検索エンジンにおける繰り返し検索に対する文書の順位付け

酒井 義文[†]，佐藤 永欣[††]，上原 稔[††]

[†]東北大学大学院農学研究科

[††]東洋大学工学部情報工学科

現在インターネットで利用されている検索エンジンのほとんどは，検索結果における文書の順位付けに文書の価値（の推定値）に関する降順を採用している．しかし，同じ検索質問で繰り返し検索を行った場合，上位に順位付けされる文書がほとんど既閲覧であることも多い．一方，文書の公開時刻の降順をオプションとして採用している検索エンジンも存在する．しかし，検索のタイミングによっては，価値の低い文書ばかりが上位に順位付けされることもある．そこで，これら二つの方式の中間の性質をもつ順位付けの方式を提案し，繰り返し検索における有効性を計算機シミュレーションにより検証する．

Ranking Documents for Repeated Retrieval in Search Engines

Yoshifumi Sakai[†], Nobuyoshi Sato[††], and Minoru Uehara[††]

[†]Graduate School of Agricultural Science, Tohoku University

[††]Department of Information and Computer Sciences, Toyo University

Most of search engines nowadays used in Internet adopt the decreasing order according to (estimated) value of retrieved documents for displaying the result. However, in case of repeated retrieval with the same query, many of previously highly ranked documents, which were browsed by the user before, are still highly ranked. On the other hand, there exist search engines which adopt, as an option, the decreasing order according to published time of documents. But, using this option, most of highly ranked documents might have little value. This paper proposes a ranking method which has the intermediate characteristics of both rankings mentioned above, and test the effectiveness of the method for repeated retrieval by computer simulations.

1 はじめに

インターネットで用いられている検索エンジンのほとんどは，ユーザから指定された検索条件を満たす文書の一覧表を検索結果として出力する．一般に，多くのユーザは一覧表の上位から順に文書を閲覧するため，一覧表における文書の順位付けはユーザにとっての利便性に大きく影響する．

現在インターネットにおいて利用されている多くの検索エンジンは，ユーザによって指定された検索質問による検索要求に対して，*tf*·*idf* 重み付け [1] や PageRank[5] などを用いて文書の価値の推定値を求め，その降順に検索結果の文書を順位付ける．この順位付けのもとでは，価値が高いと推定される文書がネットワーク上に新規に公開されない限り，検索結果の上位に順位付けされる文書は変化しない．したがって，同じ検索質問による２回目以降の検索に

おいて，たとえ前回の検索時以降に新しい文書が公開されたとしても，それらの文書が上位に順位付けられる保証はない．

一方，公開されてからの経過時間の長さの昇順に検索結果の文書を順位付ける方式では，前回の検索時以降に新しい文書が公開された場合に，次回の検索において未閲覧の文書が最上位に順位付けられる．この順位付けの方式による検索をオプションとして提供している検索エンジンとしては，フレッシュアイ [2]，goo[3] などがある．しかし，この方式では文書の価値を考慮しないため，たとえ前回の検索時以降に価値の高い文書が公開されたとしても，検索をするタイミングによって，その文書が公開された後に公開された価値の低い文書のみが上位に順位付けられる可能性がある．

検索エンジンにおける第一義の目的を，ユーザに対してネットワーク上に存在する未閲覧かつ価値の高い文書を紹介することであるとするならば[*1]，上に述べた両者の順位付けの方式は，未閲覧であるかどうかという条件と，価値が高いかどうかという条件のどちらか一方のみを極端に重視したものであるとみなすことができる．すなわち，前者の順位付けは文書の価値の高さのみを重視し，その文書がユーザにとって未閲覧かどうかは考慮されていない．しかし，前回の検索以降に公開された文書が少数ならば，後者の順位付けを採用したほうが，たとえ価値は低くても未閲覧の文書をユーザに提供できる可能性がある．一方，後者の順位付けは文書が未閲覧であることを重視し，その文書の価値の高さは考慮されない．しかし，前回の検索時以降に公開された文書が十分に多数であるならば，前者の順位付けを採用したほうが，より価値の高い未閲覧文書をユーザに提供できる可能性がある．

