28 日のダンプデータを利用した．また，データ抽出に不適切な「出典を必要とする記事」などの「総記カテゴリ」以下のいくつかのカテゴリを排除した．

実験に利用する検索エンジンとして「Google 日本語検索」を利用した．また，6 名の被験者にそれぞれ 1 語・2 語・3 語の Wikipedia の見出し語を 1 つ以上含む検索クエリを 17 個ずつ，計 51 個のクエリとその目的を用意してもらった．さらに，曖昧さ回避のページに出現するクエリは手動でどの意味かを選んでもらった．そして，それぞれのクエリの検索結果 100 件に対して，以下の基準で 4 段階評価を行った．

4. 高適合（このサイトさえ見れば他のサイトを見る必要はほとんどない）
3. 適合（7 割程度の情報は得られるが，他のサイトも見ておきたい.）
2. 部分適合（目的に対して部分的な情報しか載っていない，あまり役に立たないサイト.）
1. 不適合（目的に関する情報が全く載っていない．無関係なサイト）

次に，検索結果の各 web サイトから，Wikipedia の見出し語を抽出し，Wikipedia の見出し語を 50 以上含まないサイトを排除した．排除したサイトの多くは，Html の構文解析に失敗したものであった．その結果，検索結果件数の平均は 93.86 件となった．また，全クエリでの 3 以上の評価を持つサイトの平均数は 18.02 件であり，評価値 4 のサイトの平均数は 6.340 件であった．

さらに，文書の文字数による影響を受けないようにするため，Wikipedia の見出し語を 500 以上含む文書であっても，500 語までを解析の対象とした．500 に設定した理由は，全文書の相乗平均が約 500 であったためである．

実験の評価方法として，上位 K 件の精度 (Precision at K) と平均平均適合率 (Mean Average Precision, MAP) を用いた．MAP とは，各クエリごとに精度と再現率を考慮した AP(Average Precision) を計算し，全クエリの AP を平均した値である．これらの評価方法は，適合か不適合かのどちらかで評価を行うので，3 以上が適合文書と考えた場合と 4 のみが適合文書であると考えた場合のそれぞれで評価を行った．また，Precision at K の K は 10 とした．

我々は，3 章で複数の手法を提案した．そこで，初めにカテゴリの拡張手法の評価を行い，次に，各モデルを利用した手法の評価と検索語の重要度設定手法の評価を行う．そして，最後に初期検索結果との比較と，提案手法の傾向について評価・分析する．

但し，各手法を組み合わせるときは，各手法で算出したベクトルの 2 乗和を 1 になるように cosine 正規化し，加算合成を行った．

4.2 カテゴリの拡張手法の比較

3.1 節で提案した，カテゴリの拡張手法における評価値を表 1 に示す．$preK$ が評価値 3 以上を適合文書にした時の Precision at K による結果で，H がつくものは評価値 4 のみを適合文書にした時の結果である．リンク共起のウインドウサイズは 10 と設定した．また，リランキング時は評価値の高い上位 20 カテゴリを利用した．

この結果から，単独の手法を比較した場合，outlink による手法が最も良い結果を残したことがわかる．一方で，リンク共起に関してはセンテンスを用いた手法が最も結果がよく，全てのリンクを共起していると判断した手法が最も悪い結果であった．段落を用いた手法がセンテンスを用いた手法より悪くなってしまった理由として，ウインドウサイズ K 以上の段落内にあらわる共起語をカウントしたので，ページによって評価する共起語数にばらつきが出てしまったためだと考えられる．

次に，2 つの手法を組み合わせた場合は，outlink とカテゴリによる手法が最も良い評価値であった．これは，元々結果が良かった outlink にカテゴリによる大局的な情報が加わったためだと考えられる．また，inlink とリンク共起を組み合わせた手法が最も悪い結果であった．これは，inlink とリンク共起は近い関係にある情報であり，ベクトルの合成が良い方向に働かなかったのではないかと考えられる．

表 1 カテゴリの拡張手法による評価値

	preK	preK$_H$	MAP	MAP$_H$
inlink	0.29	0.121	0.315	0.19
outlink	**0.304**	0.128	**0.325**	0.196
リンク共起 (全て)	0.265	0.105	0.288	0.167
リンク共起 (センテンス)	0.281	0.109	0.3	0.172
リンク共起 (段落)	0.273	0.104	0.296	0.167
outlink+カテゴリ	**0.306**	0.122	**0.329**	0.197

4.3 各モデルを利用した手法及び検索語の重要度手法の評価

3.3 節で提案した各モデルを利用した手法の評価値を表 2 に示す．カテゴリの評価値は最も良かった「outlink+カテゴリ」による評価値である．こちらの結果でも，単独の手法では outlink による手法が最も良い結果を残した．また，全体の傾向としてより多くの情報を用いたものや，複雑な手法ほど悪い結果となっている．例えば，リンク共起 (二次共起) や outlink(tfidfVec) の結果は単純な手法より結果が劣っている．特に，二次共起は著しく結果

Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

を落とした．

2 つの手法を組み合わせた場合は outlink とカテゴリまたはリンク共起を組み合わせたものが良い結果であった．これは，カテゴリ拡張の時と同じく，元々結果が良かった outlink にカテゴリによる大局的な情報が加わったためだと考えられる．また，クエリの元々所属するカテゴリのみを利用した場合と拡張した場合との比較は大きな差が出なかった．

表 2 各モデルを利用した手法の評価値

	preK	preK$_H$	MAP	MAP$_H$
カテゴリ	0.306	0.122	0.329	0.197
inlink	0.301	0.127	0.321	0.208
outlink(tfidf)	**0.31**	0.129	**0.334**	0.222
outlink(tfidfVec)	0.301	0.126	0.318	0.2
リンク共起 (cosine)	0.299	0.121	0.318	0.192
リンク共起 (二次共起)	0.267	0.108	0.283	0.159
リンク共起 (NGD)	0.284	0.119	0.297	0.179
outlink+カテゴリ	0.318	0.133	**0.344**	0.224
outlink+カテゴリ (クエリ)	0.320	0.132	0.339	0.220
outlink+リンク共起 (cosine)	**0.321**	0.136	0.342	0.225

次に，3.2 節で提案した検索語の重要度を設定した手法による評価値を表 3 に示す．outlink を用いた手法をベースにして，Wikipedia の見出し語を 2 つ以上含むクエリに関してのみ比較を行った．結論から述べると，いずれの手法もあまり効果が上がらなかった．後に詳しく述べるが，提案手法は 2 語以上のクエリの時に，複数のクエリの関係性を上手く捉えられないために，このような結果になったと考えられる．他の要因として，エントロピーは 1 つのページが複数のカテゴリに属すことや，カテゴリが細分化されているためにカテゴリの偏りを検出出来なかったのではないかと考えられる．

表 3 検索語の重要度設定手法の評価値

	preK	preK$_H$	MAP	MAP$_H$
重みなし	0.297	0.121	0.312	0.176
out/in	0.300	0.122	0.316	0.176
エントロピー	0.296	0.122	0.311	0.178

4.4 初期検索結果との比較

初期検索結果の評価値との比較を表 4 に示す．表のとおりに，提案手法は著しく初期検索結果に比べて評価値を落としてしまった．無作為に選んだ場合に上位 10 件に評価値 3 以上の適合文書が現れる確率は，18.02/93.86 = 0.192 であるので，初期検索結果が非常に良かったとも言える．

次に，1 語のクエリでの評価値の比較を行う．1 語のクエリの検索結果件数の平均は 91.28 件で，3 以上の評価を持つサイトの平均数は 15.88 件であり，評価値 4 のサイトの平均数は 6.029 件であった．

全クエリの場合に比べ，提案手法の評価値が著しく向上していることがわかる．これは，クエリが 2 語以上の場合に 2 つのクエリの意味が非常に近いものでないと，上手く関連語を抽出出来ないためだと考えられる．例えば,「C++,Java」のような検索クエリだと上手くいくが,「Ipod, バックアップ」の様な検索クエリだとあまり良い結果にならなかった．

ここで，無作為に選んだ場合に上位 10 件に評価値 3 以上の適合文書が現れる確率は，15.88/91.28 = 0.174 であるので，無作為に比べて提案手法は 2 倍以上の精度を残していることがわかる．

しかし，Google による初期検索結果の方が良い結果であった．ここで，図 2 では 1 語のクエリに対してのリランキング結果の MAP 値の比較を行っている．

比較は，既存の検索エンジンと比較して精度が上昇しているかどうかを念頭に置き，まず Google の AP を計算する．次に，各クエリを Google の AP によって分類する．分類の仕方は，Google の AP が 0.1 以下，0.2 以下，0.3 以下，…，1.0 以下の 10 通りである．そして，分類されたそれぞれにおける各評価方法の MAP を求める．つまり，Google の AP が 0.2 以下の各手法の MAP とは，Google ではあまり良い結果が得られなかった時の各手法の精度であり，Google の AP が 1.0 以下の各 MAP では，Google で良い結果が得られている時も含めた全体の各手法の精度である．

提案手法が初期検索の MAP 値を逆転する地点は「0.6 ≧」である．この結果から，初期検索では不十分な場合に提案手法が有効に働くと言える．

また，今回 6 名の被験者に評価をしてもらったが，1 名の被験者の評価値だけ非常に悪い結果であった．その被験者の結果は無作為に選んだものと変わらない程度の精度になってしまっていた．

5. おわりに

今回，Wikipedia の特徴を利用して，様々な手法でのリランキングを行った．実験から，単純な手法が最も有効であり，outlink または inlink のような局所的な情報にリンク共起や

Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

表 4 初期検索結果との比較

	preK	preK_H	MAP	MAP_H
outlink	0.31	0.129	0.334	0.222
outlink+カテゴリ	0.318	0.133	0.344	0.224
outlink+リンク共起 (cosine)	0.321	0.136	0.342	0.225
初期検索結果	0.494	0.239	0.505	0.417

