

アクセス履歴を考慮したファイル間の 関連度を用いたデスクトップ検索

井ノ口 伸人[†] 吉川 正俊[‡]

[†]名古屋大学大学院情報科学研究科 〒464-8603 名古屋市千種区不老町

[‡]京都大学情報学研究科 〒606-8501 京都市左京区吉田本町

E-mail: [†] inoguchi@dl.itc.nagoya-u.ac.jp, [‡] yoshikawa@i.kyoto-u.ac.jp

あらまし 近年、パーソナルコンピュータに保存可能なデータ量が飛躍的に増加しており、デスクトップを対象とする検索に対する需要が高まっている。デスクトップ検索はウェブ検索とは異なった性質を備えているが、既存のデスクトップ検索はウェブ検索に近い手法を用いているものが多い。本研究では、デスクトップ上のファイル間には関連が存在するという仮説に基づき、関連度のその強さに基づくデスクトップ上のファイル集合モデルを提案する。この関連度を利用者のファイル利用履歴から得られる、アクセスの順序や共起、密度を用いてモデル化した。また、本モデルを用いた、デスクトップ検索システムを作成し、関連度抽出実験を行った。

キーワード デスクトップ検索, アクセス履歴

The Desktop Search Using Relativity of Files Obtained from Access History

Nobuto INOGUCHI[†] Masatoshi YOSHIKAWA[‡]

[†] Graduate School of Information Science, Nagoya University Furo-cho, Chikusa-ku, Nagoya 105-0123 Japan

[‡] Graduate School of Informatics, Kyoto University Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501 Japan

E-mail: [†] inoguchi@dl.itc.nagoya-u.ac.jp, [‡] yoshikawa@i.kyoto-u.ac.jp

Abstract Quantity of information recorded in a personal computer has been increasing rapidly. It promoted needs of desktop search systems. Desktop search has different context from it of the web search. However, existing desktop search systems use the method like the web search systems. In this study, we propose a hypothesis that desktop files has some hidden links and modeled the link using information from history of file accesses, for example, access order, co-occurrence and density. Then, we implemented a desktop search system which finds files in new respect based on our model.

Keyword Desktop Search, Access History

1. はじめに

近年、ハードディスクの容量増加と価格低下に伴い、パーソナルコンピュータに保存可能な情報量が飛躍的に増加している。さらに、様々な情報が電子化され、パーソナルコンピュータ上に保存されるようになってきている。そのため、デスクトップ上に大量のファイルが存在するようになり、利用者自身による整理・分類が困難になりつつある。そのため、デスクトップから目的のファイルを発見するためのデスクトップ検索に対する需要が高まった。

デスクトップ検索は、大きく二つの点でウェブ検索や既存の文書検索と異なる。一つは利用者が求めている情報が既知であるかどうかである。ウェブ検索や文書検索が、未見の情報を対象としているのに対し、デスクトップ検索は、主に、既に見たことがある情報を対象としている。もう一つの相違点は、対象とするフ

ァイルの性質である。ウェブ文書はハイパーリンク構造を有し、情報間の関連が明示されている。また、文書検索は検索対象の分野が限定されている場合が多い。それに対して、デスクトップ上の情報は、種々の分野の情報が混在するにも関わらず、ファイルどうしの関連を明示する一般的な方法が存在しない。ディレクトリによって、あるファイル集合が同じ分類に所属するということが示されているに留まっている。しかしながら、既存のデスクトップ検索は、ウェブ検索や文書検索の手法を踏襲している部分が多い。

本研究では、デスクトップ上の情報にも、明示はされないが関連が存在するという仮説を立て、その強さを観点とした、デスクトップ上のファイル集合モデルを提案した。そして、その関連を、利用者のファイル利用履歴から得られる情報、特にアクセスの順序、共起、密度の三種類の情報を用いてモデル化した。さら

