Web ページのグループ化による静的動的ランキング

大阪教育大学大学院教育学研究科総合基礎科学数理情報コース

039606 中窪 仁

指導教官 佐藤 隆士 教授

2005 年 1 月 13 日提出

概要

インターネットの普及に伴い，WWW 空間上に蓄積される情報が急増している．この情報群は非常に大規模であり，情報の質も様々であるため，インターネット利用者が必要とする情報を抽出する事は非常に困難となっている．この困難な作業を補助するためのシステムとして，Web 検索システムが存在するが，精度的に十分とはいえない．本論文では，Web 検索システムの精度向上を図るため，Web ページのグループ化とリンク構造を併用した手法を提案する．また，提案手法について実験を行い，その結果を報告する．

目次

1. はじめに

インターネットの普及に伴い，WWW 空間上に蓄積される情報が急増している．2004 年 12 月時点での Web ページ総数は 114 億 Web ページと推定[1]され，さらに増加傾向にある．この情報群は非常に大規模であり，情報の質も様々であるため，インターネット利用者が必要とする情報を抽出する事は非常に困難となっている．この困難な作業を補助するためのシステムとして，Web 検索システムが存在する．

Web 検索システムには大別して 2 種類の方式が存在する．Web ページの情報を手動で蓄積するディレクトリ方式と，Web ページの情報を自動で蓄積するロボット方式である．

ディレクトリ方式は，Web 検索システム管理者が重要と判断した Web ページを Web 検索システムに登録することで情報を蓄積する．その際，各 Web ページは情報の内容によりジャンル別に分類されて登録される．この方式は，Web 検索システムが有しているジャンルに沿った検索語句を用いて検索した場合にはある程度の精度を確保することが可能である．また，ジャンル名をインデクスとして情報が整理されているため，リンク構造を辿っていくことによる情報抽出も比較的容易である．しかし情報蓄積を手動で行うため，WWW 空間上に存在する情報を網羅することは困難であり，情報量は不足しがちである．なお，この方式の代表的な Web 検索システムに Yahoo![3]がある．

ロボット方式は，Web 探索ロボットプログラムが自動で Web ページのリンク構造を辿り情報を蓄積する．この蓄積された情報群に対して全文検索を適用し，検索語句に適合する検索結果集合を抽出することで検索を行う．この検索結果集合に含まれる各 Web ページは，検索語句に対する適合度が高い順にランキング付けされて出力される．この方式はリンク構造を辿ることにより到達可能な Web ページを全て蓄積するため，WWW 空間上に存在する多くの情報から検索を行うことが可能である．しかし，蓄積された情報量が膨大であるために，検索語句の意図する検索結果を出力することは困難である．なお，この方式の代表的な Web 検索システムに Google[4]がある．

ロボット方式では，蓄積された情報群から検索語句に適合する情報を抽出し，適合度順に出力するため，各 Web ページに対して全文検索を行う．しかし，全文検索のみを用いた検索精度には限界があり，他の手法との併用で検索精度の向上を図る必要がある．そこで現在，Web ページ特有の情報であるリンク構造を利用した手法が注目されている．このリンク構造解析による手法としては，PageRank アルゴリズム[5][6]，HITS アルゴリズム[7]などが例としてあげられる．

これらは各 Web ページ間の隣接関係を基にランキングを決定する手法であり，隣接関係にない Web ページとの関係は再帰的に解決される．しかし，現実の WWW 空間上では関連する Web ページ間で常にリンク構造が存在するとは限らない．例えば，リンク行為を Web サイトトップページにしか許可していない Web ページが存在した場合，隣接関係を基に決定したランキングでは適切な結果を得ることができないと考えられる．

この問題を軽減するため本論文では，類似情報を持つ Web ページ群をグループ化することにより，リンク構造上の隣接関係を拡張する手法を提案する．また，リンク構造隣接関係の拡張による検索精度の低下を防止するため，動的スコアリング手法を提案する．以下，2 章にて関連研究について述べる．3 章にて提案システム概要を述べ，4 章から 7 章にて提案手法の詳細を述べる．8 章から 9 章にて実験結果報告，考察を行い，10 章にてまとめる．

[1] Netcraft 社[1]の発表によると，約 5700 万台の Web サーバが存在する（2004 年 12 月時点）．また総務省情報通信政策研究所[2]の発表によると，JP ドメイン 1 サーバあたりの平均 Web ページ数は約 200 ページである（2004 年 2 月時点）．本論文では，この 2 値の積により Web ページ総数を推定した．

2. 関連研究

本章ではリンク構造解析を利用した代表的な手法について，概要と問題点を述べる．

2.1. PageRank アルゴリズム

2.1.1. 概要

PageRank アルゴリズムは，1998 年にスタンフォード大学の L. Page と S. Brin が提案した Web ページの重要度を表す指標であり，「多くの良質なページからリンクされているページは，やはり良質なページである」との思想による，「検索語句に依存しない」スコアリングアルゴリズムである．基本概念はネットサーフィンを行う人のモデル化（ランダムウォークモデル）に基づく．ネットサーフィンを行う人は各 Web ページにあるリンク構造をランダムで辿ると仮定し，各 Web ページに辿り着く確率からスコアを決定する．PageRank アルゴリズムの概念図を図 1 に示す．

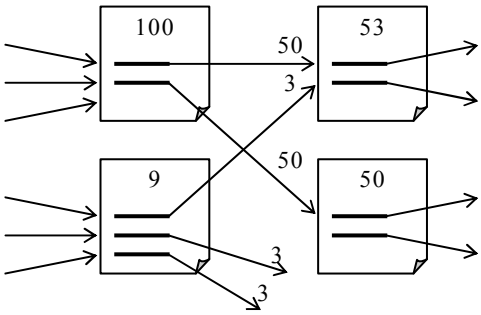


図 1 PageRank アルゴリズム概念図

ただし，このモデルでは全ての Web ページはリンク構造を辿ることにより到達可能であると仮定されているが，現実の WWW 空間においてはそうではない．他 Web ページへリンクがまったくない Web ページ（dangling page）や，逆に他 Web ページへのリンクはあるが，他の Web ページからリンクされていない Web ページも存在する．また，リンク構造が部分集合内でループになっていて外に出ることができない状況（rank sink）も存在する．さらに，ネットサーフィンの始点によってはリンク構造を幾ら辿ろうとも到達できない Web ページ群も存在する．この問題を解決するために，PageRank アルゴリズムでは「ネットサーフィンを行う人は，多くの場合は現在の Web ページに存在するリンク構造を辿って移動するが，時々はまったく無関係な Web ページに移動する」というブラウジングモデルを導入している．このブラウジングモデルにおける「時々」は，具体的には 15%となっている．

PageRank スコア算出式は以下のようになる．$R(p)$ は Web ページ p の PageRank スコアを，$R(q)$ は Web ページ q の PageRank スコアを表す．n は対象とするグラフ G（Web ページをノードとし，Web ページ間のリンク構造をエッジとしたグラフ構造）のノード総数，$outlink(q)$ は Web ページ q から他 Web ページへのリンク構造数である．ε は前述のブラウジングモデルにおける「時々」の確率であり，$\varepsilon = 0.15$ である．

$$R(p) = \frac{\varepsilon}{n} + (1-\varepsilon) \cdot \sum_{(p,q) \in G} \frac{R(q)}{outlink(q)}$$

この数式により算出されるスコアは，WWW 空間上の各 Web ページの遷移確率を表す固定値となり，各 Web ページの被参照度を端的に表す静的スコアとなる．また，数式中にてリンク元 Web ページのスコアを加味していることにより，リンク自体の質も考慮したスコアとなっている．これは，特定 Web ページのスコアを上げるために複数のダミーページからリンクを行うという構造（scam web）が存在しても，その影響を受けにくいというメリットを持つ．

2.1.2. 問題点

PageRank アルゴリズムにおけるリンク行為は，スコア算出式より「Web ページがリンク先 Web ページにスコアを分け与える行為」と考えることが可能である．これは，リンク構造上隣接関係にある Web ページに大きな影響を与えるが，反面，リンク構造上隣接関係にない Web ページは再帰的な関係を解決する必要があるため，受ける影響は小さくなる．つまり，このアルゴリズムはリンク構造上の隣接関係を重視した手法であることが判る．

しかし現実の WWW 空間においては，「スコアを分け与えたい」Web ページおよび情報にリンクを行うことが不可能な場合も存在する．例えば，「リンクはトップページにお願いします」のような記述が Web サイトにあった場合，その Web サイト内の特定の Web ページにスコアを分け与えることはできない．動的生成される Web ページや掲示板のような揮発性の情報についても同様である．これら場合は，PageRank アルゴリズムの想定とのズレが発生し，精度面に影響を及ぼすものと考えられる．この問題の例を図 2 に示す．

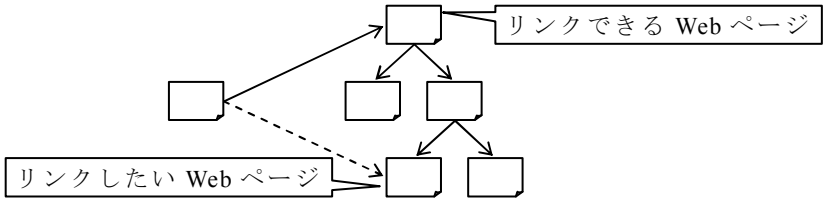


図 2 PageRank アルゴリズム問題点

2.2. HITS アルゴリズム

2.2.1. 概要

HITS アルゴリズムは，1997 年に IBM の J. M. Kleinberg により提案されたコミュニティ抽出のための「検索語句に依存する」スコアリングアルゴリズムである．これは Web ページの重要度を表す指標として Authority と Hub を定義し，この二つの関係を「よい Authority は複数の良質の Hub によってリンクされ，また良質の Hub は複数のよい Authority にリンクをしている」と定義している．Authority とは特定のトピックにおける的確な情報を持つと承認された Web ページ群を意味する．Authority ページは互いの存在を承認したがらない性質を持ち，Hub ページの仲介により関係しあう．HITS アルゴリズムの概念図を図 3 に示す．

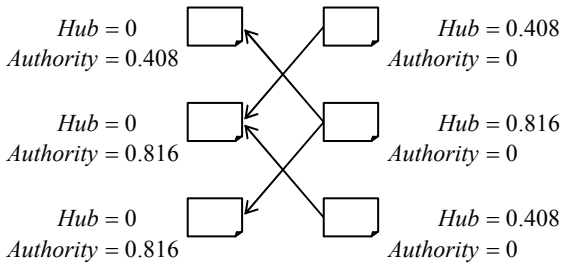


図 3 HITS アルゴリズム概念図

HITS アルゴリズムにおける Authority スコア，および Hub スコアの算出式は以下のようになる． $Authority(p)$ は Web ページ p の Authority スコアを， $Hub(p)$ は Web ページ p の Hub スコアを表す． $p \to q$ は Web ページ p が Web ページ q にリンクしていることを表す．

$$Authority(p) = \sum_{p \to q} Hub(q)$$

$$Hub(p) = \sum_{p \to q} Authority(q)$$

HITS アルゴリズムにおける各スコア算出までの手順を以下に示す.

1. 検索語句を Web 検索システムに与え，その検索語句を含む Web ページを一定数 *r* 件収集し，root 集合とする.
2. root 集合に含まれる Web ページからリンクされている全ての Web ページ，および root 集合に含まれる Web ページにリンクしている Web ページ最大 *n* 件を収集し，root 集合に追加して大きさ *n* の base 集合を作成する.
3. base 集合内の Web ページ間のリンク構造を全て抽出し，グラフ構造を作成する. この際，同ドメイン下にある Web ページ同士のリンク（intrinsic link）を全て削除し，異ドメイン間のリンク（transverse link）のみを残す.
4. グラフ構造より隣接行列を作成，前述の数式を用いて各スコアを算出する. 算出結果のうち，絶対値の大きいものを Authority として抽出する. また，Authority に対応した Hub を抽出する.

この手順および数式により算出されるスコアは，検索語句によって値が変動する動的スコアであり，特定の検索語句に関する有用な Web ページを抽出することが可能である.

2.2.2. 問題点

HITS アルゴリズムには，常に適切なコミュニティを抽出することができるとは限らないという既知の問題が存在する. これは，検索語句に適合しない Web ページが base 集合に存在することが原因であり，このような Web ページが密なリンク構造を持っている場合には適切なコミュニティを抽出できない. この問題の例を図 4 に示す.

