

可変なカテゴリ構造を用いた WWW 検索支援方法

3 L-8

仲川こころ 高田喜朗 関 浩之

奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

大量の情報が散在する現在の WWW では、必要な情報を的確に探し出すことが難しい。情報検索を支援するために、近年様々な WWW 文書の検索サービスが公開され、広く利用されている。しかしいずれのサービスを用いても、必要な情報が簡単に見つからないことがある。

そこで本研究では、検索ごとに適切なカテゴリ構造を提供することで、WWW 検索を支援する一手法を提案する。

1.1 現在の WWW 検索サービス

現在の検索サービスは大きく、キーワード検索サービスとディレクトリサービスの2種類に区分できる。

キーワード検索サービスは、ユーザにキーワードまたはキーワードの組合せ（検索式）を入力してもらい、入力語を多く含む文書を出力するものである。一方ディレクトリサービスは、データベース中の文書を分類したカテゴリ階層構造を提供する（図1(a)）。

現状では両手法とも、必ずしも精度のよい検索を容易に行えるとは限らない。原因として以下の問題点が考えられる。

- (1) 適切なキーワードを考える事は難しい。データベースから、不必要な出力に埋もれることなく、ちょうど正解だけを抽出するような入力語を考えることは、一般に難しい。
- (2) ディレクトリサービスの提供するカテゴリ構造は、巨大で見通しが悪い。常に構造の一部しか見えていないユーザは、カテゴリの選択に迷いや誤解を生じやすい。
- (3) 分類のしかたは様々あり、適切なカテゴリ構造は検索ごとに異なる。提供される構造は常に固定されているが、ユーザが期待する構造がそれとは異なる場合もある。例えば「湯豆腐」というカテゴリに対して、図1(a)のように配置したい場合や、(b)のように配置したい場合など、様々考えられる。固定的な構造ではすべての要求に対処できない。

ディレクトリサービスは(1)の問題を軽減しているが、(2)、(3)のようなカテゴリ構造に関する問題を持っている。

1.2 目的

上記の考察に基づき、本研究では、ユーザの検索目的に合わせたカテゴリ構造を提供する方法を考える。

WWW Retrieval Method Using Flexible Category Structure.
Kokoro Nakagawa, Yoshiaki Takata, Hiroyuki Seki
Graduate School of Information Science, Nara Institute of Science and Technology

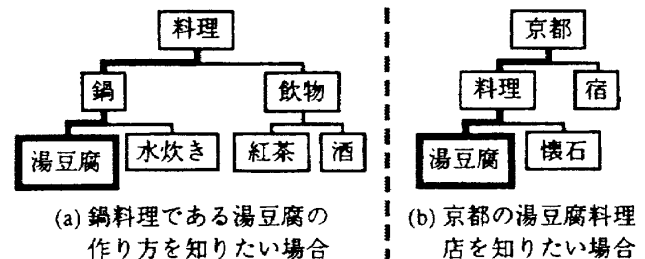


図 1: カテゴリ階層構造の例

検索のたびに、それに応じて適切なカテゴリ構造を構築すれば、上の(2)、(3)で述べたようなディレクトリサービスの問題を解決できる。さらに、例えば図1(a)の構造では「湯豆腐」の周辺に「水炊き」や「飲物」、(b)の構造では「温泉」や「宿」というように、目的のカテゴリの周辺にもユーザの興味に近いものが配置されることが考えられる。

またユーザに要求する操作に関しては、簡単なキーワード入力とカテゴリの選択程度に抑え、現状に比べ負担をあまり増やさずに実現したい。以上の方針に沿って設計した検索支援システムについて、以下説明する。

2 システム設計の方針

2.1 基本方針

提案システムは、以下の2段階の処理を行う。

- [i] ユーザの興味を反映する文書の集合を大まかに決める。
- [ii] その文書集合に依存してカテゴリ構造を作る。

[i] の段階で対象とする文書集合を限定し、その文書集合に依存したカテゴリ構造を作ることで、検索ごとの目的に合わせたカテゴリ構造を構築する。同時に、得られるカテゴリ構造をユーザの興味の範囲に限定し、巨大にならないようにする。[ii] のカテゴリ構造の構築には、設計者がシステムに予め与える分類観点という予備知識を用いる。

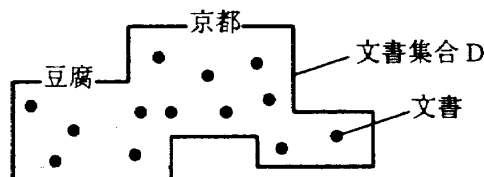
分類観点 分類観点とは、ディレクトリサービスで言うと、カテゴリ階層構造に現れる兄弟ノード（兄弟カテゴリ）の集合である。例えば「京都」「大阪」「奈良」などのカテゴリからなる「地域」という分類観点などが考えられる。