に、このモデルを用いて、既存のデスクトップ検索とは異なる観点に基づいたデスクトップ検索システムを実装し、評価を行った。

本論文の残りは以下のように構成される。2節で関連研究について述べる。3節では、デスクトップ上のファイルに特有の性質について述べる。4節で、デスクトップ検索の課題について説明する。5節で新しいファイル集合モデルについて示し、6節でそのモデル化に必要な具体的な関連度モデルを定義する。7節では、そのモデルを用いて行った関連度抽出実験について説明する。最後に8節、9節で今後の課題とまとめについて述べる。

2. 関連研究

2.1. デスクトップ検索

デスクトップ上の情報は利用者にとって既知であり、その整理・分類は利用者が管理することが前提であった。しかし、多様なデータが電子化され、パーソナルコンピュータに大量に保存されるようになった現在、人手による分類は困難になってきている。そこに、デスクトップ検索に対する需要が生じた。こうした経緯から、デスクトップ検索はウェブ検索や文書検索に比べて新しい分野である。

既存のデスクトップ検索として、Google デスクトップ検索 [2]、Windows デスクトップ サーチ [3]、Spotlight [4] が挙げられる。Google デスクトップ検索はデスクトップファイルに対しても履歴を保存している点が特徴である。Windows デスクトップ サーチ、Spotlight は ファイルに付与されたメタデータを利用している点が特徴として挙げられる。これらのデスクトップ検索システムはいずれも利用者から与えられた文字列に応じた結果を出力し、ファイル間の関連については考慮していない。その点が本研究と異なる点である。

2.2. パーソナライズド検索

パーソナライズド検索とは、利用者の選好に応じて検索結果がカスタマイズされる検索サービスのことである。あるいはコンテキストを反映した検索と表現することもできる。例えば、二人の利用者が同じ「フィッシング」という検索語を入力した場合、既存のウェブ検索は、利用者の意図に関わらず同一の結果を出力する。しかし、理想的にパーソナライズされていれば、一方は釣りに関する選好が強いので釣り (fishing) に関連したページが上位に表示されるが、一方はインターネット犯罪についての選好が強かったのでフィッシング (phishing) 詐欺に関するページが上位に表示される。この例から分かるように、パーソナライズド検索は、利用者ごとに適した結果を出力するための技術

である。

代表例として、Google パーソナライズド 検索 [8] が挙げられる。Google パーソナライズド 検索の場合、利用者の選好は、検索の際に入力された語や、検索結果のうち実際にクリックした結果から自動生成される。利用者のアクセス履歴を用いる点が類似しているが、対象とする文書集合が異なる点と、履歴を検索結果に補正をかけるためのみに用いている点において、履歴を情報の編成に用いる本研究と大きく異なる。

2.3. 情報の集約

デスクトップ上には様々な情報が体系化されないまま散在している。またある目的の情報が、一つのファイルに対応しているとは限らない。ある一つの情報が、複数のファイルに分散している場合がある。例えば、比較すべき二つの実験の出力ファイルが、異なったファイルとして存在している場合である。また、逆に目的の情報が、大きなファイルの一部として含まれていることもある。この場合、そのファイルが見つかったとしても、その中において検索を行わなければならない。

これらのように保存されている情報を繰り返し利用する場合、何度もファイル内検索をしたり、複数のファイルを見比べたりするなど煩雑な操作が必要になる。こうした問題を、情報を必要な部分だけ集約し、検索性の高い状態で管理することで解決しようというアプリケーションが、Office OneNote [5] や Google Notebook [6] である。これらは、本研究とは、手法は大きく異なるが、目的は本研究と同じデスクトップ上の情報へのアクセサビリティの向上である。

3. デスクトップ上のファイルの性質

デスクトップ上のファイルは、ウェブ文書や文書集合とは異なる性質を持つ。本節では、均質性、リンク情報の有無、完結性の三つの相違点について説明する。

3.1. 均質性

ここでは均質性を、文書長の分布と、言葉遣いが定型かどうかで評価する。

従来の文書検索が対象とする文書集合は、均質性が高い。典型的な例として論文の集合を挙げて説明する。論文は一定のルールに基づいて記述されている。そのため、誰が書いたかに関わらず文書長はほぼ一定である。また、公的なものであるため、言葉遣いも定型である。そのため、tf-idf モデルを始めとする単語の出現頻度を利用した特徴量によるモデルが、実際の文書集合をよく近似する。