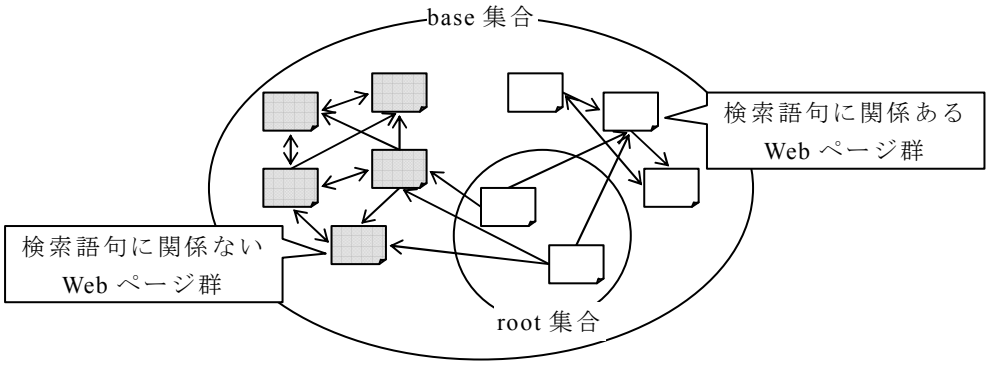


図 4 HITS アルゴリズム問題点

3. 提案システム

本章では，Web ページのグループ化による静的動的ランキング手法を提案する．

3.1. 概要

提案手法は,「類似情報を持つ Web ページ集合間のリンク構造解析スコアと,検索語句に関する情報を持つ Web ページ間のリンク構造解析スコアを併合することにより，検索語句に関連する Web ページ集合を上位に，中でも有用な情報をより上位にランキングする事が可能である」という基本概念に基づいたアルゴリズムである．このアルゴリズムでは,「類似情報を持つ Web ページ集合間のリンク構造解析スコア」を静的スコアと定義し,「検索語句に関する情報を持つ Web ページ間のリンク構造解析スコア」を動的スコアと定義する．

静的スコアは検索語句に依存しないスコアであり，Web ページ集合の持つリンク構造を解析することにより算出する．ここでは PageRank アルゴリズムの問題点を解決するために，Web ページをグループ化することによりリンク構造上隣接関係の拡張を行う．この手順を踏むことにより，得られる静的スコアは Web ページ集合ごとのスコアとなる．なおグループ化処理は，類似情報を持つ Web ページを抽出し，抽出した Web ページ集合内のリンク構造を削除することにより行う．この処理により，Web ページ集合を一つの意味のあるノードとして扱うことが可能になる．

動的スコアは検索語句に依存したスコアであり,検索結果集合に含まれる Web ページ間のリンク構造を解析することで算出する．これにより高いスコアが得られる Web ページは，検索語句に適合した Web ページ中での被参照度が高い Web ページであるといえる．しかし，検索結果集合に含まれる Web ページ間のリンク構造が少数しか存在しない場合も考えられ，その場合はリンク構造解析が期待通りに機能しない可能性がある．また，PageRank アルゴリズムの問題点も解消されない．そこでグループ化を用い，検索結果集合に一部でも含まれる Web ページ集合に対しても動的スコアを算出し，二つの動的スコアを併合して最終的な動的スコアとすることでこれらの問題を解消する．なお，グループ化にて各 Web ページ集合に意味を持たせているため，HITS アルゴリズムの問題点は発生しないと考えられる．

最終的なランキングは，静的スコア，動的スコア，および全文検索スコアを併合することにより得られる最終スコアを用いて行う．これにより得られるランキングは，前述の基本概念に基づいた上で既存手法の問題点を解決したものになると考えられる．

3.2. 処理

提案手法によるランキング算出までの手順を以下に示す．また，システム構成を図 5 に示す．

1. WWW 空間上より Web ページを収集する．収集データよりリンク構造を全て抽出し，グループ化を行う．
2. 手順 1 にてグループ化したリンク構造に対してリンク構造解析スコアリングを行い，グループ化後静的スコアを得る．
3. 検索語句を Web 検索システムに与え，検索結果集合，および全文検索スコアを得る．
4. 手順 3 にて得られた検索結果集合より，検索結果集合に含まれる Web ページ間のリンク構造を抽出する．抽出したリンク構造に対してリンク構造解析スコアリングを行い，グループ化前動的スコア（動的スコア#1）を得る．
5. 手順 3 にて得られた検索結果集合より，検索結果集合に一部でも含まれる Web ページ集合間のリンク構造を抽出する．抽出したリンク構造に対してリンク構造解析スコアリングを行い，グループ化後動的スコア（動的スコア#2）を得る．
6. 手順 2～5 で得られたスコアを併合し，最終スコアを得る．
7. 最終スコア降順にソートを行い，ランキングを得る．

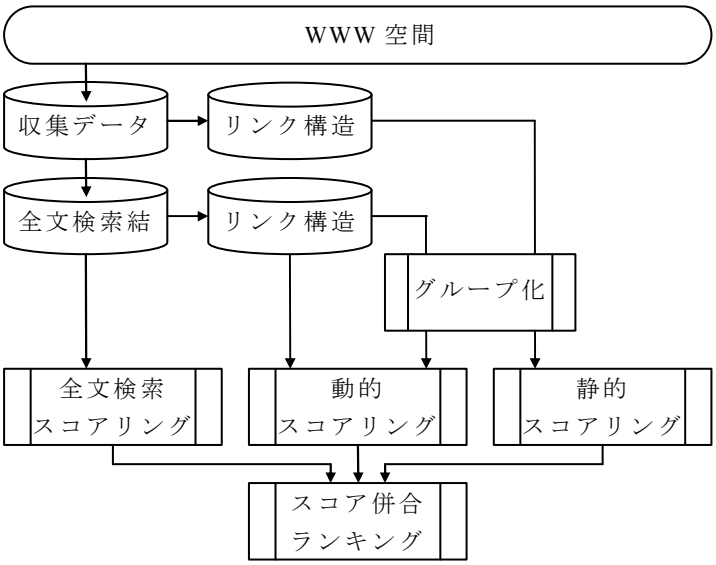


図 5 システム構成

4. グループ化

本章では，グループ化手法について述べる．

4.1. 概要

グループ化の主な目的は，リンク構造上の隣接関係の拡張，および Web ページ集合単位の意味付与の二つである．これを実現するため，各 Web ページを大まかに分類する必要，およびリンク構造の変更方法を決定する必要がある．以下にてそれぞれについて説明する．

4.1.1. Web ページの分類方法

分類を行う前にカテゴリを設定する必要があるが，WWW 空間上に蓄積された全情報を分類可能なカテゴリを設定することは困難である．そのため，詳細なカテゴリ分類は行わないこととし，Web ページの構成より大まかな分類を行う．本提案では，グループを「同一の作成者による類似情報を持つであろう Web ページ集合」と定義した．この定義における「同一の作成者」については，一般に認識されている Web サイトの概念[2]を利用し，「Web サイト内の Web ページ集合は同一の作成者により作成されている」と仮定することにより判別可能である．「類似情報を持つであろう Web ページ集合」に関してもほぼ同様に，「Web サイトの各コンテンツ内の Web ページ集合は類似情報を有する」と仮定することで判別可能となる．

また，分類を行う際に各 Web ページの内容を解析して分類する手法を採用した場合，自然言語解析を行う必要があり，非常にコストがかかる．そこで，リンク構造解析や URL 文字列解析を利用して Web ページの分類を行う手法を採用する．

4.1.2. リンク構造の変更方法

分類を行った Web ページ集合には，Web ページ集合内の Web ページ間のリンク構造，Web ページ集合外からのリンク構造，および Web ページ集合外へのリンク構造が混在する．これらリンク構造のうち Web ページ集合内の Web ページ間で構成されるリンク構造に関しては，Web ページ集合自体が類似情報を持つグループとして認識可能であるためリンク構造を削除して問題ないと考えられる．この操作により，Web ページ集合を単一のノードと認識することが可能である．これに伴い，Web ページ集合外からのリンク構造，および Web ページ集合外へのリンク構造に関してはそれぞれ，リンク先 Web ページをリンク先 Web ページ集合に，リンク元 Web ページをリンク元 Web ページ集合に変更することが必要となる．

この操作を全ての Web ページ集合に適用することにより，Web ページ集合間のリンクのみで構成されるリンク構造の作成が可能である．

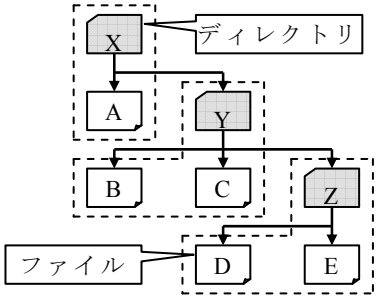
[2]一定の内容，ルールおよびデザインで書かれたウェブページ群を指す．
（出典元：-, “Wikipedia” 〈http://ja.wikipedia.org/wiki/Web%E3%82%B5%E3%82%A4%E3%83%88〉）

4.2. 処理

Web ページをグループ化する手法には，2 種類の方法が考えられる．ディレクトリ構造方式とリンク構造方式である．以下でそれぞれについて説明する．

4.2.1. ディレクトリ構造方式

ディレクトリ構造方式は，類似情報を持つであろう Web ページ集合の抽出にディレクトリ構造を用いる方式である．これは，「Web サイト内の各 Web ページは，Web サイト運営者によって適切なディレクトリ構成の基に管理されている」という仮定に基づくものである．この方式によるグループ化は，ディレクトリ構造を木構造として考え，葉を枝に統合することで実現する．これは，Web ページの URL を親ディレクトリの URL に置き換えるという行為である．例を図 6 に示す．



	Document URL	Group URL
A	X/A.html	X/
B	X/Y/B.html	X/Y
C	X/Y/C.html	
D	X/Y/Z/D.html	X/Y/Z
E	X/Y/Z/E.html	

図 6 ディレクトリ方式概念

この方式を適用した場合，全ての Web ページがディレクトリ URL で置き換えられるため，Web サイト区切りを意識する必要がない．また，URL 文字列のみからグループ化が可能であるため，非常に低コストである．しかし現状の WWW 空間では，「Web サイト内の各 Web ページは，Web サイト運営者によって適切なディレクトリ構成の基に管理されている」という仮定が必ずしも正しくないため，グループ化の精度面では不安が残る．

4.2.2. リンク構造方式

リンク構造方式は，類似情報を持つであろう Web ページ集合の抽出にリンク構造を用いる方式である．これは，「Web サイト内の各 Web ページは，Web サイト運営者の意図通りのリンク構造を構成している」との仮定に基づくものである．この方式によるグループ化は，リンク構造を有向グラフ構造と捉え，その中から部分グラフ構造を抽出することで実現する．例を図 7 に示す．

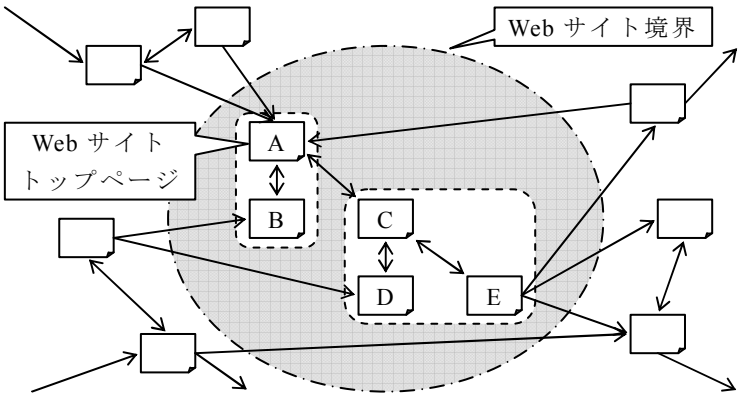


図 7 リンク構造方式概念

この方式は，現状の WWW 空間と比較した場合にも仮定が満たされており，高精度のグループ化が可能である考えられる．しかし，部分グラフ構造の抽出に関してはいまだに研究され続けている分野であり，この方式でグループ化することは現時点では非常に困難である．また，この方式は Web サイト区切りを明確にしておく必要があり，その手法として外部サーバからの被リンク数を基に判断する手法，URL 文字列から判断する手法などが存在するが，これらも精度的に不安が残る．これらの現状を考慮し，本論文ではリンク構造方式についてこれ以上扱わないこととする．

5. 静的スコアリング

本章では，グループ化を用いた静的スコアリングについて述べる．

5.1. 概要

本手法での静的スコアリングは，WWW 空間上の各 Web ページ集合の被参照度を明確にすることが目的である．スコア算出アルゴリズムには，すでに実績ある手法である PageRank アルゴリズムを利用することとし，グループ化を行った後のリンク構造に対して適用する．

5.2. 処理

静的スコアリングの手順を以下に示す．

1. WWW 空間上より収集した Web ページ集合より，リンク構造を抽出する．このリンク構造より Web ページをグループ化し，グループ化済みリンク構造を得る．
2. グループ化済みリンク構造に PageRank アルゴリズムを適用し，各 Web ページ集合のスコアを算出する．
3. 算出されたスコアを各 Web ページにマッピングし，静的スコアを得る．

静的スコアリングの適用例を図 8 に，静的スコアリングにより得られるスコア例を表 1 にそれぞれ示す．この例における Web ページ A～E は同一 Web サイト内の Web ページであり，図 6 で示したディレクトリ構造を持つものとする．Web ページ F～H は Web サイト外の Web ページである．スコア例では Web ページ G，H のスコアが上昇しているが，これはグループ化によるリンク構造上の隣接関係が拡張されたために，与えられるスコアが増えたためと考えられる．しかしその弊害として．Web ページ B，C，および Web ページ D，E のスコアが等価になる，Web ページ A～E のスコア差が減少する，などの現象が発生していることがわかる．

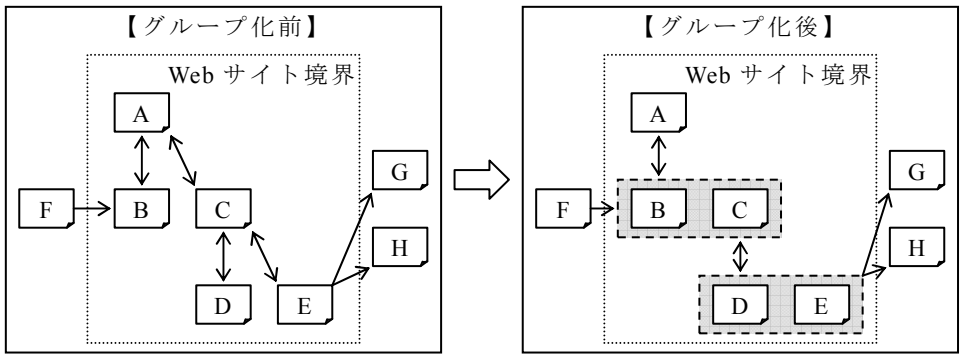


図 8 静的スコアリング適用例

表 1 静的スコア例

	A	B	C	D	E	F	G	H
グループ化前	0.26	0.15	0.26	0.11	0.11	0.00	0.06	0.06
グループ化後	0.22	0.38		0.22		0.00	0.09	0.09