表 5 1 語のクエリでの初期検索結果との比較

	preK	preK_H	MAP	MAP_H
カテゴリ	**0.355**	0.15	0.373	0.262
outlink	0.347	0.157	**0.38**	0.314
outlink+カテゴリ	0.365	0.167	**0.394**	0.311
outlink+リンク共起	**0.382**	0.172	**0.394**	0.312
初期検索結果	0.439	0.231	0.477	0.474

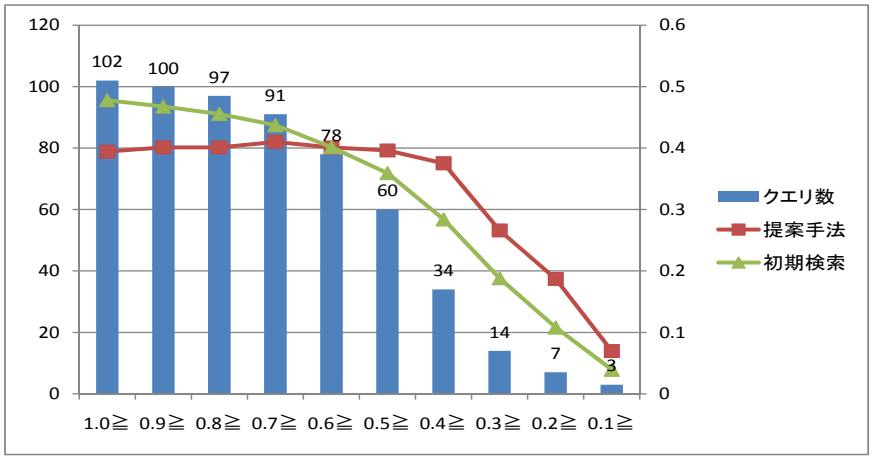

図 2 リランキング結果

カテゴリ情報を組み合わせることで評価値が向上することを確認した．また，1) とは異なり outlink の方が inlink よりやや良い結果を残した．

この結果から，Wikipedia から統計的な情報を抽出する時は，何らかの制約や有効なモデルの元で行う必要があると考えられる．例えば，15) のような機械学習による手法が有効であると考えられる．

今後の研究方針としては，Wikipedia の特徴をより生かせるようなデータ取得方法やそのデータを利用した機械学習やクラスタリングによるデータ分類手法を考えている．

謝辞 本研究は科研費（21500102）の助成を受けたものである．

参 考 文 献

1) S. Chernov, T. Iofciu, W. Nejdl, and X. Zhou. Extracting semantic relationships between wikipedia categories. In *Proc. of Workshop on Semantic Wikis (SemWiki 2006)*. Citeseer, 2006.
2) R.L. Cilibrasi, etal. The google similarity distance. *IEEE Transactions on knowledge and data engineering*, pp. 370–383, 2007.
3) C.Fellbaum, etal. *WordNet: An electronic lexical database*. MIT press Cambridge, MA, 1998.
4) L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E.Ruppin. Placing search in context: The concept revisited. *ACM Transactions on Information Systems (TOIS)*, Vol.20, No.1, pp. 116–131, 2002.
5) L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E.Ruppin. WordSimilarity-353 Test Collection. 2002.
6) E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 6–12, 2007.
7) Kentaro Hori, Tetsuya Oishi, Tsunenori Mine, Ryuzo Hasegawa, Hiroshi Fujita, and Miyuki Koshimura. Related Word Extraction from Wikipedia for Web Retrieval Assistance. pp. 192–199, 2010.
8) M.Ito, K.Nakayama, T.Hara, and S.Nishio. Association thesaurus construction methods based on link co-occurrence analysis for wikipedia. In *Proceeding of the 17th ACM conference on Information and knowledge management*, pp. 817–826. ACM, 2008.
9) B.Leuf and W.Cunningham. The Wiki way: collaboration and sharing on the Internet. *history*, Vol. 1060, p.12.
10) M.Nakatani, A.Jatowt, H.Ohshima, and K.Tanaka. Quality evaluation of search results by typicality and speciality of terms extracted from wikipedia. In *Database Systems for Advanced Applications*, pp. 570–584. Springer Berlin/Heidelberg, 2009.
11) K.Nakayama, T.Hara, and S.Nishio. Wikipedia mining for an association web thesaurus construction. *Web Information Systems Engineering–WISE 2007*, pp. 322–334, 2007.
12) L.Page, S.Brin, R.Motwani, and T.Winograd. The pagerank citation ranking: Bringing order to the web. 1998.
13) H.Schutze and J.O. Pedersen. A cooccurrence-based thesaurus and two appli-

Vol.2010-DBS-150 No.12
Vol.2010-IFAT-99 No.12
2010/8/4

cations to information retrieval. *Information Processing & Management*, Vol. 33, No.3, pp. 307–318, 1997.

14) M.Strube and S.P. Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. In *Proceedings of the National Conference on Artificial Intelligence*, Vol. 21, p. 1419. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

15) A. Sumida, N. Yoshinaga, and K. Torisawa. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in wikipedia. *Proc. of the LREC 2008*, 2008.

16) I.H. Witten and D.Milne. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pp. 25–30, 2008.