これらのことから，前回の検索時以降に公開された文書数に応じて前者と後者の順位付けの方式をうまく使い分けることができれば，ユーザに対してより高い頻度で未閲覧かつ高い価値をもつ文書を提供できると推察される．本研究では，パラメタの設定により前者と後者の中間の性質をもたせることのできる順位付けの方式を導入し，前回の検索時以降に公開された文書数に応じてパラメタの値を設定し直すことにより，ユーザに対してさらに高頻度で未閲覧かつ高い価値をもつ文書を提供するための方法について議論する．

[*1] 本稿では，ユーザにとって既閲覧の文書は検索エンジンを介することなく常時利用することが可能であり，未閲覧の文書は検索エンジンを介して閲覧するまで利用することが不可能であるような単純な状況のみを考える

2 時間経過重みによる順位付け

ネットワークに i 番目に公開された文書を d_i とし，t_i を文書 d_i の公開された時刻，v_i を d_i の価値（の推定値）とする．$w : [0,\infty) \to [0,1]$ を単調減少関数とする．このとき，現在の時刻 t に対して，

$$v_i \cdot w(t - t_i)$$

の降順に各文書 d_i を並べる順位付けを考える．

この順位付けにおいて，w として任意の実数に対して常に値として 1 をとる関数を設定すると，価値 v_i の降順に文書 d_i を並べる順位付けとなる．一方，任意の時刻 t において，単調減少関数 w が存在して，$t_i \le t$ を満たす任意の $i \ge 2$ に対して，

$$v_i \cdot w(t - t_i) \ge v_{i-1} \cdot w(t - t_{i-1})$$

が成立する．したがって，この条件を満たす w が設定されている場合は，公開時刻 t_i の降順に文書 d_i を並べる順位付けとなる．この条件を満たす w は，一般に，入力の値の大きさが増加するにしたがって急速に小さな値をとる．この意味で，常に値として 1 をとる w とは対極をなすものである．これらの二つの w の中間の性質をもつ w に対しては，v_i の降順，t_i の降順とは異なる順位付けとなる．

単調減少関数 $w : [0,\infty) \to [0,1]$ の集合 W と，関数 $\mathrm{w} : [0,1] \to W$ が与えられると，0 と 1 の間の実数値パラメタ q によって，$v_i \cdot \mathrm{w}(q)(t - t_i)$ の降順に文書 d_i を並べる順位付けが定まる．以降では，素朴な設定として，パラメタ q に対して，$\mathrm{w}(q)(t - t_i) = q^{t-t_i}$ である場合に限定して考える．$v_i \cdot q^{t-t_i}$ の降順に文書 d_i を並べる順位付けを Exp_q で表す．したがって，

Exp_1 は v_i の降順に文書 d_i を並べる順位付けを表す．また，便宜的に $t-t_i < t-t_j$ ならば $v_i \cdot 0^{t-t_i} > v_j \cdot 0^{t-t_j}$ が成立すると仮定し[*2]，Exp_0 で t_i の降順に文書 d_i を並べる順位付けを表す．

3 繰り返し検索のモデル

繰り返し検索における順位付けの有効性に関する特徴を，以下のような仮想的なモデルのもとで計算機シミュレーションにより検証する．

任意の正整数 i に対して，$t_i = i$，すなわち，文書 d_i は時刻 i に公開されるものと仮定する．また，d_i の価値は公開時に $1/2^{v_i+1}$ の確率で非負整数 v_i に定まるものとする．すなわち，価値が $0, 1, 2, \ldots$ の文書が公開される確率は，それぞれ $1/2, 1/4, 1/8, \ldots$ のように指数関数的に減少する．ユーザは検索をするたびに最上位に順位付けられた文書のみを閲覧するものとし，文書の順位付けの有効性を，時刻 t までにユーザが閲覧した未閲覧文書の価値の総和 s_t によって比較する．

文書 d_i の価値 v_i に関する仮定は，現在の検索エンジンのほとんどが，文書の価値を内容が検索質問の条件を意図するか否かによって判定するのではなく，文書中に記述されている文字列の統計量や文書同士の検索質問に依存しないリンク関係を用いて判定する方式を採用しているため，検索結果にユーザにとってほとんど価値のないものが多数含まれる可能性があるという現状を反映したものである．