2.2 動作の概要

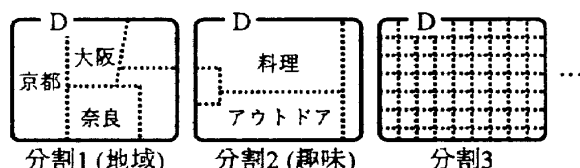
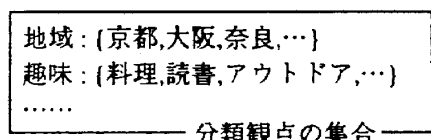
システムの動作手順は、以下のようになる（図2）。

- (1) ユーザが複数のキーワードを入力する。

- (1) キーワードを入力 (e.g., 「豆腐, 京都, 店, ...」)
- (2) 関連文書のある程度広く収集



- (3) 分類観点に沿って D を分割



- (4) 評価値の高い分類観点を提示
- (5) ユーザーが適するカテゴリを選択 (e.g., 「京都」)
- (6) 選択されたカテゴリに対応する文書集合について, (3)~(4)を繰り返す

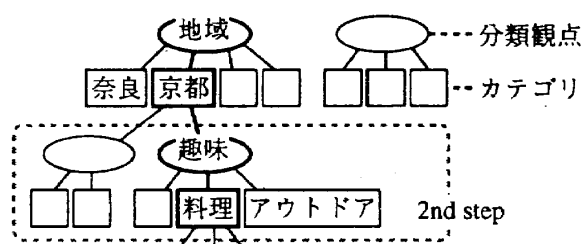


図 2: システムの動作

- (2) 文書データベースから, 入力キーワードに関連する文書を求める. この文書集合を D とする.
- (3) システムに用意した分類観点に沿って D を分割する. 同時に, カテゴリへの分類の明確さや細かさに基づいて, D に対する分類観点の評価を行う.
- (4) 上記の評価値の高い (複数の) 分類観点と, それぞれの要素であるカテゴリをユーザに提示する.
- (5) ユーザが, 提示されたカテゴリの中から適するものを選ぶ. 同時に, そのカテゴリに分類された文書集合 D' を検索結果として検索を終了するか, さらに D' の分割を行うか選ぶ.
- (6) ユーザが D' の分割を希望したときは, D' について (3) 以下を繰り返す.

以下, 本手法の各要素について説明する. 詳細については [1] を参照.

文書の収集 入力された複数キーワードのいずれか一つと関連性の高い文書を抽出する (OR 検索). ここでの目的

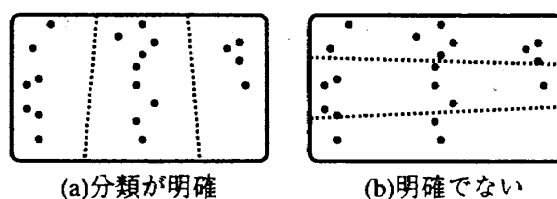


図 3: 分類観点の評価

は, D を, 後にカテゴリ構造に配置する対象として広く求める事である. 従って, 精度の良い検索をめざして慎重にキーワードを考える必要はない.

分類観点に沿った分割 上で求めた文書集合を, 用意した分類観点に沿って分割する. 直感的に, D の分割とはカテゴリ構造の 1 階層を構築することに相当する. 分類観点 S に沿った文書集合 D の分割とは, D の各要素について, それが属するカテゴリを S の要素から一つ選ぶことである. これにより, D は S の各カテゴリに属する部分集合に分割される.

文書が属するカテゴリは, S の要素のうち, その文書との関連度が最大のものとする. 本研究ではベクトル空間モデル [2] を用い, 文書とカテゴリの関連度を, 双方の特徴ベクトルのなす角に依存して定義する.

分類観点の評価 分割を行った後, (1) カテゴリへの分類が明確であること, かつ (2) 分割が細かすぎないこと, の 2 点から分類観点の適切さを評価する. 分類が明確であるとは, 文書 d が属すべきカテゴリ w_d が明確である, すなわち, d と w_d との関連度が d と他のカテゴリとの関連度よりも十分高いことである (図 3(a)). 図 3(b) のような場合は, その分割が, 文書の特徴を分類するものとは言えないことを示唆している.

また, 分割が細かすぎると, 各文書を個別に扱う状態に近づき, 分類という意味が失われる. そこでここでは, 分割が細かすぎる分類観点を除いた上で, 分類が明確なものを評価値の高い分類観点とする.

3 おわりに

可変なカテゴリ構造の提供による WWW 検索支援について, その目的と設計案を述べた. 現在, 上記の設計を基に実験システムを試作中である. 本手法の有効性の評価, 既存の WWW 検索手法との比較実験等を計画している.

参考文献

- [1] 仲川, 高田, 関: 可変なカテゴリ構造を用いた WWW 検索支援方法の提案, 電子情報通信学会第 9 回データ工学ワークショップ (DEWS'98), 講演番号 22, 1998-3.
- [2] Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition Using Text Segments and Text Themes, *Hypertext '96 Proc.*, pp.53-65, 1996.