

# Gooots-検索目的に沿ってカテゴリ名を取捨選択して クラスタリングする検索エンジン

金子 大輔†, 高山 毅†, 池田 哲夫†, 長内 亘†

岩手県立大学ソフトウェア情報学部†

## 1. はじめに

ロボット型検索エンジンの検索結果を検索者が理解しやすいようクラスタリングして表示する研究が進行している。既定のカテゴリ階層の一部を抜き取り、カテゴリ名として採用する「静的なクラスタリング」[1]や、検索結果と検索のキーワードから抽出したカテゴリ名を動的に付与する「動的なクラスタリング」[2]などである。しかし、カテゴリの分け方に対する検索者の満足度は充分とはいえない。そこで本研究では、検索のキーワードを問わず出現率が高いカテゴリの五つのうち、検索者が取捨選択し指定したのみをカテゴリとして採用する検索エンジンを提案する。本提案により検索者は、慣れ親しんだカテゴリのうち選択したカテゴリを用いたクラスタリングのもとで、必要な情報をより効率的に獲得することが可能になる。

## 2. 先行研究

### 2.1 検索結果のクラスタリングにおける問題点

静的なクラスタリングの欠点は既定のカテゴリ階層から抜き取るため、カテゴリ分けにおいて一貫性にかけることである。また、動的なクラスタリングの欠点は、有用ではないタームをカテゴリとして採用してしまう危険性があることである。

### 2.2 検索エンジンが満たすべき要件

文献[3][4]によれば要件として以下二点が重要である。

- 要件 1: 検索者が検索エンジンに対して与える平均で 2 つ程度の検索のキーワードから、検索目的に沿った的確な検索結果を返すこと[3]。
- 要件 2: 可能な限り短時間で的確な検索結果を返すこと[4]。

## 3. 提案手法

### 3.1 検索エンジンの使用目的

検索エンジンの使用目的に関する、信頼すべき調査結果は見当たらない。そこで、より広い範疇である「インターネットの利用目的」(図 1)から、本研究では検索エンジンの使用目的として以下の 2 つに注目する。

[検索目的 1] : 仕事や学習のための検索

[検索目的 2] : 商品購入のための検索

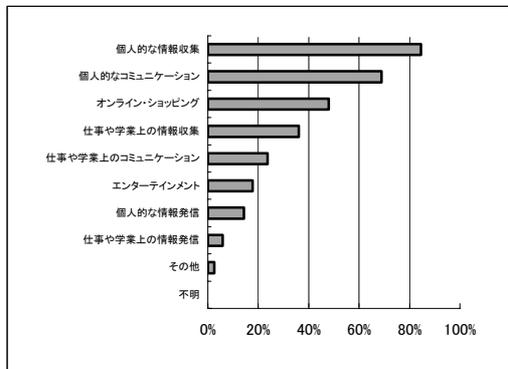


図 1 インターネットの利用目的[5].

Gooots- Search Engine Which Enables Clustering with Selected Category Name According to Retrieval Purpose  
D.Kaneko, T.Takayama, T.Ikeda, and W.Osanai  
Faculty of Software and Information Science, Iwate Prefectural University, JAPAN

### 3.2 検索のキーワードを問わず出現率の高いカテゴリ

検索のキーワードを問わず出現率の高いカテゴリを探す。表 1 に示すキーワードで google[6]で実際に検索を行い、検索結果上位 100 件を手動で分類した。なお、紙幅の都合により、分類結果の一部を表 2 に示す。

表 1 検索に用いたキーワード

1	モンテカルロ法	5	過去完了形	9	ソファ
2	江沢民	6	食物連鎖	10	ディスプレイ
3	量子力学	7	ヘッドフォン	11	自転車
4	太宰治	8	ラーメン	12	インテリア

