

カテゴリ型検索エンジンの分類体系を用いた概念的情報検索

太田黒 啓 佐藤 慶三 中島 誠 伊藤 哲郎

大分大学工学部

1. はじめに

インターネットの急速な普及と、それに伴うネットワーク技術の発展により、ユーザはウェブや電子図書館などを通じて、様々な文献情報を容易に手に入れることができるようになった。問題は、多くの不必要な情報が取り出されてしまうことである。これに関して、質問との概念的関連性を利用し、ユーザが欲する情報を効率よく取り出す手法が提案されている[3][4]。

概念的関連性を利用した情報検索では、語の概念的上位下位関係が記述された、概念辞書が用いられる。しかしながら既存の概念辞書の多くは、人手により作成されており、頻繁に更新することが難しく、時事情報などに上手く対応できない。また、専門分野に特化しているものが多く、広い分野に対応可能な辞書は少ない。

これらの問題に対処するため、ここではカテゴリ型検索エンジンの分類体系を概念辞書として利用することを考える。カテゴリ型検索エンジンでは、様々なサイトが、それぞれの内容に合ったカテゴリに分類されており、カテゴリに含まれる情報が頻繁に更新されている。また、多岐にわたる分野のカテゴリが存在しており、様々な検索要求に対応できる。以下では、カテゴリ型検索エンジンの一つである Yahoo!Japan[6] の分類体系を概念辞書として利用する方法を述べ、その有効性を示す。

2. 概念的情報検索の概要

概念辞書中の一つの概念は、記述子（概念を示すための語句あるいは記号）によって表されている。多くの場合、記述子は概念の概念体系上の位置を示す表記（今後、識別子と呼ぶ）と、その概念を表す語句（今後、概念見出し語と呼ぶ）の二つからなっている。

概念辞書を用い、質問と文献の概念的関連性をどのように捉えるかをまとめておく。質問

（あるいは文献）は重み付きのキーワードの羅列として与えられるとする。

質問と文献の表現は、これらキーワードを概念見出し語に部分文字列として含むような概念辞書中の記述子に置き換え、置き換えられた記述子を集めて得る。質問と文献の概念的関連性は、表現中に概念辞書中で、より近い位置にある記述子を多く含んでいる文献ほど、質問との関連性が高いとする。これにより、キーワードの文字面では捉えることの出来ない関連性を、概念的に捉えることができ、より優れた検索効率が期待できる。

3. Yahoo!Japan の利用

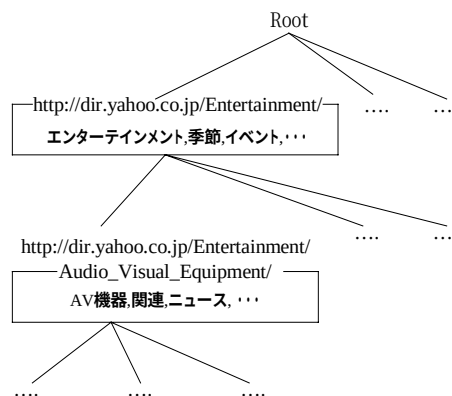


図1 Yahoo!Japan を用いた概念辞書

Yahoo!Japan の分類体系を概念辞書として利用する際には、一つのカテゴリを一つの概念とし、Yahoo!Japan のディレクトリ階層を概念の上位下位関係として捉える。その上で、各概念を表す記述子を抽出して充てる。具体的には、識別子として各カテゴリの URL を充て、概念見出し語として、各カテゴリにつけられたカテゴリ名と、カテゴリの各サイトに付加されている説明文から形態素解析によって抽出した名詞句を充てる。

一つ概念につき、複数の見出し語を用意することで、様々なキーワードを置き換えられる可能性が高くなる。上のようにして得られる概念辞書（一部）を図1に示す。深い階層の概念であるほど、下位であることを示す。

質問と文献の表現を得るには、与えられたキ

Conceptual information retrieval using a classification scheme of a category-based search engine

Akira Otaguro, Keizo Sato, Makoto Nakashima and Tetsuro Ito
Faculty of Engineering, Oita University

ワードを記述子に置き換える。キーワードと概念見出し語の類似度を式 $(m_1+2m_2)/(L-2(L-1))$ [1]で求め、値が閾値以上であるときに置き換えを行う。ここで、 m_1 はキーワードと概念見出し語の間に共通する文字の数、 m_2 は隣接する二文字を一組としたときに共通する組の数、 L はキーワードと概念見出し語の文字列長で長い方の長さである。閾値は、現段階では 0.7 としている。この方法を用いた置き換えを行うことによって、例えば「エンターテインメント」と「エンターテイメント」のような、意味が同じで、表記に多少違いがある語句同士の場合でも、置き換えができるようになる。