これに対して、ウェブ文書は均質性が低い。制約が存在しないため、文書長が広範囲に分布している。また、公的な文書と私的な文書が混在しているため、言葉遣いのばらつきが大きい。実用上では、検索結果で上

位に表示されることを目的に恣意的にキーワードを埋め込んでいる文書の存在も問題となる。

デスクトップ上のファイルもまた均質性が低い。利用者によって状況は変わるため、一概には言えないが、制約が無い場合、文書集合と比較して、文書長が広範囲に分布している。またその分布も一様ではない。デスクトップ検索が対象とするものには画像や音楽も含まれる。それらの情報には文字列がほとんど含まれない。次に、言葉遣いについて考察する。デスクトップ上には他人に見せることを前提とした文章と、利用者自身だけが読むことを前提とした文章が混在するため、言葉遣いは非定型的である。

従って、ウェブ文書とデスクトップ上のファイルを、単語の出現頻度を利用した特徴量のみによって評価することは、妥当性に欠けると考えられる。

3.2. リンク情報の有無

本節においては、ある情報が、関連する他の情報をコンピュータ上で明示することが可能である場合に限り、その情報にはリンク構造を有するとする。

まず文書集合について考察する。3.1節で例として挙げた論文集合の場合は個々の論文が参考文献情報を含んでいる。しかし、その他の場合は一般には含まれない。

インターネットはハイパーリンクによるグラフ構造を基盤としている。従って、ウェブ文書は、ほぼ例外なくハイパーリンクを有している。そのため、その内容を評価に使用せず、グラフ構造だけから特徴量を計算することが可能である[9][10]。

デスクトップ上のファイルは、一般に、リンク情報を持つことが不可能である。ファイル間の関連は、ファイルそのものではなく、ファイル・システムが表現する木構造によって暗示されるに留まる。ただし、ショートカットやハードリンク、シンボリックリンクによって、ディレクトリとファイルに関連させることは可能である。しかし、それでも巡回有向グラフであるインターネットに比較して表現能力が低い。また、同フォルダ内のどのファイルとどのファイルが関連しているかを表現することはできないという問題も抱えている。特に同一フォルダ内に多数のファイルが存在する場合、リンク情報を保持することができないため、その中から少数のファイルを発見することは困難である。

3.3. 情報の完結性

本節でいう完結性とは、個々の情報がそれ自体で完結しているかどうか、言い換えれば、他に情報を必要しているかどうか、である。

文書集合に含まれる個々の情報の完結性は高い。一般の文書は、読者が、その著者が想定している知識を

持っていれば、参考文献を参照せずに、理解できるように構成されている。

対照的に、ウェブ文書には、完結性が低いものが存在する。ウェブ文書はリンク情報を持つため、それ自身で完結していなくとも、参照先に利用者を容易に誘導できるためである。ポータル・サイトや検索エンジンがその最たる例である。それらのサイトは、利用者の誘導を最大の目的としている。

デスクトップ上のファイルにも、完結性が低いものが存在する。特に利用者だけが見ることを前提としている情報は、利用者さえ把握していれば、まとめなくとも不都合は生じないため、完結性が低くなる傾向が強いであろう。しかし、デスクトップ上のファイルはリンク情報を保持できないため、個々のファイルだけでは、情報の価値が低下する場合がある。例えば、実験の出力ファイルは、それ単体では大きな意味を成さない。評価、考察を記述したファイルと共にあって、初めて意味をなす。

4. デスクトップ検索の課題

本節の目的は、利用者の目的と、3節で述べたデスクトップ上のファイルの性質から、デスクトップ検索の満たすべき要件を導くことである。

まず、利用者がどのような意図を持ってデスクトップ検索を行っているかについて、ウェブ検索と対比させて説明する。

ウェブ検索における利用者の要求は、抽象的である。検索を行う時に、特定のウェブ・ページを結果として期待していることは少ない。期待しているのは、特定の情報である。そのため、欲している情報が存在するかどうかは利用者の最大の関心事になる。この傾向は、同じ情報が記述された文書が複数存在することに関連しているであろう。つまり、利用者にとって、正解文書は同じ情報を持つ他の文書と代替可能である。