4 シミュレーション

4.1 Exp_q の典型的な振舞い

Exp_q の q に関する特徴を概観するために，ユーザが r の倍数の時刻にのみ検索をする設定のもとで，Exp_1，$\mathrm{Exp}_{0.9}$，Exp_0 に対する s_t の値の典型例を図 1 に示す．グラフは上から順に $r = 1$，4，10，25 に対するものである．グラフ中の + 印は各時刻 i におい

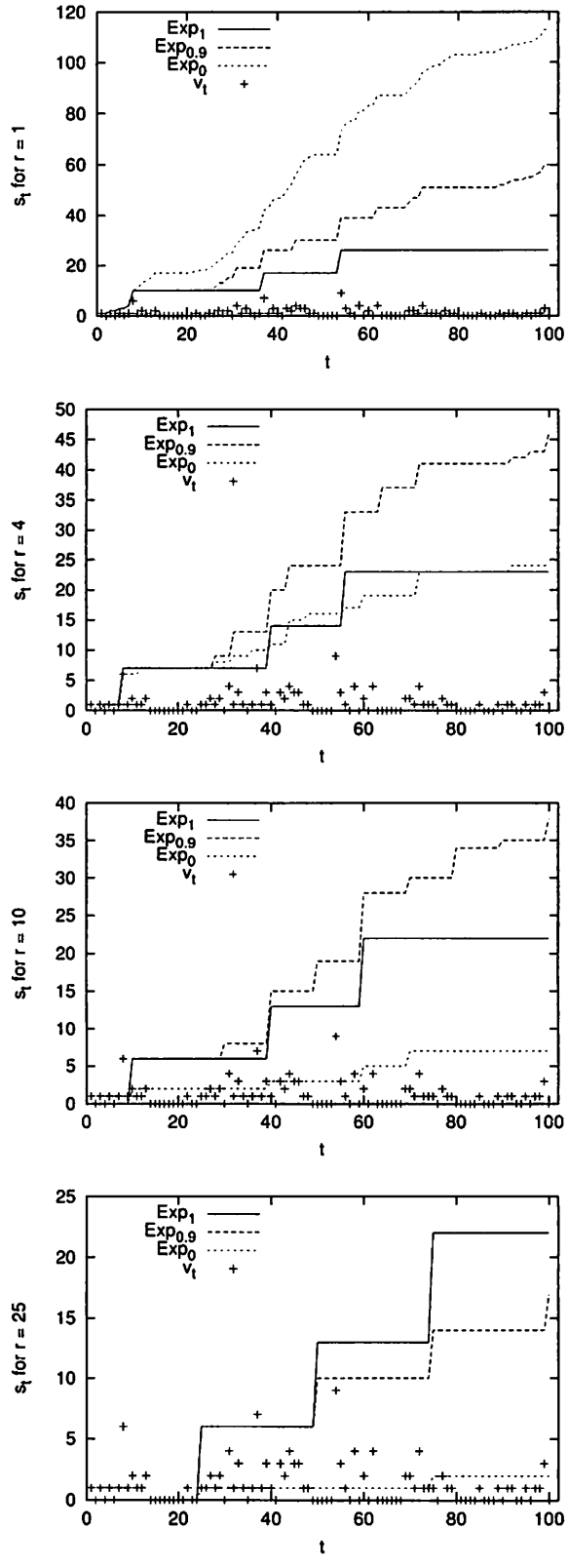

図 1 $r = 1$，4，10，25 に対する Exp_1，$\mathrm{Exp}_{0.9}$，Exp_0 のもとでの s_t の振舞い

[*2] q の値が十分 0 に近いとき $v_i \cdot q^{t-t_i} > v_j \cdot q^{t-t_j}$ が成立することに注意されたい．

て公開された文書 d_i の価値 v_i の値を表し，比較のため各グラフはすべて同一の系列を用いている．

$r = 1$ の場合（図1の一番上のグラフ）は，ユーザが文書が新しく公開されるたびに検索を行う仮想的な状況を再現したものである．

i の降順に文書 d_i を並べる順位付けである Exp_0 のもとでは，ユーザは検索のたびに公開されたばかりの新しい文書を見逃すことなく閲覧できるため，s_t の値は常に時刻 t までに公開されたすべての文書の価値の総和となる．一方，v_i の降順に文書 d_i を並べる順位付けである Exp_1 のもとでは，前回の検索までに公開された文書の中で最大の価値をもつ文書と比較して同等，あるいは，それを超える価値をもつ文書が公開されない限り，ユーザは未閲覧の文書を閲覧することができない．したがって，公開されるほとんどの文書が閲覧されることなく見逃されることになるため，s_t の値は Exp_0 の場合と比較して小さく抑えられる．図1の典型例においても，これらの傾向が確認できる．