図 2 に置き換えた記述子を集めた表現の例を示す。「識別子:(キーワード)概念見出し語:重み」という形式で表記されている。記述子の重みは、キーワードの重みを置き換わった記述子の数で等分してある。

http://dir.yahoo.co.jp/Business_and_Economy/Business_to_Business/Travel_and_Transportation/Transportation/Aircraft/Helicopters/:(阿蘇山)阿蘇山:0.071428575

... ..

http://dir.yahoo.co.jp/Business_and_Economy/Business_to_Business/Energy/Petroleum/Idemitsu_Kosan/:(ドライブコース)ドライブコース:0.0625

... ..

http://dir.yahoo.co.jp/Recreation/Automotive/Makes_and_Models/Honda/Models/S_Series/S800/:(ドライブコース)推薦ドライブコース:0.0625

図 2 置き換えを行った例（キーワード：「阿蘇山，ドライブコース」）

4. 実験

被験者 7 人にそれぞれ質問を設定してもらい、9 件の質問を収集した。また各質問について、ロボット型検索エンジン Excite[2]の検索結果上位 100 件の文献について、質問を設定した被験者に、適合、不適合の判定を行ってもらった。この際、文献のキーワードとして、名詞句の生起頻度の重み付けによる上位 20 個を利用した。結果を再現率-適合率の形で図 3 に示す。比較対象として、質問と文献のキーワードをそのまま用いたコサイン関数、質問と文献のキーワードを記述子に置き換えた後に識別子をキーワードとして用いたコサイン関数、そして、概念見出し語としてカテゴリ名のみを利用した場合の結果も示してある。

提案した手法による検索効率が他の方法より

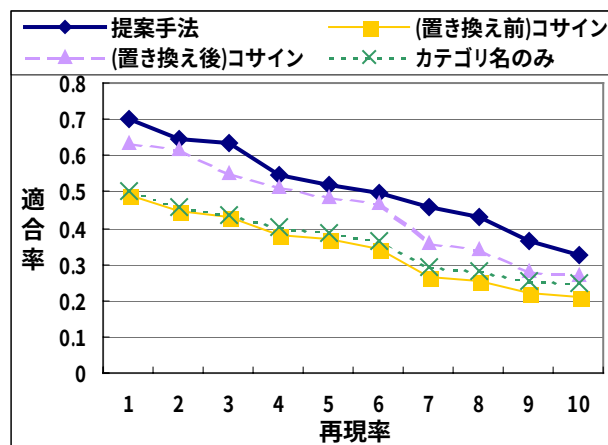


図 3 再現率-適合率グラフ

も優れていた。提案手法が、コサイン関数では捉えきれてない概念的関連性を捉えていること、Yahoo!Japan から用いた概念辞書の上位下位関係の利用が有効であることがわかる。また、カテゴリ名だけでなく、カテゴリに含まれるサイトの説明文から抽出した複数の見出し語を用いることが有効であるといえる。

5. まとめ

カテゴリ型検索エンジンの分類体系を概念辞書として利用し、質問と文献間の概念的関連性を捉える手法を述べ、その有効性を示した。

今後は、ロボット型検索エンジンの検索結果のブラウジング支援に提案手法を応用させていく。その際、[5]で提案されているようなユーザの適合判断を取り入れた手法も検討していく。

参考文献

- [1] 安部 隆之, 佐藤 浩史, 重松 修一, 中島 誠, 伊藤 哲郎: 構造マッチングによる文献の知的検索と結果の色空間表示, 情報処理学会研究報告, 人文科学とコンピュータ, vol.29-5, pp.25-30, Jan. 1996.
- [2] Excite: <http://www.excite.co.jp/>
- [3] 堀井 千夏, 今井 正和, 千原 國宏: デジタル図書館のための概念情報を用いた科学技術論文の検索, 電子情報通信学会論文誌 D-I, vol.J82-D-I, no.10, pp.1245-1255, Oct 1999.
- [4] Makoto Nakashima, Keizo sato, Yanhua Qu, Tetsuro Ito: Browsing-Based Conceptual Information Retrieval Incorporating Dictionary Term Relations, and a Use's Interest, JASIST, vol.54, no.1, pp.16-28, Jan, 2003.
- [5] 曲 艶華, 佐藤 慶三, 中島 誠, 伊藤 哲郎: 電子図書館のための適合可能性示唆によるブラウジング支援, 電子情報通信学会論文誌 D-I, vol.J84-D-I, no.7, pp.1009-1020, July 2001.
- [6] Yahoo! Japan: <http://www.yahoo.co.jp/>