6. 動的スコアリング

本章では，動的スコアリングについて述べる．

6.1. 概要

動的スコアリングは，検索結果集合内での有用な Web ページを抽出することが目的である．スコア算出アルゴリズムに PageRank アルゴリズムを用いることとし，適用範囲を検索結果集合とすることにより動的スコアを算出する．しかし，検索結果集合内でのリンク構造が存在しない場合には遷移確率は全て等しくなってしまう，検索結果集合に含まれる Web ページ間ではリンクされていないが，リンクを数回辿ることにより検索結果集合に含まれる Web ページに到達できるようなリンクを無視してしまう，などの問題が発生すると考えられる．そこで，スコア算出アルゴリズムの適用範囲を「検索結果集合に一部でも含まれる Web ページ集合」とすることにより PageRank アルゴリズム適用範囲を拡張し，この問題を軽減したスコアを算出する．これら二つのスコアを併合することにより，最終的な動的スコアを算出する．

6.2. 処理

動的スコアリングの手順を以下に示す．なお，スコア併合手法については後述する．

1. 検索語句を基に，検索結果集合を抽出する．
2. 検索結果集合よりリンク構造を抽出し，PageRank アルゴリズムを適用して動的スコア#1 を得る．
3. WWW 空間上より収集したリンク構造を利用して，検索結果集合に含まれる各 Web ページを Web ページ集合単位に置き換える．
4. 手順 3 にて置き換えたリンク構造に対して PageRank アルゴリズムを適用し，各 Web ページ集合のスコアを算出する．
5. 各 Web ページ集合のスコアを検索結果集合に含まれる Web ページにマッピングし，動的スコア#2 を得る．
6. 動的スコア#1，#2 を併合し，動的スコアを得る．

動的スコアリングの適用例を図 9 に，動的スコアリングにより得られるスコア例を表 2 にそれぞれ示す．ここで，Web ページ U～X は検索結果集合内の Web ページであり，Web ページ X～Z はグループ化により同一 Web ページ集合に属する Web ページである．動的スコア#1 では，リンクされている Web ページが Web ページ U のみであるため，それにのみスコアが与えられている．このスコアはリンク構造より被参照度を明確に表したスコアである．動的スコア#2 では，グループ化を適用することによりリンク構造上の隣接関係が拡張され，その結果 Web ページ U から二つリンクを辿ることで到達可能であった Web ページ X にもスコアが与えられていることがわかる．

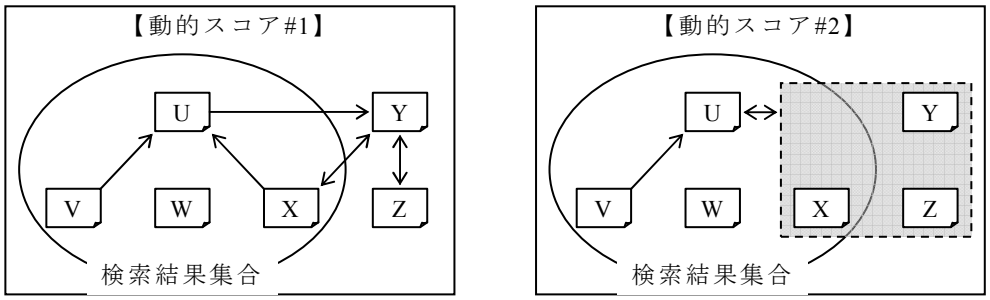


図 9 動的スコアリング適用例

表 2 動的スコア例

	U	V	W	X	Y	Z
動的スコア#1	1.00	0.00	0.00	0.00	0.00	0.00
動的スコア#2	0.50	0.00	0.00	0.50		

7. ランキング

本章ではランキング手法について述べる．

7.1. 概要

ランキングでは，最終的な出力順を決定するための各スコア併合処理を行う．スコア併合の際には各スコアの特性を生かしつつ併合を行う必要がある．各スコアの特性は，表 3 のように考えられる．

表 3 各スコアの特性

	関連度	
	検索語句	リンク構造
全文検索スコア	大	―
静的スコア	―	小
動的スコア#1	大	大
動的スコア#2	小	小

単純な併合方式として，加算方式，および乗算方式が考えられる．しかし，各スコアの特性を生かすことを考えると各スコアに重み付けができない乗算方式は不適切と考えられる．そこで加算方式を採用する．また各スコアにおける最適な重み係数については，実際に実験を行った結果を基に決定する．

7.2. 処理

最終スコア算出式を以下に示す．$Score(p), Retrieval(p), Static(p), Dynamic(p), Dynamic\#1(p),$ および $Dynamic\#2(p)$ はそれぞれ，Web ページ p の最終スコア，全文検索スコア，静的スコア，併合動的スコア，動的スコア#1，動的スコア#2 を表す．また，w_x（ただし $x \in \{r,s,d,d1,d2\}$）は各スコアの対する重み係数を表す．

$$Score(p) = w_r \cdot Retrieval(p) + w_s \cdot Static(p) + w_d \cdot Dynamic(p)$$
$$Dynamic(p) = w_{d1} \cdot Dynamic\#1(p) + w_{d2} \cdot Dynamic\#2(p)$$

8. 実験結果

本章では，提案手法について実験検証した結果を述べる．

8.1. 目的

本実験では，提案システムに含まれる各手法が有効であるかどうかの実験検証を行うとともに，既存手法との比較検証を行う．実験項目を以下に示す．

- ✓ 全文検索スコアによるランキング結果の評価実験
- ✓ 静的スコアによるランキング結果の評価実験
- ✓ 静的スコアと全文検索スコアの併合スコアによるランキング結果の評価実験
- ✓ 動的スコア#1，#2 それぞれによるランキング結果の評価実験
- ✓ 併合動的スコア（動的スコア#1，#2 の併合スコア）によるランキング結果の評価実験
- ✓ 各動的スコアと全文検索スコアの併合スコアによるランキング結果の評価実験
- ✓ 全スコアの併合スコアによるランキング結果の評価実験
- ✓ 重み係数最適値の調査実験

8.2. 環境

実験に先立ち，実験環境について述べる．

8.2.1. PC 環境

実験に使用した PC は，以下のスペックを有するものである．

- ✓ OS : free BSD 5.2
- ✓ CPU : Pentium 4 2.8GHz
- ✓ メモリ : 2GB
- ✓ HDD : 480GB

8.2.2. 対象データ

NTCIR-4 Web Task[8][9]にて提供されているテストコレクションである，NW100G-01 を実験対象データとした．これは約 100GB 分の文書コレクションであり，主に 2001 年 8 月から同 11 月にかけて JP ドメインから収集された文書である．このテストコレクションには，以下のデータが含まれる．

- ✓ 提供文書データの内容
  - ➢ 収集サイトのリスト
  - ➢ 別名サイトのリスト
  - ➢ 重複ページのリスト
  - ➢ メタデータ（収集時の URL，時刻，http ヘッダなど）
  - ➢ ページデータ（原データ）
- ✓ ページデータ処理
  - ➢ 収集したままの原データ
  - ➢ 日本語文字コードを EUC に変換したもの
  - ➢ 日本語文字コードを EUC に変換し，タグを除去したもの
- ✓ 不要ページの除去
  - ➢ 文書コレクションの作成に障害となるため，以下のものを削除済み．
    - ✧ パスのループ
    - ✧ 動的な生成ページ：10 ページを超える分
    - ✧ 明らかにテキストではない長大データ
- ✓ 提供文書データの形式
  - ➢ 1 サイトにつき 1 ファイル

これらに加え，リンク構造を抽出したリストも提供されている．これは，上記ページデータに含まれるリンク構造を抽出したものであり，ページデータ集合に含まれない Web ページへのリンク構造を含む．

ページデータには接頭辞が NW であるドキュメント ID が，ページデータ集合に含まれない Web ページには接頭辞が NX であるドキュメント ID が，それぞれ付与されている．ドキュメント ID の概念を図 10 に示す．

NTCIR-4 Web Task では，接頭辞が NW，もしくは NX のものを検索対象文書として扱っているが，本論文では検索対象を NW 集合のみとした．実験には日本語文字コードを EUC に変換し，タグを除去したページデータ，およびリンク構造データを利用した．ゆえに実験で扱うデータは，Web ページ総数は約 2370 万 Web ページ（うち NW 集合約 1100 万 Web ページ），リンク総数は約 8000 万リンクとなる．

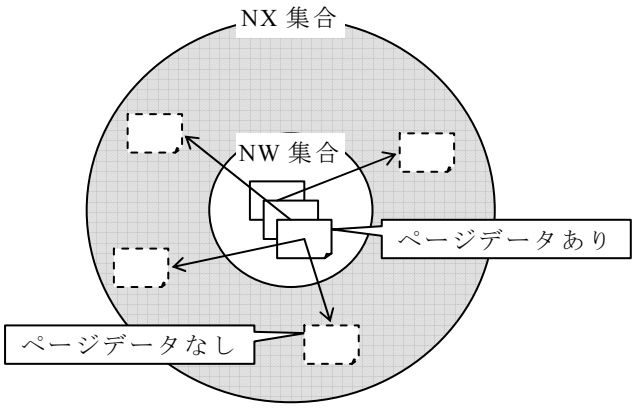


図 10 ドキュメント ID 概念

8.2.3. 検索課題

検索課題として NTCIR-4 Web Task にて用意された 300 課題のうち，実際に利用された 197 課題を用いて実験を行った．全文検索実験の結果，検索結果が十分数得られなかったものを 197 課題からさらに除外し，残った 77 課題を本論文における実験評価対象とした．

なお検索課題には検索語句のほかに，検索者の意図，検索の動機，などの多くの情報が含まれている．しかしこれらの情報を利用した場合は本論文の範疇から超えてしまうため，検索語句のみを利用して実験を行うこととした．

8.3. 評価方式

実験結果の評価は，NTCIR-4 Web Task にて採用された適合判定結果を基に行う．適合判定結果は，多値適合レベルによって高適合，適合，部分適合，不適合のいずれかに判定されている．本実験ではこれらのうち，高適合，適合，部分適合と判定された Web ページを適合文書として扱った．

評価には，WRR[10][11]，DCG[12]，11 点平均適合率および累積適合課題数の 4 種類の評価方式を利用する．以下でそれぞれの評価方式について述べる．なお，以降では高適合 Web ページ集合を H，適合 Web ページ集合を A，部分適合 Web ページ集合を B と表記する．

8.3.1. WRR

WRR（Weighted Reciprocal Rank）は，主に初出の適合 Web ページが検索結果のどの程度上位に現れるかを評価する尺度であり，MRR（Mean Reciprocal Rank）[13]を多値適合レベルに対応するように拡張した評価手法である．MRR は，しばしば質問応答システムの評価に利用される評価方式であり，各質問に対する実行結果リストにおける初出解答のランクの逆数を，全質問にわたって平均した値である．WRR は，以下の算出式で定義される $wrr(m)$ の全検索課題にわたる平均値として求められる．

$$wrr(m) = \max(r(m))$$

$$r(m) = \begin{cases} \delta_h / (i - 1/\beta_h) & \text{if}(d(i) \in H \wedge 1 \le i \le m) \\ \delta_a / (i - 1/\beta_a) & \text{if}(d(i) \in A \wedge 1 \le i \le m) \\ \delta_b / (i - 1/\beta_b) & \text{if}(d(i) \in B \wedge 1 \le i \le m) \\ 0 & \text{otherwise} \end{cases}$$

ここで，$d(i)$ は上位 i ランクの Web ページを，m は実行結果リストにおいて着目するランクの最大値を表す．重み係数は，$\delta_h \in \{1,0\}, \delta_a \in \{1,0\}, \delta_b \in \{1,0\},$ および $\beta_h \ge \beta_a \ge \beta_b > 1$ を満たすものとする．なお，β_x（ただし $x \in \{h,a,b\}$）の値が十分大きい場合，$(-1/\beta_x)$ の項は省略できる．本論文では簡単のため，$(\beta_h, \beta_a, \beta_b) = (\infty, \infty, \infty)$ とした．また，重み係数 δ_x は $(\delta_h, \delta_a, \delta_b) = (1,1,1)$ として評価を行った．図 11 に，本実験における WRR 評価の理想値を示す．

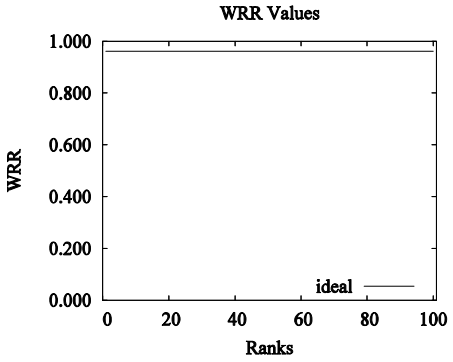


図 11 本実験における WRR 評価の理想値

8.3.2. DCG

DCG（Discounted Cumulative Gain）は，K. Järvelin と J. Kekäläinen によって考案された評価尺度である．これは多値適合レベルに適した評価尺度であり，適合 Web ページのランクを考慮することにより適合度順の評価も可能である．算出式は以下になる．

$$dcg(i) = \begin{cases} g(1) & \text{if}(i = 1) \\ dcg(i-1) + g(i)/\log_b(i) & \text{otherwise} \end{cases}$$

$$g(i) = \begin{cases} h & \text{if}(d(i) \in H) \\ a & \text{if}(d(i) \in A) \\ b & \text{if}(d(i) \in B) \end{cases}$$

ここで，$d(i)$ は上位 i ランクの Web ページを表す．本論文では，対数関数の底は $b = 2$ とし，$(h, a, b) = (3,2,1)$ として評価を行った．図 12 に，本実験における DCG 評価の理想値を示す．