一方、デスクトップ検索においてはそうではない。利用者はどこかに保存したと考えているある特定のファイルを求めているので、正解は代替不可能である。また、多数の人間が生成するウェブとは違い、個人が利用するコンピュータ上において、同じ情報が複数のファイルに存在していることは少ないであろう。このことも、正解が代替不可能であるという性質を強めている。

3節で述べたように、デスクトップ上の情報は十分体系化されていない。そのため、目的の情報が一つファイルに含まれているとは限らない。また、リンク情報を持たないため、一つのファイルを手掛りに他のファイルを発見することは困難である。従って、デスクトップ検索における利用者の目的は、特定のファイル

の集合を発見することである。デスクトップ上の情報は均質性が低いため、目的のファイル集合に含まれるファイルのいずれにも同じ語が含まれているとは限らないことに注意されたい。

5. ファイル集合モデル

本節では、デスクトップ上のファイル集合モデルについて説明する。3節で説明したように、デスクトップ上のファイルは、完結性が低く、かつリンク情報を持たないため、複数のファイルに分散した情報を再発見することが難しい。そこで、我々は、各ファイルは、明示的なリンク情報を持たないが、意味においては関連を持つファイルが存在すると考えた。デスクトップ上のファイル集合を、関連した情報の集合と捉えると、ファイル集合は、重み付き有向グラフで表現される(図1)。ファイル集合をモデル化したグラフは以下に挙げる性質を持つ。

- 各頂点はファイルと一対一に対応する。
- 各辺は重みを持つ。
辺の存在は関連の存在を示す。その重みは関連の強さを表現する。
- 頂点 a から頂点 b への辺 e_1 が存在するならば、頂点 b から頂点 a への辺 e_2 も存在する。
頂点 a がファイル A に、頂点 b がファイル B に対応する。このとき、ファイル A がファイル B と関連を持つならば、ファイル B はファイル A と関連を持つ。

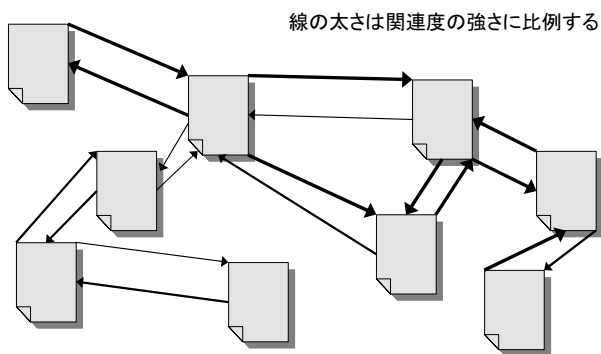


図 1 ファイル集合の概念図

また、各辺の重みを決定するための重み関数 C は、以下の性質を持っていないなければならない。

- 関連の強さと正の相関を持つ。
- 一般に、 $C(A, B) \neq C(B, A)$ である。
ファイル A に対するファイル B の関連度と、ファイル B に対するファイル A の関連度は異なる。

二つ目の性質について説明する。例えば、実験結果の出力ファイルに対する評価や考察を記述したファイルの関連は強いが、評価や考察には他の実験について

も言及しているため、これに対する一つの実験結果の出力ファイルの関連の強さは低下する。

本研究では、この関連度を、利用者のアクセス履歴から得られる情報に基づいて計算する。履歴情報を用いることで、利用者の利用状況を反映したファイル集合モデルを動的に生成することが可能になる。