また，$0 < q < 1$ に対する Exp_q については，その定義より，Exp_1 のもとでユーザが閲覧できる未閲覧文書はすべて Exp_q でも閲覧できる．また，ある時点において最上位に順位付けされた文書の価値と比較して，同等，あるいは，それを超える価値をもつ文書が以降に現れない場合であっても，十分に時間が経過すると最上位の文書が変化する可能性があるため，s_t は Exp_1 より大きな値をとる．しかし，十分な時間が経過する前に公開された価値の小さな文書は閲覧できないため，Exp_0 と比較すると s_t は小さな値となる．このように，$r = 1$ の設定のもとでは，$0 < q < 1$ に対する Exp_q は，Exp_0 と Exp_1 の中間的な性質をもつ．図1の典型例においても，$\mathrm{Exp}_{0.9}$ に対する s_t にこの傾向が確認できる．

$r = 25$ の場合（図1の一番下のグラフ）は，ユーザが検索を行う間隔が比較的長く，前回の検索との間にある程度多数の文書が公開される状況を再現したものである．グラフに示すとおり，s_t は Exp_1，$\mathrm{Exp}_{0.9}$，Exp_0 の順に大きな値をとり，$r = 1$ の場合と比較して逆転していることが確認できる．一般に r の値が十分大きな場合はこの傾向が現れる．これは，q の値が小さくなるにしたがって，たとえ前回の検索の後に価値の大きな文書が公開されたとしても，わずかな時間の経過によってその後に公開された価値の低い文書が最上位に順位付けられやすくなるためである．特に，Exp_0 のもとでは，検索時に公開された文書が無条件に最上位に順位付けられるため，たとえ検索の直前に価値の高い文書が公開されたとしても，ユーザはこれを閲覧することができない．

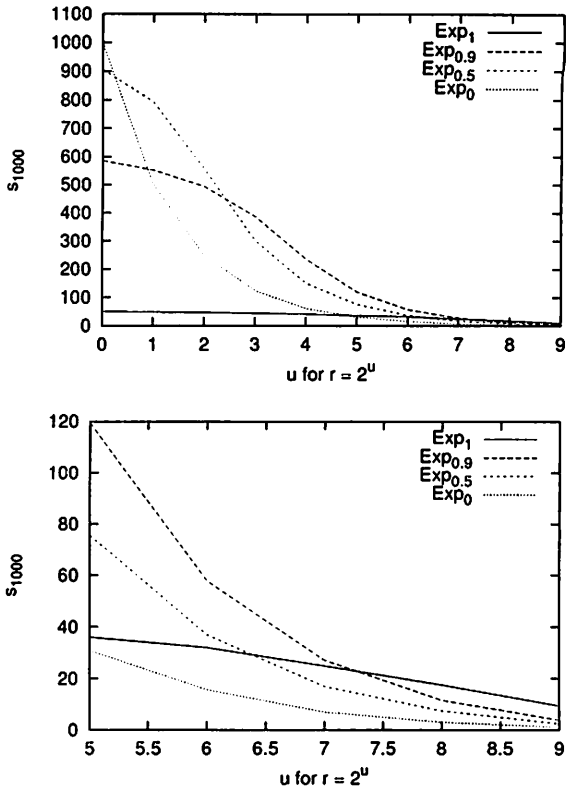

図2　Exp_1，$\mathrm{Exp}_{0.9}$，$\mathrm{Exp}_{0.5}$，Exp_0 のもとでの r に対する s_{1000} の1000回の試行による平均値

上に述べた二つの場合については，$r = 1$ の場合は Exp_0 が最も効果的であり，$r = 25$ の場合は Exp_1 が最も効果的であった．このことは，現実の検索において，ユーザが検索を頻繁に繰り返し，新しく公開される文書の個数が少数である場合は公開された新しい順に並べる順位付けが効果的であり，一方，検索の間隔が長く，新しく公開される文書が多ければ多いほど文書の価値の大きい順に並べる順位付けが効