表 2 表 1 での各検索結果の分類の抜粋

検索キーワード	カテゴリ	出現率
モンテカルロ法	特定のアプリケーションを必要とするページ(pdf,doc,ppt,rtf,ps,xls)	9%
	ショッピングサイト	6%
	書籍	8%
江沢民	特定のアプリケーションを必要とするページ(pdf,doc,ppt,rtf,ps,xls)	4%
	書籍	4%
	掲示板,日記	5%
量子力学	特定のアプリケーションを必要とするページ(pdf,doc,ppt,rtf,ps,xls)	4%
	ショッピングサイト	6%
	書籍	16%
	シラバス	11%
ヘッドフォン	ショッピングサイト	35%
ラーメン	掲示板,日記	3%
	ショッピングサイト	14%

以上の結果から、以下の五つのカテゴリは検索のキーワードを問わず頻出するカテゴリとわかった。

1. 特定のアプリケーションを必要とするページ (pdf, doc, ppt, rtf, ps, xls)
2. ショッピングサイト
3. 書籍
4. 掲示板, 日記
5. シラバス

本研究ではこれ以降、これら 5 つのカテゴリのことを、「基本カテゴリ」と呼ぶことにする。

### 3.3 カテゴリの取捨選択

#### 3.3.1 検索者がカテゴリを選択可能な Gooots 通常検索

表 1, 2 の調査結果として、[検索目的 1]のとき、検索結果に'シラバス'などは出現する確率が高いが、[検索目的 2]の結果にはあまり出現しない。このようにあまり出現しないカテゴリ名まで採用してクラスタリングすることは、効率的とはいえない。また、カテゴリの採否に関しては検索者側にも希望の個人差があると考えられる。そこで本研究では検索のキーワードに加え、検索者に基本カテゴリの中からどれをカテゴリとして採用するかを指定させる。これを、「Gooots 通常検索」と呼び試作システムを作成した。

Gooots 通常検索は、選択されたカテゴリに属する Web ページのみをカテゴリでまとめる。どのカテゴリにも登録されなかった Web ページはリスト表示する。

### 3.3.2 検索目的に沿ってカテゴリが自動選択される Gooots 目的別検索

基本カテゴリの中で[検索目的 1, 2]のそれぞれに対して、選択すべきカテゴリをアンケート調査した(表 3)。これに基づき、検索目的を選択すると、自動的にカテゴリが選択される「Gooots 目的別検索」を作成した。

表 3 カテゴリと検索目的の対応関係に関するアンケート結果 (岩手県立大学の学生 126 名が有効回答)

選択すべきカテゴリ\検索目的	勉強や仕事目的	買い物目的
特定のアプリケーションを必要とするページ	67.46%	21.43%
掲示板や日記	35.71%	26.19%
書籍の紹介	65.08%	63.49%
ショッピングサイト	30.95%	22.22%
シラバス	75.40%	75.40%

### 3.4 クラスタリングの実現方法

まず、各カテゴリの特徴語は経験則により設定する。そして、googleAPI[9]を用いて検索を行い、設定した特徴語を含む Web ページをそのカテゴリに登録する。

## 4. 評価

### 4.1 被験者による主観的評価

[検索目的 1, 2]に対して、以下のような問題を設定した。

- [検索目的 1]の例として、大学入試および Web ラーニングの問題 4 問
- [検索目的 2]の例として、指定した商品を検索する問題 4 問

設定した問題を以下の四つの検索エンジンを使用し回答してもらった。

1. vivisimo[7](動的なクラスタリングの例)
2. yahoo[8](静的なクラスタリングの例)
3. Gooots 通常検索
4. Gooots 目的別検索

以下の四つの尺度で 7 段階(0:とても悪い~7:とても良い)の主観的評価を行った(図 2, 3)。

- 尺度 1: カテゴリの妥当性
- 尺度 2: 検索結果から得られた情報の満足度
- 尺度 3: 検索目的とは関係の無いページがどのくらい目に付いたか
- 尺度 4: 目的の情報の見つけやすさ

[検索目的 1]では四つの尺度全てにおいて'vivisimo', 'yahoo'より Gooots2 つが高い評価を得た。[検索目的 2]においても、微差ではあるが、Gooots2 つに優位性が見られる。