6. 関連度

本節では、利用者のファイル利用履歴から得られる情報を用いて、関連度を算出するための具体的な重み関数について説明する。

6.1. 順序モデル

順序モデルでは、関連は、アクセスの順序と強い相関があると考えられる。ファイル A に対するファイル B の関連度を、利用者がファイル A にアクセスした時、その直前あるいは直後にファイル B にアクセスした割合で定義する(式1)。

$$C_0(A, B) = \frac{A \text{ の前後での } B \text{ のアクセス数}}{A \text{ のアクセス数} \times 2} \quad (1)$$

順序モデルには、各アクセスにおいて、関連を取得できるのは、最大3ファイルまでである。そのため、4つ以上のファイルが相互関連している場合、それを発見しにくいという問題がある。

6.2. 時刻モデル

時刻モデルでは、関連は順序よりも時刻と強い相関があると考えられる。ファイル A とファイル B のアクセス時刻が近ければ近いほど関連が近いと見なす。A へのアクセス時刻を 0、B へのアクセス時刻を t としたとき、 $t < t_1$ において最大値 1 をとる。以降は、二次関数で減少し、 $t = t_2$ において 0 に達するという定式化を本研究では採用した(式2)。

$$C_t(A, B) = \begin{cases} 1 & (0 \leq t < t_1) \\ \frac{1}{(t_1 - t_2)^2} (t - t_2)^2 & (t_1 \leq t < t_2) \end{cases} \quad (2)$$

式2は A へのある1度アクセスに対する関連度の定義である。ある期間 T における A に対する B の関連度を、式2の平均で定義する(式3)。

$$C_T(A, B) = \frac{\sum_T C_t(A, B)}{A \text{ のアクセス数}} \quad (3)$$

時刻モデルは、ほぼ同時に多数のファイルがアクセスされた場合、それらのファイルの関連度が大きくなる。そのため、ランク上位がそれらのファイルによって占められてしまい、その他の正解ファイルのランクが大きく下がるという問題を持っている。

6.3. アクセス密度

アクセス密度 n を、あるファイルのアクセス時刻か

ら、時刻 t 以内にアクセスされたファイル数として定義する。我々は、密度の大きいアクセスがあった場合、それぞれのファイルの関連度は、密度の小さい場合より小さくなると考えた。密度の概念を順序モデルと時刻モデルに適用し、それぞれのモデルが持つ問題を解決することを試みた。

6.3.1. 密度モデル

密度が大きいアクセスにおいて、各ファイルの関連度は小さいと考えた。そこで、各アクセスにおける重みの和が等しくなるとし、それを順序に応じて配分するモデルを考案した。このモデルを順序・密度モデルと呼ぶ。あるファイル A のアクセスにおいて、関連度は、密度 n とファイル B の順位 r によって定義される (式 4)。

$$C_{do}(A, B) = \frac{n-r+1}{\sum_{i=1}^N i} \quad (4)$$

期間 T におけるファイルの全アクセスについて、 C_{do} を計算し、その平均をファイル A と B の関連度とする (式 5)。

$$C_{Do}(A, B) = \frac{\sum_T C_{do}(A, B)}{\text{ファイル}A\text{のアクセス数}} \quad (5)$$

さらに、 C_t と C_{do} の積から、時刻を加味した指標が得られる。これの平均を取ることで、時刻、共起、密度を考慮した関連度モデルが得られる (式 6)。

$$C_{DOR}(A, B) = \frac{\sum_T C_{do}(A, B) \times C_t(A, B)}{\text{ファイル}A\text{のアクセス数}} \quad (6)$$

これを複合モデルと呼ぶ。

7. 実験

6 節で説明した関連度モデルのそれぞれについて、関連度抽出実験を行った。7.1 節で実験の概略について説明し、7.2 節で実験結果について考察する。

7.1. 実験概略

あるファイルを入力として与え、それに対し、その他のファイルの関連度を計算する。出力は、それらのファイルの関連度順のランキングとした。

C_t を計算するために必要なパラメータ t_1, t_2 はそれぞれ 600 秒、300 秒とした。これは予備実験の結果に基づき決定した値である。

関連度を計算するための情報として、2005 年 1 月から 2006 年 4 月にかけて蓄積したアクセス履歴を用いた。アクセス数は 10811 件である。

評価指標は、 m ($= 1, 3, 5$) 個目の正解が見つ

かった時の平均逆順位とした。ただし、正解が見つからない場合の逆順位は 0 とする。 m は利用者が求める情報が分散しているファイルの最大数を表現している。平均逆順位の満点は式 7 によって与えられる。 m によって満点が異なることに注意されたい (式 7)。