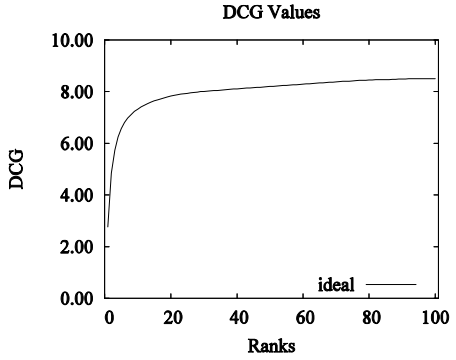


図 12 本実験における DCG 評価の理想値

8.3.3. 11 点平均適合率

11 点平均適合率は，再現率（Recall）と適合率（Precision）より， $Recall = \{0.0,0.1,...,1.0\}$ のそれぞれに対する適合率の平均を算出したものである[3]．再現率は適合 Web ページを洩れなく検索できる度合いを，適合率は検索結果が検索語句に適合しているかどうかを表す．以下に再現率と適合率の算出式を示す．

$$Recall = \frac{|Z|}{|X|}, \quad Precision = \frac{|Z|}{|Y|}$$

ここで， X は適合 Web ページ集合を， Y は検索された Web ページ集合をそれぞれ表し， Z は． $Z = X \cap Y$ を満たす集合とする．図 13 に，本実験における 11 点平均適合率評価の理想値を示す．

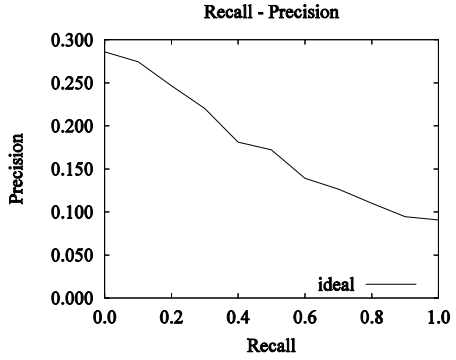


図 13 本実験における 11 点平均適合率評価の理想値

[3] 当該の評価尺度は，TREC[14]にて開発されたツールである trec_eval を用いて算出することが可能である．trec_eval は次の URL より入手可能である．〈http://trec.nist.gov/trec_eval/trec_eval.7.3.tar.gz〉

8.3.4. 累積適合課題数

累積適合課題数（cumulative number of topics which one or more relevant documents were retrieved）は，全検索課題中何課題について検索できているかを表す評価尺度である．この値は，以下の算出式にて定義される *Relevance*(*m*) を全ての検索課題にわたって加算することで求められる．

$$Relevance(m) = \max(c(m))$$

$$c(m) = \begin{cases} 1 & \text{if}\,(d(i) \in \{H, A, B\} \wedge 1 \le i \le m) \\ 0 & \text{otherwise} \end{cases}$$

ここで，$d(i)$ は上位 i ランクの Web ページを，m は実行結果リストにおいて着目するランクの最大値を表す．図 14 に，本実験における累積適合課題数評価の理想値を示す．

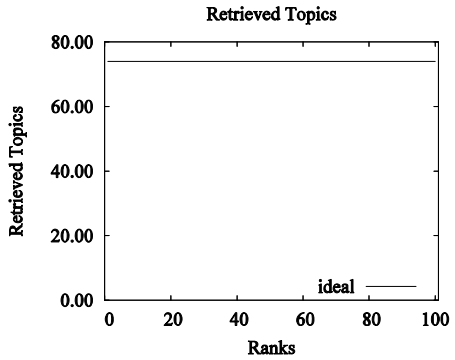


図 14 本実験における累積適合課題数評価の理想値

8.4. 全文検索

提案手法における全文検索結果の抽出には，可変長グラムベースインデクスを利用した全文検索システム[15]を使用した．このシステムの特徴は，文字の出現頻度に基づき各文字にハフマン符号を割り当て，グラムを可変長に扱っていることである．また，データ圧縮と 3 段階の木構造を持つデータ構造により，従来のグラムベースインデクスの問題点であった索引サイズの削減を実現しつつ高速な検索を可能としている．

しかし，実験に利用したシステムは半角文字の利用ができない仕様となっていたため，実験対象データに含まれる半角文字を全角文字に変換することによりこの問題を回避した．また，半角文字変換の際に空行および重複する空白文字の削除を行うことにより，データサイズ削減を図った．インデクス作成に関する情報を表 4 に，検索に関する情報を表 5 にそれぞれ示す．

表 4 元データ→インデクス変換処理

データ	サイズ [GB]	処理時間 [H]
未加工	100.0	－
HTML タグ除去	44.0	－
半角文字変換	37.8	20.0
インデクス作成	30.2	14.7

表 5 検索処理

検索課題数	300 [課題]
検索語句数	497 [語句]
総検索時間	212 [sec]
検索語句あたり検索時間平均値	42 [msec]
検索語句あたり検索時間中央値	20 [msec]
検索課題あたり検索時間	707 [msec]

全文検索スコアは $tf \cdot idf$ 法を利用して重み付けを行い，その後確率モデルを適用することにより算出した[16]．

$tf \cdot idf$ 法の計算式を以下に示す．なお，N は全文書数，tf_{ij} は文書 d_j 中に出現する語 t_i の出現数，df_i は語 t_i を含む文書数を表す．

$$W_{ij} = tf_{ij} \cdot idf_i$$
$$idf_i = \log\left(N / df_i\right)$$

確率モデルによるスコア算出式を以下に示す．なお，w は検索語句の重み係数を示すものとし，定数 k は 2 とした．

$$s = \frac{tf}{k + tf} \cdot idf \cdot w$$

算出した全文検索スコアの情報を表 6 に，全文検索スコアによる評価結果を図 15 に示す．

表 6 全文検索スコア

最小値	2.2831
最大値	30.2596
平均値	10.3889
中央値	9.4818

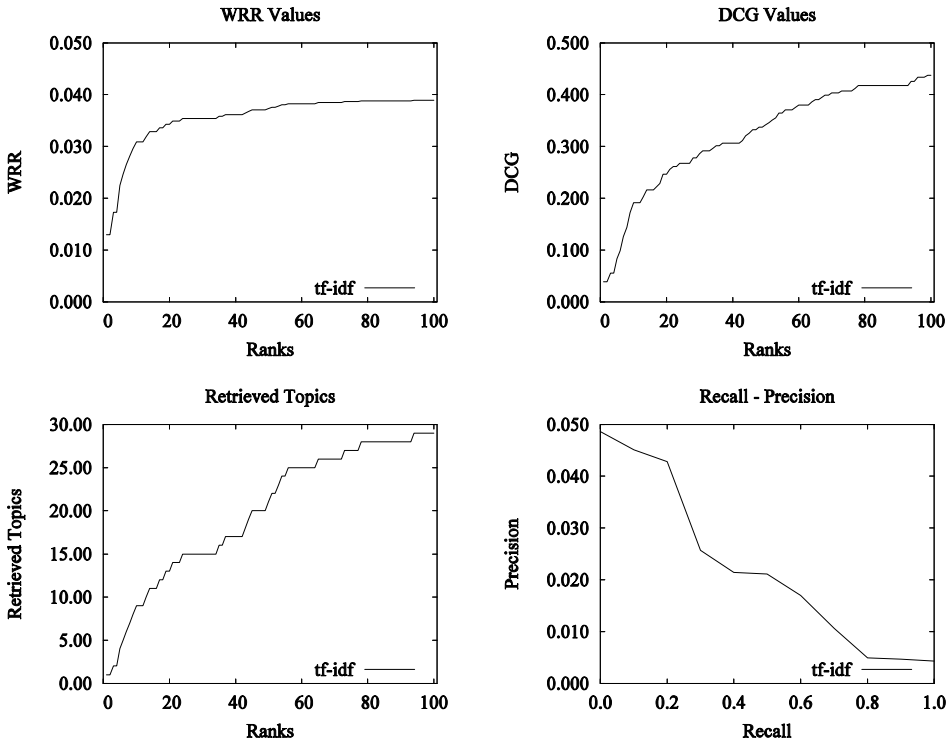


図 15 全文検索スコアによるランキング評価

8.5. 静的スコアリング

ここでは，提案手法のひとつである静的スコアリング手法の効果を判断するために実験を行った．

8.5.1. グループ化および静的スコアに関する実験

全リンク構造を対象にグループ化を行った結果を表 7 に示す．この結果より，各グループに含まれる Web ページ数が非常にかたよっていることがわかる．平均して 5 つの Web ページがディレクトリ内に格納されていることを読み取ることができるが，大半のディレクトリには Web ページがひとつしか格納されておらず，一部のディレクトリに非常に大量の Web ページが格納されていることも同時に読み取ることができる．このような Web ページ数のかたよりはリンク数に影響を及ぼすと考えられる．その結果，リンク構造解析によるスコアリングに影響を及ぼすと考えられ，ディレクトリ方式によるグループ化の性能は低いと考えられる．

表 7 グループあたりの Web ページ数

最小値	1
最大値	30,466
平均値	5
中央値	1

次に，グループ化前後それぞれのリンク構造に PageRank アルゴリズムを適用した結果を表 8 に示す．グループ化により，Web グループ総数は Web ページ総数の 19%まで，リンク総数は 23%まで減少していることがわかる．そしてグループ化前のスコアと比べて，グループ化後のスコアは平均化されていることがわかる．これは，グループ化により Web グループ数が Web ページ数に比べて大幅に減少したため，各 Web グループへの遷移確率が増加したこと，および各 Web グループ内の Web ページ間で PageRank スコアが分散してしまったことが原因であると考えられる．またグループ化後のスコアは，グループ化前と比べ最大値は下回っているが最小値および平均値は上回っていることがわかる．グループ化前後の静的スコア比較結果を図 16 に示す．このグラフは，グループ化前後それぞれのスコアを昇順にプロットした場合の近似曲線である．各スコアは正規化のため 16 乗根をとり，10^{-1} のオーダーまで範囲を縮小してある．また X 軸は Web ページ総数，Web グループ集合総数を 100%とした百分率を，Y 軸はスコアをそれぞれ表している．この比較結果より，全体の約 1/4 ではグループ化前静的スコアが高スコアを，残り 3/4 ではグループ化後静的スコアが高スコアを示していることがわかる．

表 8 グループ化前後による静的スコア比較

	グループ化前	グループ化後
ノード総数	23,670,000	4,500,000
リンク総数	79,700,000	18,140,000
スコア最大値	2.6126E-04	4.1985E-07
スコア最小値	7.3143E-09	3.3442E-08
スコア平均値	4.2231E-08	2.2230E-07
スコア中央値	8.3860E-09	2.2612E-07

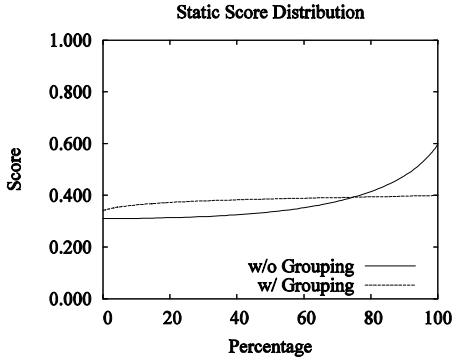


図 16 グループ化前後によるスコア分布比較図

次に, 静的スコアのみでランキングを行った評価を図 17 に示す. なお, StaticN はグループ化前静的スコアを, StaticG はグループ化後静的スコアをそれぞれ表す. StaticN の結果に比べ, StaticG の結果は非常に低い結果となっていることがわかる. これは, スコアが平均化してしまっているために本当に優れている Web ページを抽出できていないためと考えられる.

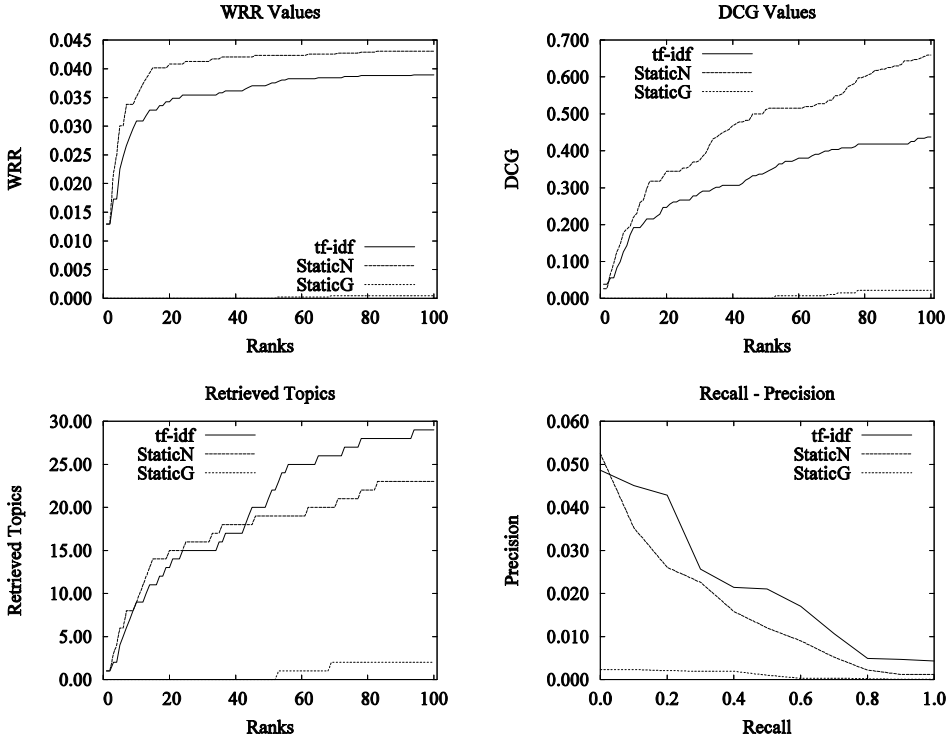


図 17 静的スコアのみによる評価結果

図 17 の結果を詳細に調査するため, グループ化前後それぞれの検索結果の比較を行った. 比較結果を表 9 に示す. この結果より, グループ化後のみで適合する課題が存在することがわかる. またグループ化前後双方で適合する場合について, グループ化前後それぞれのランク比較を行った結果を表 10 に示す. この結果より, グループ化前後双方で適合する場合においてグループ化後ランクが優位になる場合が, 適合課題では 10%, 適合文書では 25%あることがわかる.