果的であるという直感と一致する．また，$\mathrm{Exp}_{0.9}$ はどちらの場合も Exp_1 と Exp_0 の中間的な効果しか期待できない．

しかし，図 1 の典型例における 2 番目と 3 番目のグラフに示すように，r の値が極端でない場合には，Exp_1 と Exp_0 の中間的な特徴を示すのではなく，s_t の値が最も大きくなる傾向が現れた．このことから，r の値に対して最も効果的な Exp_q による順位付けが，Exp_1 と Exp_0 の二つで不連続に切り替わるのではなく，$0 \le q \le 1$ の範囲で連続的に変化すると考えられる．そこで，r の値を変化させた場合に s_{1000} がとる値の平均値を，Exp_1，$\mathrm{Exp}_{0.9}$，$\mathrm{Exp}_{0.5}$，Exp_0 に対して 1000 回の試行により求めたところ，図 2 のようになった．ただし，グラフの横軸は $r = 2^u$ における u の値である．この結果から，検索の間隔の長さによっては，Exp_1，Exp_0 のみを用いるのではなく，q が適切に設定された場合に，$0 < q < 1$ に対する Exp_q による順位付けがより効果的であることがわかる．

4.2 適応的 Exp による順位付け

前節のシミュレーションの結果より，効果的な順序付けを行うためには Exp_q における q の値をどのように設定したらよいかが問題となる．たとえユーザによる繰り返し検索が毎日決まった時刻，あるいは，毎週決まった曜日のように一定の周期をもってなされたとしても，一般に，新しい文書がネットワークに公開される頻度は一定ではないため，q の値を特定の値に固定できないからである．この問題に対する素朴な解決方法の一つとして，ユーザからの検索のたびに q の値を適応的に変化させた場合について，その効果を計算機シミュレーションによって検証する．

シミュレーションの設定は以下のとおりである．新しい文書がネットワークに公開される頻度に変化をもたせるため，文書 d_i は 1/4 の確率でのみネットワークに公開されるものとする．したがって，各時刻 i において，d_i は $1/8, 1/16, 1/32, \ldots$ の確率でそれぞれ $0, 1, 2, \ldots$ の価値をもつ文書として公開される．

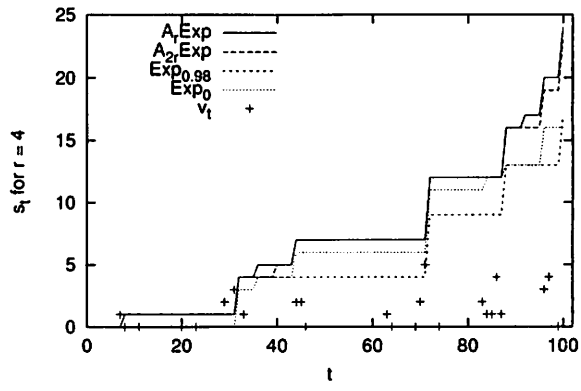

図 3 $r = 4$ に対する $\mathrm{A}_r\mathrm{Exp}$，$\mathrm{A}_{2r}\mathrm{Exp}$，$\mathrm{Exp}_{0.98}$，$\mathrm{Exp}_0$ のもとでの s_t の振舞い

文書 d_i が 0 の価値をもつ文書として公開される場合と，d_i がネットワークに公開されない場合において，Exp_0 の順位付けに違いが生じることに注意されたい．すなわち，時刻 $t = i$ における Exp_0 の順位付けのもとでは，前者の場合は価値 0 の文書 d_i が最上位に順位付けされ，後者の場合は時刻 t 以前において最後に公開された文書が最上位に順位付けられる．

$\mathrm{A}_r\mathrm{Exp}$ を，時刻 t において時刻 $t - r + 1$ から t までの間にたかだか一つしか文書が公開されなかった場合に Exp_0 を用いて文書を順位付けし，少なくとも二つ以上文書が公開された場合に $\mathrm{Exp}_{0.98}$ を用いて文書を順位付ける文書の順位付けを表す．したがって，前回の検索時以降に公開された文書の個数に応じて Exp_0 と $\mathrm{Exp}_{0.98}$ を適応的に使い分ける順序付けである．また，$\mathrm{A}_{2r}\mathrm{Exp}$ を，時刻の範囲を $t - 2r + 1$ から t までとする以外は $\mathrm{A}_r\mathrm{Exp}$ と同様の順位付けを表す．これはユーザの検索周期が一定ではないことを考慮し，適応の精度を落としたものである．以上の設定のもとで，4 の倍数の時刻にのみ検索を行った場合の s_t の値についてシミュレーションを行ったところ，q の値が固定された Exp_0 や $\mathrm{Exp}_{0.98}$ と比較して，$\mathrm{A}_{2r}\mathrm{Exp}$ や $\mathrm{A}_{2r}\mathrm{Exp}$ のもとで s_t はより良好な値をとる傾向があることが確認された．その一例を図 3 に示す．