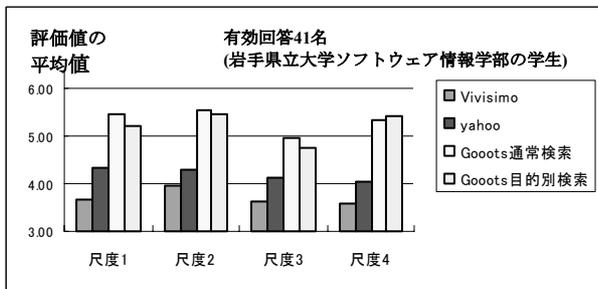


図 2 [検索目的 1]での評価結果。

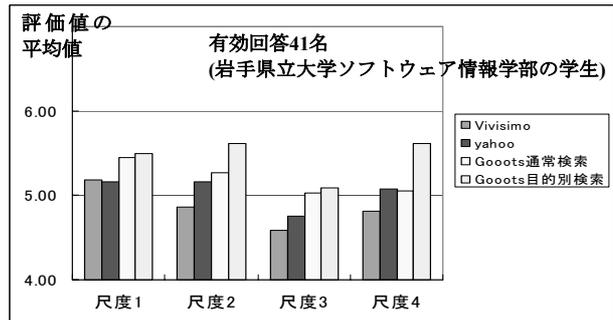


図 3 [検索目的 2]での評価結果。

### 4.2 客観的評価

また、問題を回答する際に参考にした Web ページが検索結果の上位何位のものであったかを、四つの検索エンジンで調査した。上位 60 位までに適切な Web ページが出現しなかった場合は 60 位での出現としている。結果、Gooots2 つが'vivisimo', 'yahoo'より良い結果が得られた。

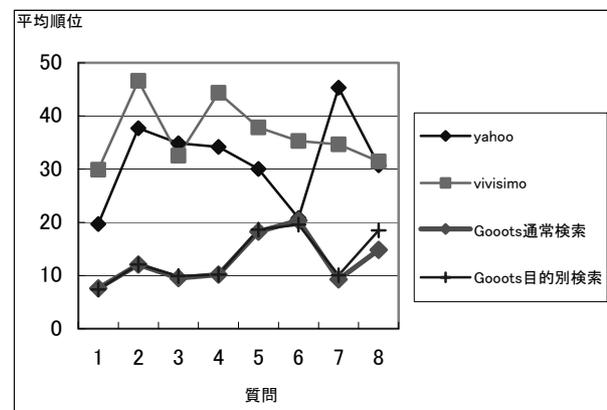


図 4 参考にした Web ページの平均順位。

## 5. 結論と今後の展望

本研究では検索のキーワードを問わず出現率が高い五つのカテゴリから、検索者の選択したカテゴリのみを用いてクラスタリングする検索エンジン Gooots を作成した。評価実験の結果、仕事や学習のための検索において、Gooots 通常検索、Gooots 目的別検索の双方で同様に有効性を確認できた。今後、商品購入のための検索など、多目的で有効性の向上を図る予定である。

### 参考文献

- [1] 安形輝, 石田栄美, 久野高志, 野末道子, 上田修一: 「WWW ページの自動分類:NDC の分類体系と Yahoo のカテゴリを使った分類」, 情処研報 FI-54, 1999.
- [2] 成田宏和, 太田学, 片山薫, 石川博: 「階層的クラスタリングを利用したメタ検索エンジンの提案」, 情処研報, DBS-128-50, 2002.
- [3] 江口浩二: Web 検索の技術動向と評価手法: 情報処理, Vol. 45, No. 6, pp. 569-573, 2004.
- [4] 岸田和明, 賀沢秀人: 探しもの見つけます~情報化社会に役立つ情報検索の技術動向~, 情報処理, Vol. 44, No. 6, pp. 615-621, 2003.
- [5] 情報通信総合研究所, [http://www.commerce.or.jp/result/min30/2\\_1.html](http://www.commerce.or.jp/result/min30/2_1.html)
- [6] google, <http://www.google.com/>
- [7] vivisimo, <http://vivisimo.com/>
- [8] yahoo, <http://www.yahoo.co.jp/>
- [9] googleAPI, <http://www.google.com/apis/index.html>