以上の調査結果をまとめると, グループ化前後それぞれの特徴は表 11 のようになる. それぞれの項目に関し

て，グループ化前とグループ化後は逆の性質をもっていると考えられる．よって，グループ化前後それぞれの利点を併せ持つスコアを得るためには，これら二つの静的スコアを併合する必要があると考えられる．

表 9 グループ化前後による検索結果比較

適合手法	グループ化前	双方	グループ化後
適合課題数 / 総検索課題数	38 / 77	10 / 77	9 / 77
適合文書数 / 総適合文書数	195 / 560	47 / 560	22 / 560

表 10 双方適合文書におけるグループ化有無によるランク比較

	グループ化前 < グループ化後	グループ化前 > グループ化後
課題数 / 総適合課題数	9 / 10	1 / 10
文書数 / 総適合文書数	35 / 47	12 / 47

表 11 グループ化前後それぞれの特徴

	グループ化前	グループ化後
スコア分散	大	小
スコア優位帯	高スコア帯	低スコア帯
ランクへの影響度	大	小

静的スコア併合は，スコアの 16 乗根をとることにより各スコアを正規化した後，加算方式で算出する．併合式を以下に示す．なお，Web ページ p について，$StaticN(p)$ はグループ化前静的スコアを，$StaticG(p)$ はグループ化後静的スコアをそれぞれ表す．w_x（ただし $x \in \{sn, sg\}$）は加算時の重みを表す．ただし重みについては，二つの静的スコアがそれぞれ別の特徴をもち，有用性は等価であると考えたため，双方とも 1 とする．

$$Static(p) = w_{sn} \cdot StaticN(p) + w_{sg} \cdot StaticG(p)$$

図 18 に，上記併合式で二つの静的スコアを併合した評価結果を示す．Merged が併合した静的スコアを表す．併合によるノイズの影響で多少の相違はあるが，どのグラフにおいてもグループ化を行わなかった場合とほぼ同等の精度であることがわかった．また 11 点平均適合率のグラフより，僅かな精度向上が確認できた．

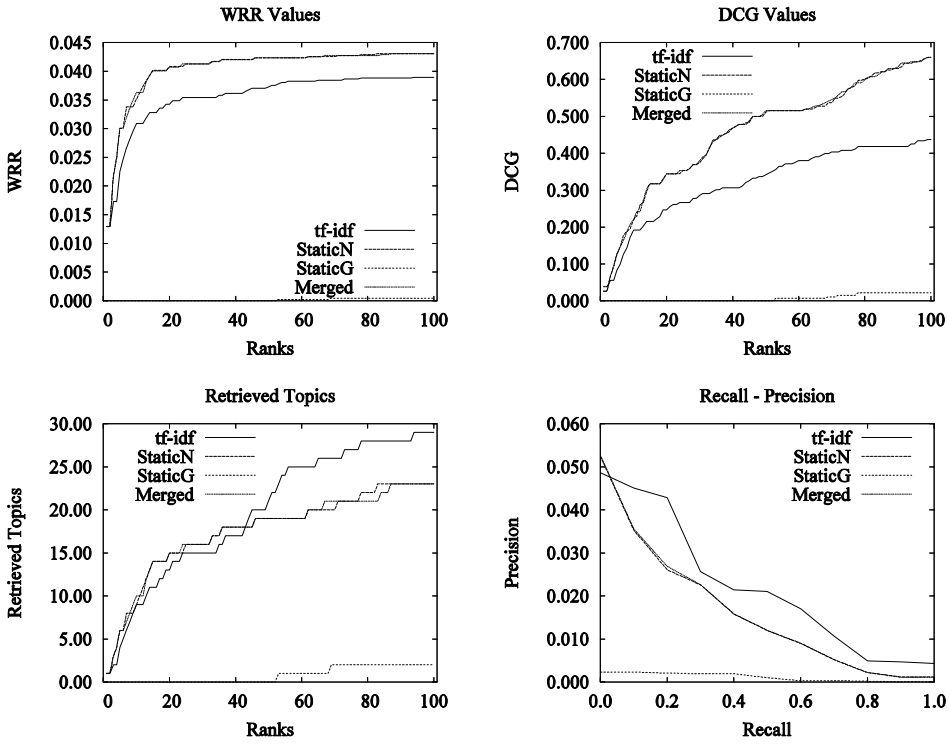


図 18 静的スコアを併合した場合の評価結果

8.5.2. 全文検索スコアとの併合に関する実験

提案手法である静的スコアリングの検証として，全文検索スコアと静的スコアの併合実験を行った．併合スコア算出式を以下に示す．式中 Web ページ p について，全文検索スコアを $Retrieval(p)$，グループ化前後の静的スコアをそれぞれ $StaticN(p)$， $StaticG(p)$ と表す． w_x（ただし $x \in \{r,sn,sg\}$）は重み係数を表す．なお，全文検索スコアは 2 乗根をとることにより 10^1 のオーダーに，PageRank スコアは 16 乗根をとった値を 10 倍することにより 10^0 のオーダーに，それぞれ正規化を行うものとする．

$$Score(p) = w_r \cdot Retrieval(p) + w_{sn} \cdot StaticN(p) + w_{sg} \cdot StaticG(p)$$

静的スコアリングの評価結果を図 19 に示す．この結果より静的スコアを併合した場合に最も精度が高くなることが確認できた．

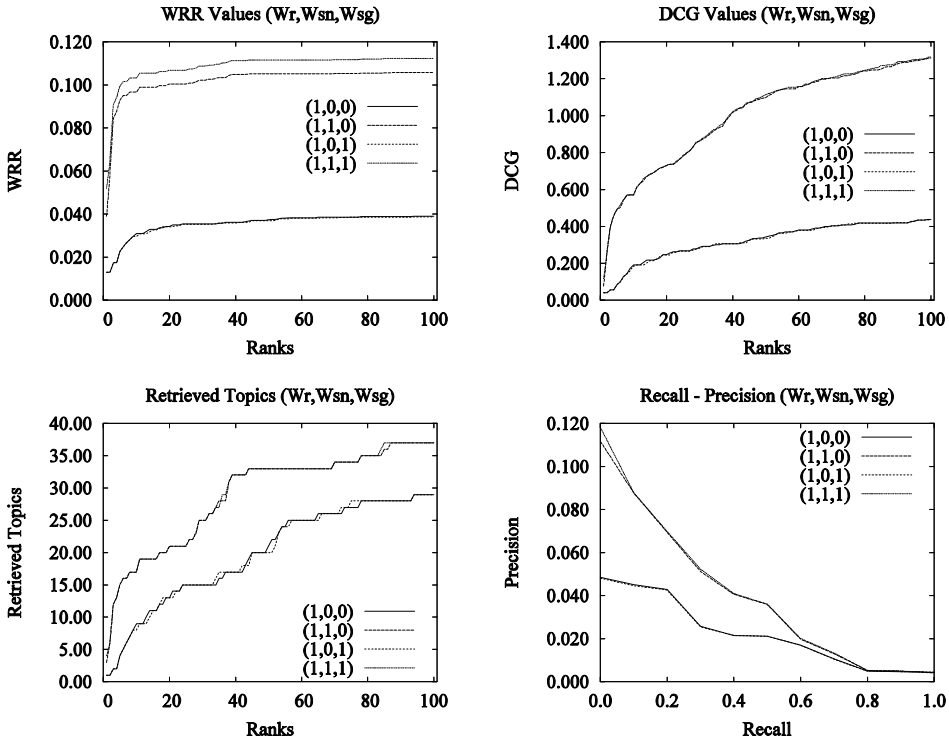


図 19 静的スコアリング評価結果

8.6. 動的スコアリング

ここでは，動的スコアリング手法の効果を判断するための実験を行った．

8.6.1. 動的スコアに関する実験

全文検索結果集合，およびそれに含まれるリンク構造より動的スコアを算出した結果を表 12 に示す．動的スコア#2 にて，リンク数が 25%増加していることがわかる．これはグループ化によりスコアリング対象が増加したためであり，静的スコアリングにおけるグループ化と異なる部分である．また，動的スコアの値についても静的スコアリングにおける傾向と異なり，グループ化後静的スコアと似た傾向となっている．動的スコアの比較結果を図 20 に示す．このグラフは静的スコアリングにおけるスコア比較図（図 16）と同様の手法で各スコアを正規化し，昇順でプロットした場合の近似曲線である．グループ化によるスコアの平均化現象はあるが，非常に似た曲線になっていることが確認できる．

表 12 動的スコア#1，#2 による比較結果

	動的スコア#1	動的スコア#2
ノード総数	192,500	124,041
検索課題あたりノード数	2,500	1,611
リンク総数	95,848	120,292
検索課題あたりリンク数	1,245	1,562
スコア最大値	4.8634E-01	5.6874E-02
スコア最小値	6.8460E-05	7.6747E-05
スコア平均値	4.0000E-04	6.3694E-04
スコア中央値	7.0123E-05	5.1010E-04

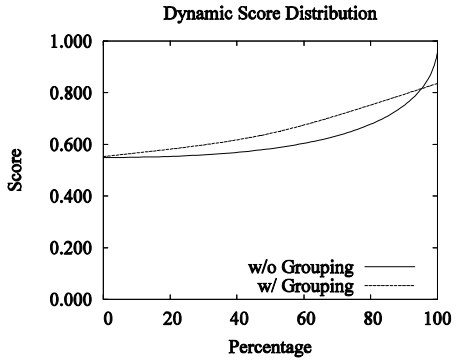


図 20 動的スコア#1（w/o Grouping），#2(w/ Grouping)によるスコア分布比較図

動的スコア#1，#2 をそれぞれ個別に評価した結果を図 21 に示す．11 点平均適合率のグラフより，動的スコア#2 に比べ動的スコア#1 の方が全体的に高性能であるといえる．しかし，他のグラフにおける Rank10 までの評価をみた場合，動的スコア#2 の方が優れていることがわかる．

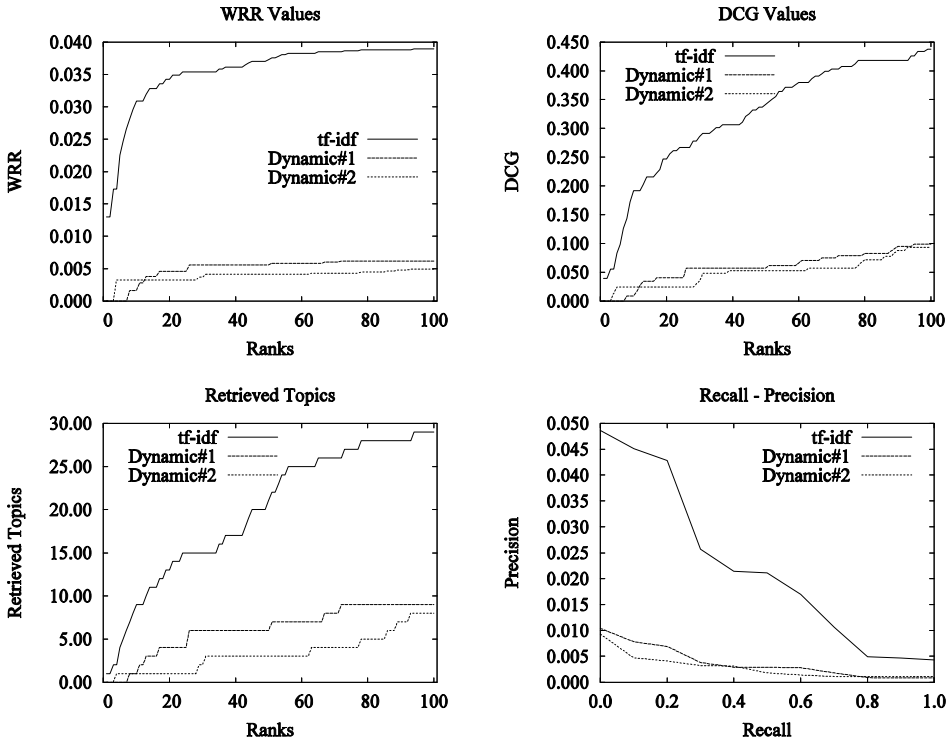


図 21 動的スコアのみによる評価結果

図 21 の結果を詳細に調査するために,動的スコア#1,#2 それぞれの検索結果の比較を行った.比較結果を表 13, および表 14 に示す．適合課題数についてはそれほど差がないが，適合文書数については動的スコア#2 は動的スコア#1 の 30%程度しか抽出できていないという結果になっている．これが 11 点平均適合率グラフでの差として現れていると考えられる．また，双方適合文書について見た場合，動的スコア#2 を適用した場合に精度が上がるケースが非常に多いことがわかる．これは，グループ化によりリンク構造上隣接関係が拡張された結果，与えられるスコアが増加したためと考えられ，これにより動的スコア#1 では低ランクであった適合文書が動的スコア#2 では高ランクに押し上げられ，11 点平均適合率以外のグラフでの高評価となったと考えられる．

表 13 動的スコア#1，#2 による検索結果比較

適合手法	動的スコア#1	双方	動的スコア#2
適合課題数 / 総検索課題数	15 / 77	23 / 77	11 / 77
適合文書数 / 総適合文書数	135 / 560	37 / 560	41 / 560

表 14 双方適合文書における動的スコア#1，#2 によるランク比較

	動的スコア#1 < 動的スコア#2	動的スコア#1 > 動的スコア#2
課題数 / 総適合課題数	10 / 23	13 / 23
文書数 / 総適合文書数	4 / 37	33 / 37

以上の調査結果をまとめると，以下のようになる．

- ✓ PageRank スコア全体における特徴…動的スコア#1，#2 ともによく似た傾向を示す．比較した場合，動的スコア#1 の PageRank スコアの分散度が大きく，動的スコア#2 の分散度は

小さい．

- ✓ 評価における特徴…動的スコア#1 の方が全体的に高性能であり，動的スコア#2 はそれに劣る．しかし，動的スコア#2 の方が一時的に高性能になるケースもある．