5 公開時間指定検索との比較

未閲覧かつ価値の高い文書を上位に順位付ける最も有効な方法の一つとして，検索エンジンが各ユーザの検索結果の閲覧状況に関する履歴を保持し，過去に閲覧されたことのない文書のみを価値の高い順に並べる順位付けが考えられるが，不特定多数のユーザによって利用されることを想定した場合，検索エンジンがユーザ一人ひとりの検索の履歴をすべて保持するのは現実的ではない．これに代わる方法として考えられるのが，各ユーザの前回検索した時刻に関する情報を用いて，その時刻以降に公開された文書のみを検索の対象とする方法である．ユーザから入力として前回の検索時刻が与えられるならば，不特定多数のユーザ一人ひとりの前回の検索時刻に関する情報を保持することなく，そのユーザにとっての未閲覧文書のみを検索の対象とすることができる．ユーザから時刻に関する情報を入力として受け取り，その時刻以降に公開された文書のみを検索の対象とする公開時間指定検索をオプションとして採用している検索エンジンとしては，goo[3]，インフォシーク[4]などが挙げられる．公開時間指定検索において指定された時刻以降に公開された文書が価値の高い順に順位付けられるならば，この順位付けは未閲覧かつ価値の高い文書を上位に順位付ける．

公開時間指定検索のもとでの文書の価値の降順に並べる順位付けと適応的 Exp 順位付けにおける特徴の注目すべき違いは，未閲覧文書の選別に関する厳密さにある．適応的 Exp 順位付けのもとでは，文書の価値の高さと公開された時刻によっては，ユーザにとってすでに閲覧したことのある文書が上位に順位付けられる可能性がある．これに対して，公開時間指定検索における順位付けのもとでは，指定された時刻以前に公開された文書は確実に検索の対象外になるため，前回の検索時刻が正確に指定された場合に s_t は，適応的 Exp 順位付けと比較して同等，あるいは，それを超える大きな値をとる．この意味で，前回の検索時刻が正確に与えられるならば，適応的 Exp 順位付けと比較して，公開時間指定検索の順位付けがより効果的である．しかし，厳密さの代償として，前回の検索時刻より後の時刻が指定されると，前回の検索時刻と指定された時刻との間にどれだけ価値の高い文書が公開されていたとしても，その文書が検索結果に含まれることはなく，ユーザは閲覧することができない．一方，適応的 Exp 順位付けのもとでは，公開時間指定検索の順位付けのもとで検索漏れとなる文書であっても，その文書の価値が十分高いならば上位に順位付けられる．

6 まとめ

本研究では，検索エンジンの検索結果の文書における順位付けに関して，文書の価値の降順と文書の公開時刻の降順の中間的な性質をもつ順位付けの方式として Exp_q を提案し，仮想的なモデルのもとでの計算機シミュレーションにより，繰り返し検索における Exp_q 順序付けの有効性に関する特徴を検証した．とくに継続的に大量の文書が出現するトピックの繰り返し検索において，文書の価値順，公開時間順に代わる順位付けのオプションとして有効性を有するものと推察される．実際のネットワークにおいて最適な q を適応的に効率よく設定するための手法の開発が今後の課題である．

参考文献

[1] R. Beaza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval," ACM Press, New York, 1999.

[2] 株式会社ニューズウォッチ, "フレッシュアイ," http://www.fresheye.com/

[3] 株式会社エヌ・ティ・ティ エックス, "goo," http://www.goo.ne.jp/

[4] 株式会社インフォシーク, "インフォシーク," http://www.infoseek.co.jp/

[5] L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank citation ranking: bringing order to the Web," http://dbpubs.stanford.edu/pub/1999-66, 1998.