$$r(m) = \frac{1}{m} \sum_{i=1}^m \frac{1}{i} \quad (7)$$

各出力ファイルが正解かどうかの判定は利用したアクセス履歴を生成した利用者が行った。判定基準は参照の有無である。すなわち、出力ファイルが、入力ファイルと参照、非参照の関係にあるとき正解とした。

7.2. 実験結果

図 2, 3, 4, いずれにおいても、一般に順位モデルが最も高い平均逆順位を持つ。時刻モデルが最も低い平均逆順位を持ち、密度を考慮した二つのモデルはほぼ同じ平均逆順位を持つことを示している。こうした傾向は、全体の平均についても現れる (表 1)。

続いて、各 m の値について、問合せ結果を比較する。問合せ 4, 6 においては、いずれのモデルも高い平均逆順位を持つ。さらに詳しくログを調査したところ、これらのファイルは、アクセスされている時期が限られており、さらにその時期は関連のない他のファイルへアクセスが少なかった。つまり、本モデルにとって理想的なアクセス履歴が残っていた。また、問合せ 1, 9 においては、 m の増加に伴って、順序モデルと他のモデルの差が小さくなっている。この特徴は、順序モデルでは発見できなかった正解を、他のモデルが発見したことが原因である。ただし、これらの正解は 10 位以降の低いランクで発見された。

実験結果から得られた知見についてまとめる。正解数が小さい場合、順序モデルが優れた結果を返す。しかし、正解数が多い場合は、順序モデルでそれらの正解を発見することは困難である。時刻モデルを用いれば、多数の正解を発見することができるが、出力は上位に不正解を多く含む。密度を考慮することで、逆数順位を向上させることができるが、順序モデルを越える性能は得られない。

表 1 平均逆順位

	順序	時刻	順・密	複合
$m=1$	0.79	0.54	0.64	0.66
$m=3$	0.39(0.64)	0.26(0.43)	0.31(0.51)	0.31(0.51)
$m=5$	0.25(0.54)	0.17(0.37)	0.20(0.44)	0.20(0.45)

() 内は満点を 1.0 に正規化した値

平均逆順位の観点から比較する限り、順序モデルが最も優れている。しかし、 $m=5$ における正解数の平均で比較すると、順序モデルは他のモデルに劣る (表 2)。ここで、パラメータ t_2 が同じならば、他の三つのモデルは、いずれも同じ正解数を持つので区別され

ないことに注意されたい。この実験結果は、6.1節で述べた、順序モデルは4ファイル以上の関連を発見しにくいという性質と矛盾しない。この傾向に関わらず、時刻モデルと密度を考慮したモデルの平均逆順序が、順序モデルより小さい原因は二つある。一つは、上位に不正解ファイルが出力されている点である。もう一つは、3つ目以降の正解が10位以降と低い順位に出現している点である。