次に，提案手法である動的スコア#1，#2 の併合について考える．動的スコア併合は，各スコアの 16 乗根をとることによって正規化を行った後，加算方式で算出する．併合式を以下に示す．なお，Web ページ p について，$Dynamic\#1(p)$ は動的スコア#1 を，$Dynamic\#2(p)$ は動的スコア#2 をそれぞれ表す．また，w_x（ただし $x \in \{d1, d2\}$）は加算時の重み係数を表す．

$$Dynamic(p) = w_{d1} \cdot Dynamic\#1(p) + w_{d2} \cdot Dynamic\#2(p)$$

加算時の重み係数については，両スコアの特徴が似ているために固定値として決めることは難しい．そこで，何パターンかの重み付けにより実験評価を行い，最適な重みを調査した．重み係数 $(w_{d1}, w_{d2}) = (1,1),(2,1),(1,2)$ の 3 パターンでの実験評価結果を図 22 に示す．$(w_{d1}, w_{d2}) = (1,2)$ については，適合文書を上位にランクすることが可能であることが WRR グラフより，常に適合文書を上位にランクすることが不可能であることが累積適合課題数グラフよりわかる．$(w_{d1}, w_{d2}) = (2,1)$ については，適合文書を上位にランクすることが不可能であることが WRR グラフより，多くの課題において適合文書を抽出可能であることが累積適合課題数グラフよりわかる．$(w_{d1}, w_{d2}) = (1,1)$ については，他 2 パターンの中間的な特徴をもつことがわかる，また，3 パターンとも性能的にはあまり相違が無いことが 11 点平均適合率グラフよりわかる．よって重み (w_{d1}, w_{d2}) は，求める性能によって決定する必要があると考えられる．

次いで，併合スコアと動的スコア#1，#2 の比較結果を図 23 に示す．この比較結果より，$(w_{d1}, w_{d2}) = (1,2)$ における WRR が非常に優れていること，累積検索課題数が非常に劣っていることがわかる．また動的スコア#2 単体にくらべ $(w_{d1}, w_{d2}) = (1,2)$ の評価が上昇していることから，動的スコア#2 の特徴を活用するためには動的スコア#1 との併合が必要であると考えられる．一方，$(w_{d1}, w_{d2}) = (2,1)$ については動的スコア#1 単体と比較しても大差ない結果となった．

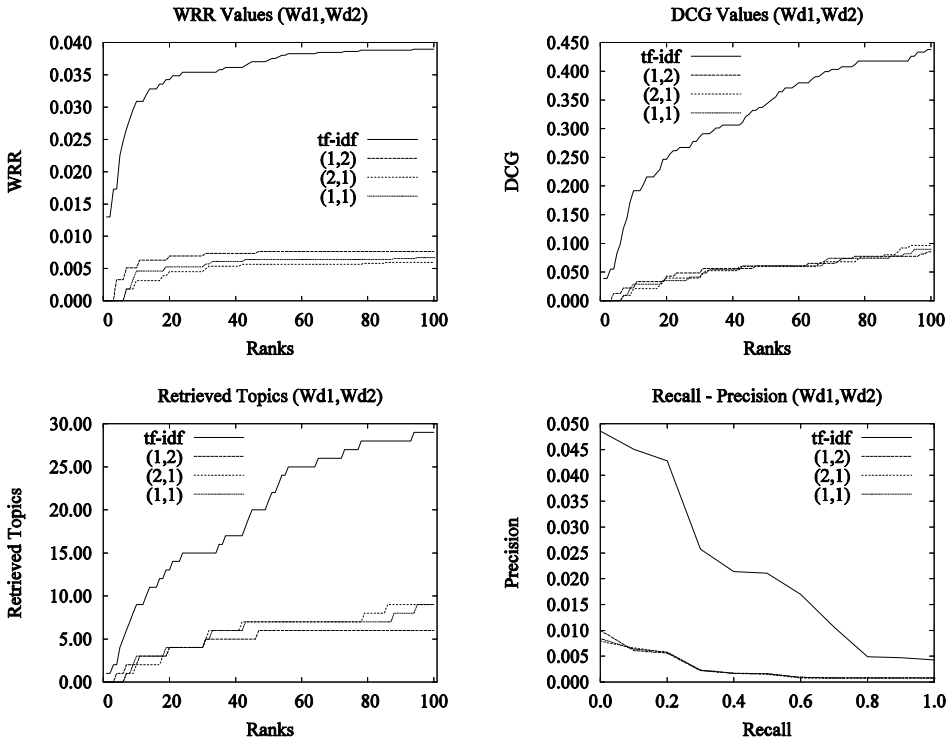


図 22 動的スコアを併合した場合の評価結果

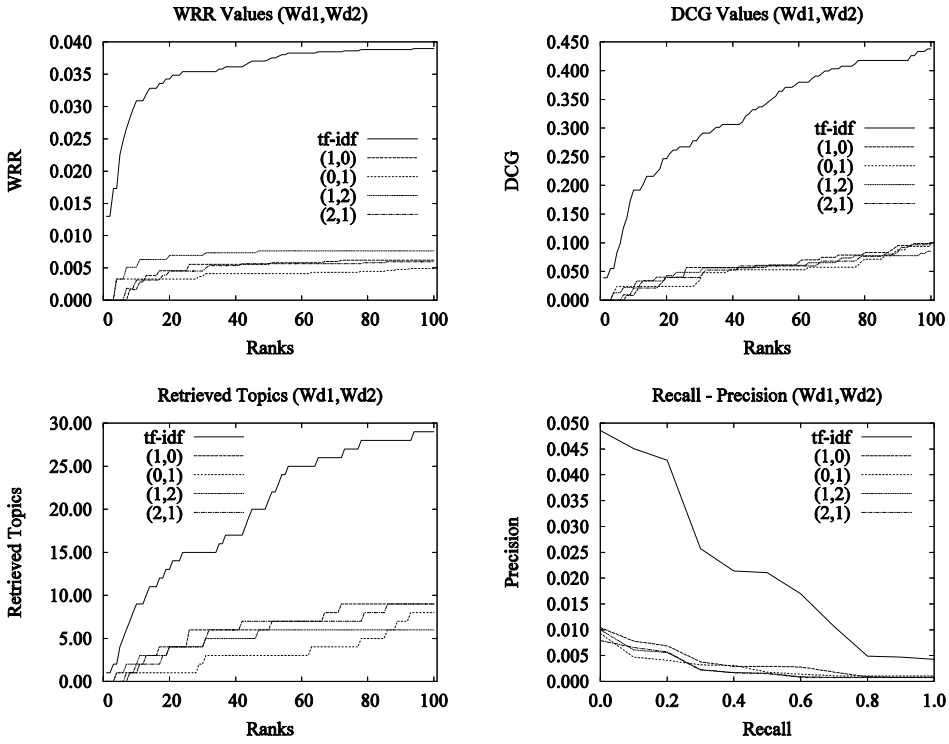


図 23 併合有無による評価結果比較

8.6.2. 全文検索スコアとの併合に関する実験

ここでは動的スコアリングにより得られた結果と全文検索スコアの併合実験を行った．併合には静的スコアリングの実験で使用したものと同様の併合式を用いた．併合式を以下に示す．

$$Score(p) = w_r \cdot Retrieval(p) + w_{d1} \cdot Dynamic\#1(p) + w_{d2} \cdot Dynamic\#2(p)$$

評価結果を図 24，および図 25 に示す．結果より，全文検索スコアのみによる評価が最も良いことがわかる．提案手法では 11 点平均適合率グラフより，適合率では動的スコア#1 が，再現率では動的スコア#2 がそれぞれ良い評価であり，動的スコアを併合したパターンは評価が悪いことがわかる．

WRR グラフでは，動的スコア#1 のみのケースが良い評価を得ていることがわかる．動的スコア#2 については，動的スコアのみによる評価結果(図 21)での特徴であった，上位ランクでの優位性が失われていることがわかる．図 24 と図 21 を比較した場合，WRR グラフにおいて，全文検索スコア併合により動的スコア#1 のグラフ全体が左上方向に移動していることがわかる．これは，動的スコア#1 が上位に抽出していた適合文書と全文検索スコアが上位に抽出していた適合文書が同じであったため，併合によって適合文書のスコアが底上げされた結果であると考えることができる．逆に動的スコア#2 については，全文検索スコアが上位に抽出していた適合文書と異なる文書を上位に抽出していたために，併合による適合文書のスコア底上げ効果を得られなかったと考えることができる．この考えのイメージを図 26 に，全文検索結果と動的スコア#1，#2 より決定したランクを比較した結果を表 15 にそれぞれ示す．この表では，「動的スコア#1，#2 により決定されたランク」から「全文検索結果により決定されたランク」を減算した値をランクの差とした．この結果より，動的スコア#1 についてはランク差が小さく，全文検索結果の方がより高ランクになっていることが，動的スコア#2 についてはランク差が大きく，全文検索結果の方が低ランクになっていることが確認できる．

また動的スコア#1，#2 を併合したパターンは評価が悪く，全文検索スコア併合を行わない場合の評価結果（図 23）とはまったく異なった傾向を示していることがわかる．しかしこれについても上記同様，動的スコア#1，#2 それぞれの抽出した適合文書の差によるものであると考えられる．

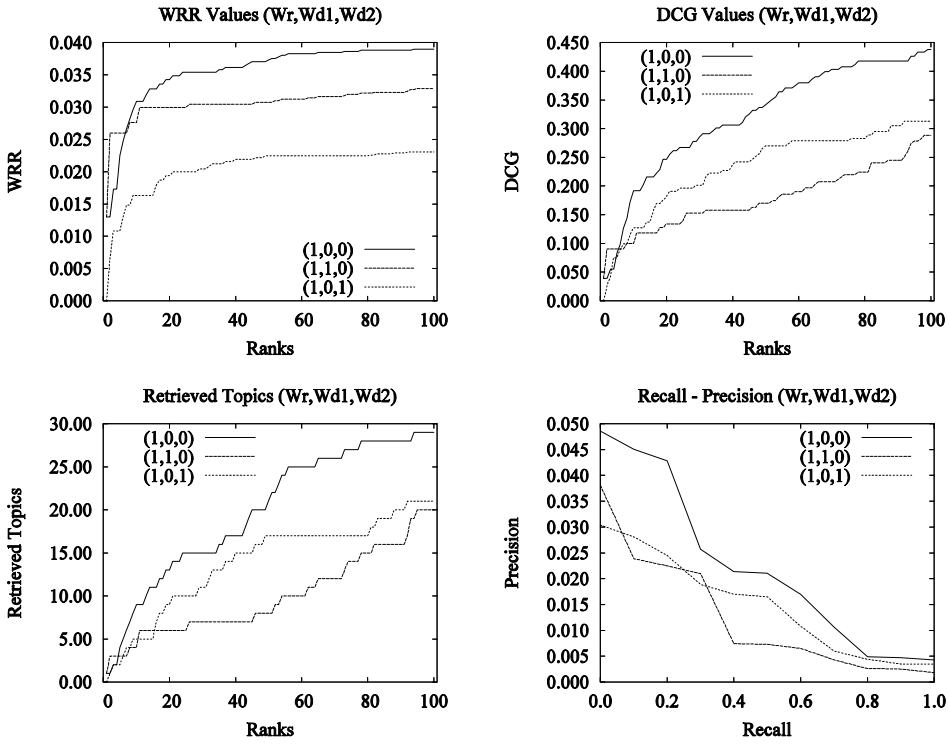


図 24 動的スコアリング評価結果(1)

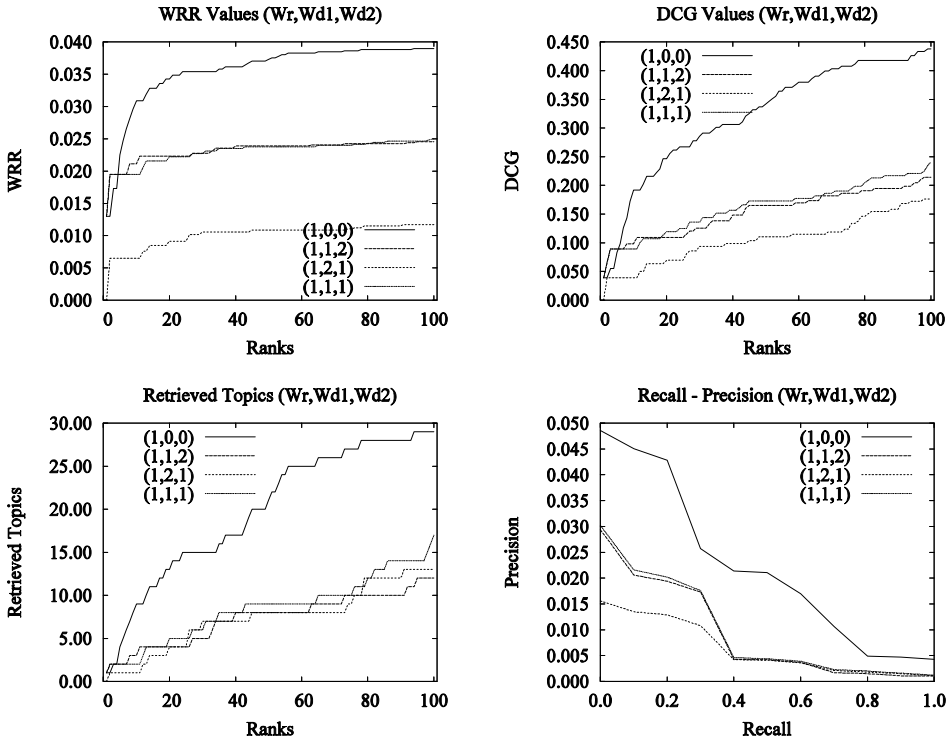


図 25 動的スコアリング評価結果(2)

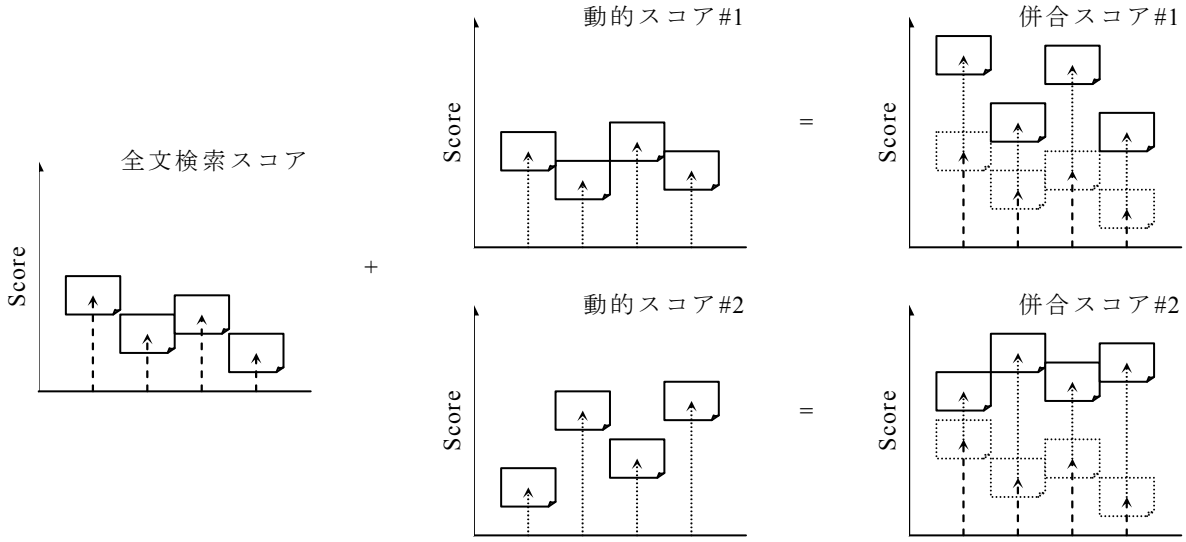


図 26 全文検索結果と動的スコア#1，#2 のランクイメージ

表 15 全文検索結果と動的スコア#1，#2 のランク差比較

	動的スコア#1	動的スコア#2
全文検索結果と同文書の抽出数	157	65
全文検索結果とのランク差合計値	5,712	-12,324
全文検索結果とのランク差平均値	38.3	-189.6

8.7. ランキング

ここでは，これまで検証してきた各手法を併合することによる評価の変化，および最適な重み係数調査に関する実験を行った．

8.7.1. 全スコア併合に関する実験

最終スコアの算出式は，静的スコアリング実験，および動的スコアリング実験で使用した算出式を併合したものとなる．算出式を以下に示す．

$$\begin{aligned} Score(p) = & w_r \cdot Retrieval(p) \\ & + w_{sn} \cdot StaticN(p) + w_{sg} \cdot StaticG(p) \\ & + w_{d1} \cdot Dynamic\#1(p) + w_{d2} \cdot Dynamic\#2(p) \end{aligned}$$

動的スコアの変化による評価結果の変動を確認するための実験として，重みを $(w_r, w_{sn}, w_{sg}) = (2,2,2)$ に固定した状態で，(w_{d1}, w_{d2}) を変化させた評価評価結果を図 27，図 28，および図 29 に示す．この結果より，動的スコアを併合した場合に精度が低下していることがわかる．これは，動的スコアの性能が他スコアに比べ大きく劣ることが原因であると考えられる．そこで各スコアの性能差を比較するため，適合課題についてランク分布の調査を行った．調査結果を図 30 に示す．この結果より，動的スコア#1 は適合率にばらつきがあり，動的スコア#2 は再現率が非常に低いことがわかる．動的スコア#1，#2 ともに，一部の適合文書については他手法より高精度ではあるが，全体的に再現率，適合率が低いため総合的に劣る結果となっていることがわかる．

動的スコア#1，#2 それぞれ単体で併合した場合の傾向として，動的スコア#2 の方が良い評価を得られており，動的スコアと全文検索スコアの併合実験で得られた結果と逆になっていることがわかる．これについては，平均化されていた動的スコア#2 の特性が，静的スコア併合により再びスコア差として現れた結果であると考えることができる．

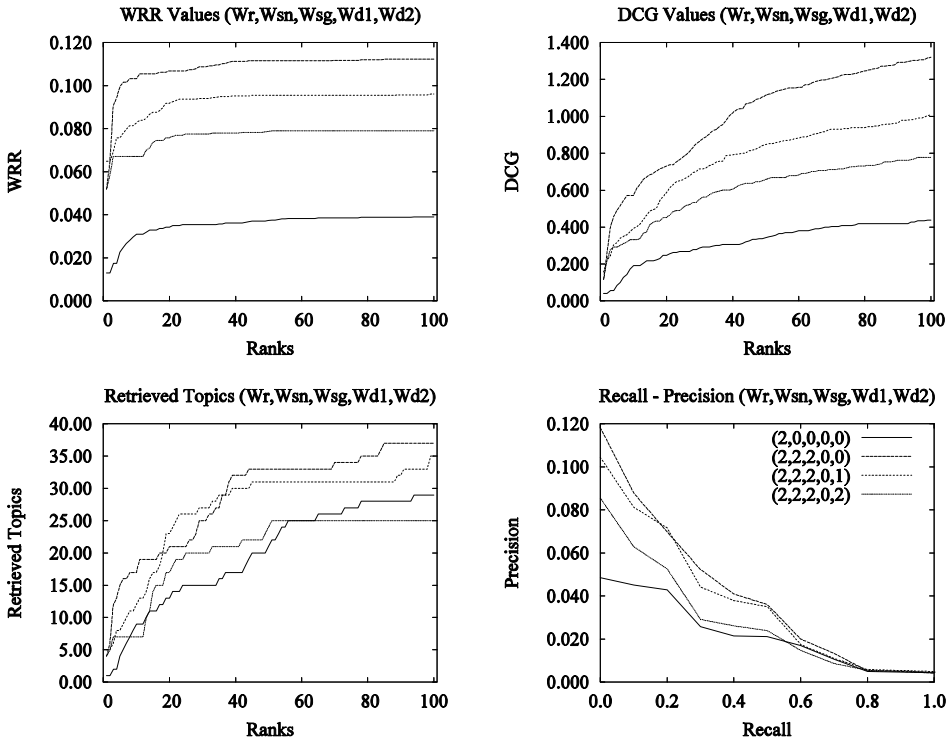


図 27 $(w_{d1}, w_{d2}) = \{(0,0),(0,1),(0,2)\}$ によるランキング評価結果

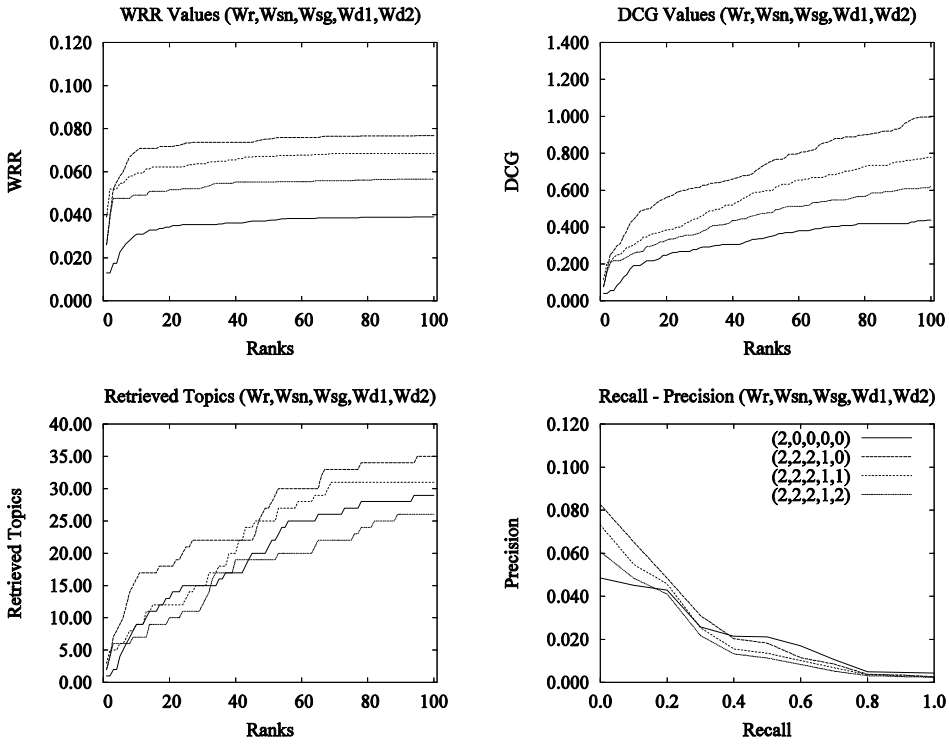


図 28 $(w_{d1}, w_{d2}) = \{(1,0),(1,1),(1,2)\}$ によるランキング評価結果

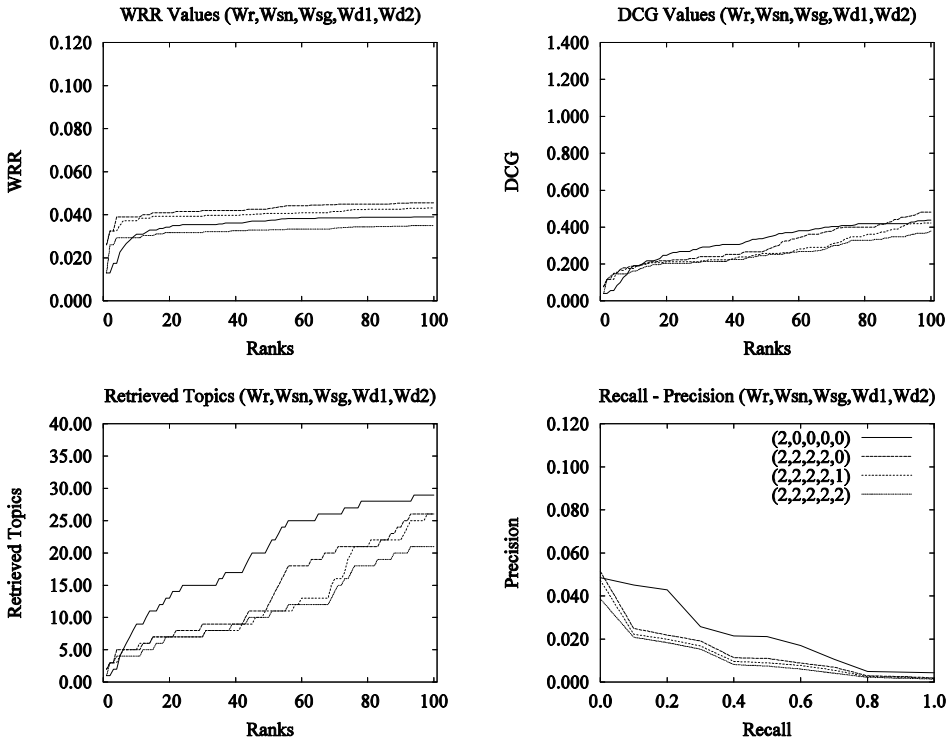


図 29 $(w_{d1}, w_{d2}) = \{(2,0),(2,1),(2,2)\}$ によるランキング評価結果

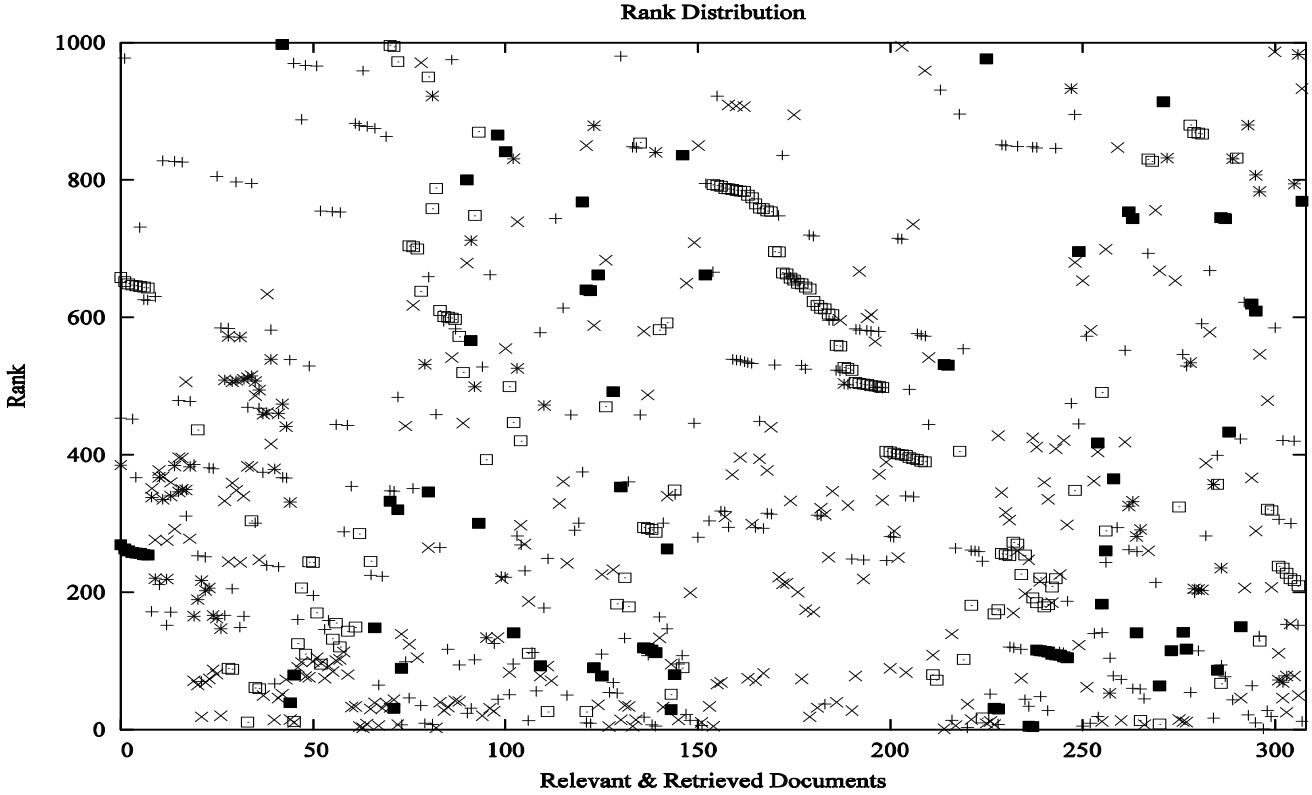


図 30 適合課題ランク分布

（＋: tf-idf, ×: StaticN, ＊: StaticG, □: Dynamic#1, ■: Dynamic#2）

8.7.2. 重み係数と評価結果に関する実験

最適な重み係数を決定するため，各スコアにかかる重み係数が(0,1,2)の 3 値をとるものとして，全パターンの評価を行った. WRR 評価における実験結果上位 30 件の比較結果を図 31 に示す. 表中の×印は上位 100 ランク，＋印は上位 10 ランクをそれぞれ評価対象にした場合の WRR 評価結果を表す. 結果より，全文検索スコアと静的スコアを併合したパターンにて，ほぼ等しい重み係数与えた場合に最も良い評価が得られることがわかった. 実験結果のうち，WRR 評価における上位 3 パターンを，全文検索スコアのみによる評価とともに図 32 に示す. 全文検索スコアのみによる精度に比べ，検索精度が大幅に向上していることがわかる.