表 2 m = 5 における正解数の平均

	正解数の平均
順序モデル	2.4
その他のモデル	3.1

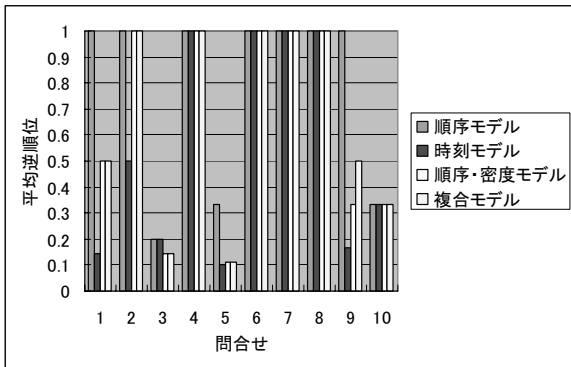


図 2 m=1 における平均逆順位

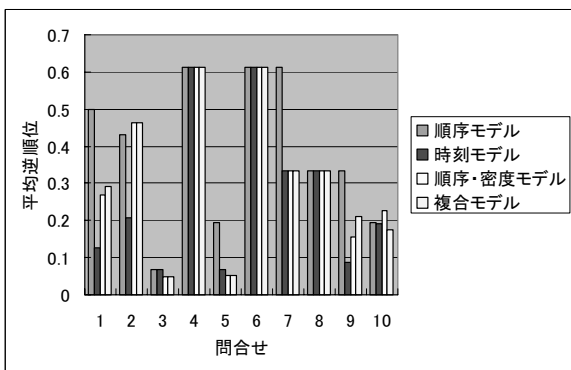


図 3 m=3 における平均逆順位

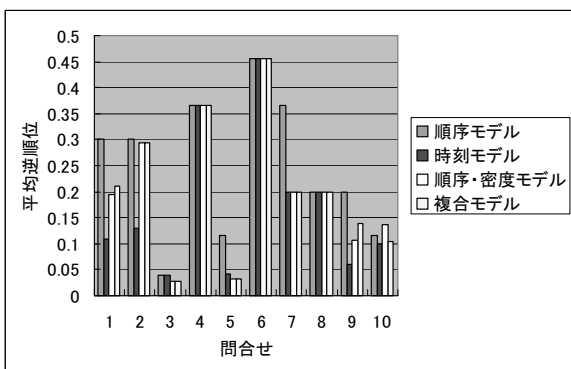


図 4 m=5 における平均逆順位

8. 今後の課題

実験では、入力ファイルに対する出力ファイルの関

連度のみを評価しており、出力ファイルに対する入力ファイルの関連度（逆関連度）を評価していない。そのため、頻繁にアクセスされる汎用的なファイルが、出力結果に現れている可能性がある。こうしたファイルはアクセス回数が多いため上位に出現するが、特定のファイルと強い関連を持つわけではない。逆関連度を評価することで、こうしたファイルをファルタリングしたいと考える。

次に、出力ファイルどうしの関連度を求めることで、結果をクラスタリングしたい。現在はスコア順にランキングしているため、互いに関連のあるファイルが、離れて表示されている場合がある。それらのファイルに利用者が求める情報が分散している場合、利用者は上位だけを見て十分な情報を得られないかもしれない。出力結果をクラスタリングすることで、利用者は分散した情報を容易に集約することが可能になる。

9. まとめ

本研究では、デスクトップ上のファイル集合モデルを提案した。そのモデルにおいて、各ファイルは他のファイルとの関連を持つ。その関連の強さを、利用者のファイル・アクセス履歴から得られる情報を用いて、計算するための4つのモデルを示した。そして、それらの手法を用いた抽出関連度実験を行い、提案したそれぞれのモデルが、どの程度実際のファイル集合に適合するかについて検討した。

文 献

- [1] 酒井哲也, “よりよい検索システム実現のために,” IPSJ Magazine, vol.47, no.2, pp.147-158, Feb.2006.
- [2] Google Desktop Search, <http://desktop.google.com/>
- [3] Windows Desktop Search, <http://www.microsoft.com/windows/desktopsearch/>
- [4] Spotlight, <http://www.apple.com/macosx/features/spotlight/>
- [5] Office OneNote, <http://office.microsoft.com/onenote/>
- [6] Google Notebook, <http://www.google.com/notebook/>
- [7] Paul-Alexandru Chirita, Rita Gavriloaie, Stefania Ghita, Wolfgang Nejdl, and Raluca Paiu, "Activity Based Metadata for Semantic Desktop Search," Proc. European Semantic Web Conference, pp. 439-454, Crete, Greece, Spring 2005.
- [8] Google Personalized Search, <http://www.google.com/psearch>
- [9] Larry Page, Sergey Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report, Stanford University, January, 1998.
- [10] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," JACM, vol.46, no.5, pp.604-632, September 1999.