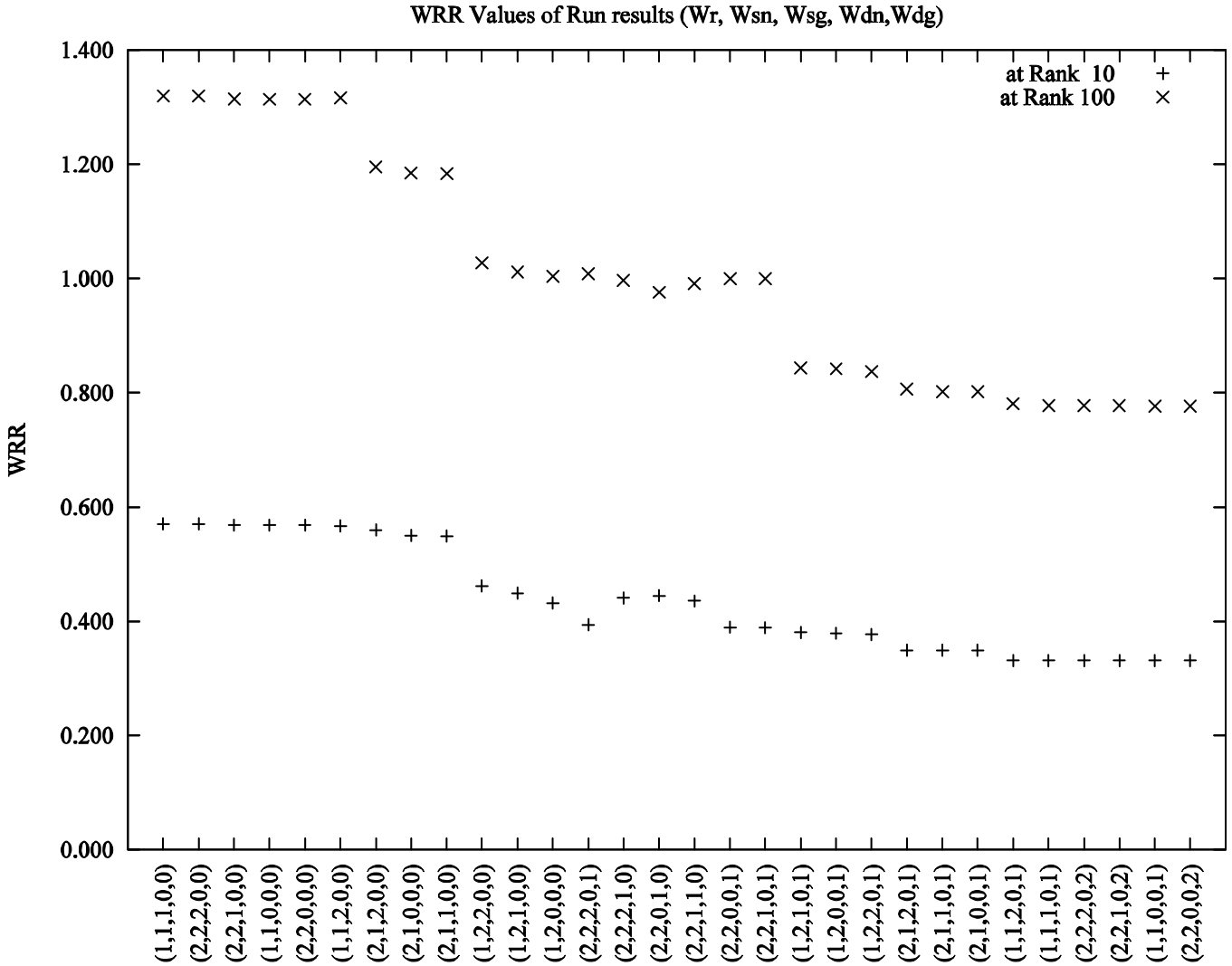


図 31 WRR 評価結果上位 30 件比較結果

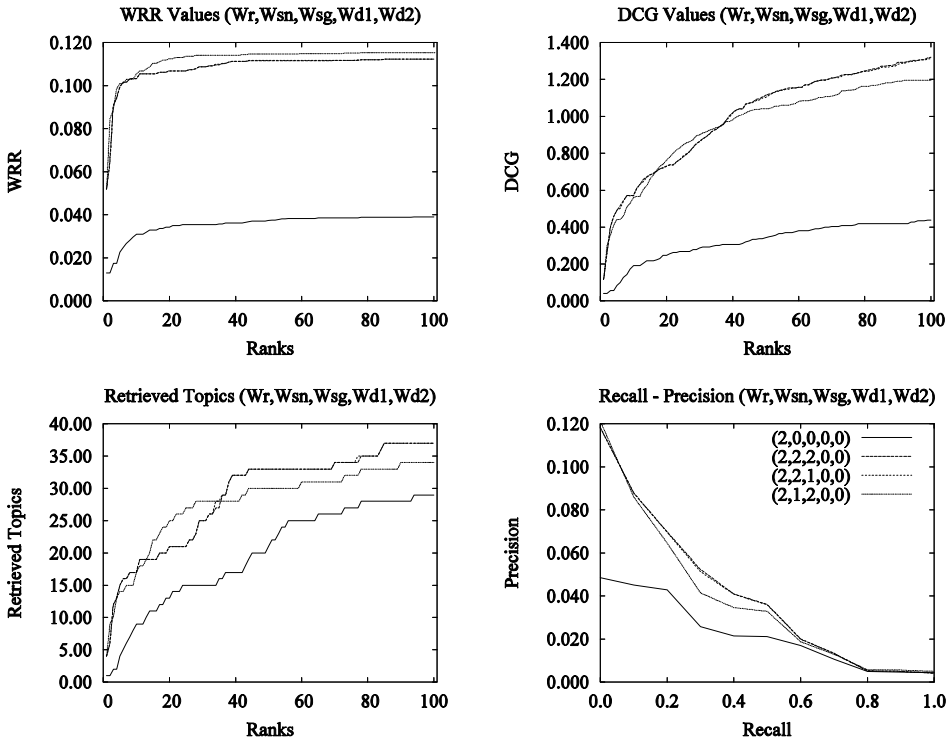


図 32 提案手法による評価上位 3 パターンの評価結果

9. 考察

各実験結果に対する考察を以下に述べる.

- ✓ グループ化
  - ➢ グループ化を行った状態でリンク構造解析によるスコアリングを行った場合に,グループ化を行わない状態でのスコアリング結果とは異なった適合文書を抽出することができた.これにより,リンク構造上隣接関係の拡張手法としてグループ化が有効であることを確認できた.しかし,ディレクトリ方式によるグループ化ではグループあたり Web ページ数のばらつきが非常に大きくなってしまった.これは被リンク数のばらつきに影響するため,最終的には検索精度に影響を及ぼすと考えられる.グループ化は提案手法全体に影響する処理であるため,より精度の高いグループ化手法を検討する必要がある.
- ✓ 静的スコアリング
  - ➢ グループ化有無それぞれのリンク構造解析スコアを併合することにより,グループ化を行わない場合のリンク構造解析スコアを上回る精度を得ることができた.
- ✓ 動的スコアリング
  - ➢ 全体的に精度が低下する結果となった.精度低下の原因としては,適合率,再現率ともに低いスコアリング結果しか得られなかったことにあると考えられる.再現率については,リンク構造解析スコアリングに必要な情報であるリンク数を増加させることにより向上させることが可能であると思われる.リンク数を増加させる簡単な方法として全文検索結果集合を大きくとることが挙げられる.
  - ➢ 動的スコアリング#2 についてはグループ化手法の精度による影響を受けるため,今回の実験では有効性を確認することができなかった.グループ化手法を再検討した後に再実験を行う必要があると思われる.
- ✓ ランキング
  - ➢ 全スコア併合による評価結果では,全文検索スコアと静的スコアをほぼ等倍で加算した際に高評価を得ることができた.その際の評価を全文検索スコアのみを利用したスコアリング結果と比較した場合,WRR および DCG では 200%程度,累積適合検索課題数では 30%程度,再現度 0.0 における適合率では 100%程度,それぞれ精度向上を確認することができた.

10. おわりに

本論文では，類似情報をもつ Web ページ群をグループ化することによりリンク構造上の隣接関係を拡張し，リンク構造解析スコアリングを静的，動的に算出する手法を提案した．また提案手法について実験検証を行い，提案手法による精度向上を確認するとともに，グループ化手法を再検討する必要があることを確認した．

今後は，グループ化手法自体の再検討として，本論文では実験を行わなかったリンク構造方式による手法のほか，ディレクトリ方式にリンク構造解析を加味する手法，代表性ヒューリスティックを URL 文字列に適用する方式などを検討していく．また，さらなる精度向上のため最終スコア算出式における各スコア重み係数の検証，スコア算出式自体の検証を行っていく．

謝辞

本研究を進めるにあたり，担当教官として終始御指導いただいた佐藤隆士教授をはじめとした諸先生方に心より感謝致します．ならびに良き助言をくださった各研究室の諸氏に対して厚く御礼申し上げます．

参考文献

[1] -, “Netcraft 社” 〈 http://news.netcraft.com/ 〉

[2] -, “総務省情報通信政策研究所” 〈 http://www.soumu.go.jp/iicp/ 〉

[3] -, “Yahoo!” 〈 http://www.yahoo.com/ 〉

[4] -, “Google” 〈 http://www.google.com/ 〉

[5] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” In *Proceedings of the 7th International World Wide Web Conference (WWW7)*, pp.107-117, 1998.

[6] L. Page, “The PageRank Citation Ranking: Bringing Order to the Web,” http://google.stanford.edu/~backrub/pageranksub.ps, 1998.

[7] J. M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment,” *Journal of the ACM*, vol.46, no.5, pp.604-632, 1999.

[8] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa, “Overview of WEB Task at the Fourth NTCIR Workshop,” *In Working Notes of the Fourth NTCIR Workshop Meeting*, pp.ov1-ov2, Tokyo, Japan, 2004.

[9] K. Eguchi, K. Oyama, A. Aizawa, and H. Ishikawa, “Overview of the Information Retrieval Task at NTCIR-4 WEB,” In *Working Notes of the Fourth NTCIR Workshop Meeting*, pp.ov3-ov15, Tokyo, Japan, 2004.

[10] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama, “Overview of the Web Retrieval Task at the Third NTCIR Workshop,” *NII Technical Report*, NII-2003-002E, 2003.

[11] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama, “Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure,” *IEICE Transactions on Information and Systems*, vol.E86-D, No.9, pp.1804-1813, 2003.

[12] Kalervo Järvelin and Jaana Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.41-48, New York, 2000.

[13] E. Voorhees, “The TREC-8 Question Answering Track Report,” In *Proceedings of TREC-8, NIST Special Publication 500-246*, pp.77-82, 1999.

[14] -, “Text Retrieval Conference (TREC)” 〈http://trec.nist.gov/〉

[15] T. Sato, T. Satomoto, and K. Han, “NTCIR-3 PAT Experiments at Osaka Kyoiku University,” In *Working Notes of the 3rd NTCIR Workshop Meeting Part III: Patent Retrieval Task*, pp.21-24, Tokyo, Japan, 2002.

[16] S.E. Robertson and S. Walker, “Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval,” In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[17] 中窪仁, 居行和, 佐藤隆士, “Web 検索におけるリンク構造解析,” 平成一五年電気関係学会関西支部大会講演論文集, no.G12-3, p.G241, Nov. 2003.

[18] 中窪仁, 居行和, 佐藤隆士, “Web 検索におけるリンク構造解析 －Web サイトのグループ化と動的スコアリングー,” 電子情報通信学会第 15 回データ工学ワークショップ, Mar. 2004.

[19] H.Nakakubo, P.Zhang, and T.Sato, “NTCIR-4 WEB Experiments at Osaka Kyoiku University -Static/Dynamic Scoring Using Link Structure Analysis and Web Page Grouping-,” In *Working Notes of the Fourth NTCIR Workshop Meeting (Vol. Supl.1)*, pp.22-25, Tokyo, Jun. 2004.

[20] 中窪仁, 佐藤隆士, “Web 検索におけるリンク構造解析を利用したランキング法,” 信学技報 Vol.104 No.177, no.DE2004-65, pp99-103, Jul. 2004.

[21] 中窪仁, 佐藤隆士, “Web 検索におけるページのグループ化,” 平成十六年電気関係学会関西支部連合大会講演論文集, no.G12-02, p.G282, Nov. 2004.