2ZE-8

地域ポータルサイトにおける概念体系間の関係と文書分類技術を用いたカテゴリ分類モデルの構築

小林　拓也† 　竹野　健夫† 　岡本　東† 　堀川　三好† 　菅原　光政†

†岩手県立大学ソフトウェア情報学部

1. はじめに

現在，Web 上に存在する大量の情報の中から利用者の検索目的に応じた情報を提供するため Google などの検索エンジンや，Yahoo などのポータルサイトが利用されている．一般的に検索にはカテゴリ検索やキーワード検索が用いられる．人手による分類が行われるカテゴリ検索では，情報の増加，内容の変化によって分類手法やカテゴリの概念体系の変更が重要となる．

これらの問題点について文書分類を用いたカテゴライズや，クラスタリングによる分類などの先行研究が多くなされている [1) 2)]．しかし，自動的に分類を行う際，複数のカテゴリに分類されることで起きる横断性や，属した先のカテゴリが適切であるかという妥当性の問題は残っている．

本研究では，文書分類に基づいた，階層的なカテゴリ構造を示す概念体系の自動生成・更新を行う手法を提案する．また，提案手法に基づいた機能を持った地域ポータルサイトのプロトタイプを構築する．

2. 分類手法の提案

変化する情報の分類を行う為，収集する情報の対象は，特定の Web サイト（以下，ターゲットサイト）から配信されている RSS 中の文書ベース情報（以下，コンテンツ）とする．

本研究では概念体系をある一つの上位カテゴリと n 個の下位カテゴリからなるカテゴリ間の関係とし階層は 2 階層に限定する．

2.1 カテゴリ候補の決定

一般的な社会の事象を捉える事ができるカテゴリ名を m 個（$n \leq m$）のカテゴリ候補とする．

2.2 概念体系を用いた分類

上位下位カテゴリの繋がりを考慮した分類と概念体系の生成を行う．

(2-1) ターゲットサイトのカテゴリを決め，それを上位カテゴリとする．

(2-2) 各コンテンツについて文書分類を行い，所属する数の上限と閾値を設定し，それに基準にカテゴリ候補を所属させる．

文書分類は TFIDF 法とベクトル空間モデルを用いる [1)]．

(2-3) 閾値を設定し，それを基準に(2-2)のカテゴリ候補から下位カテゴリを n 個決定する．

概念体系の妥当性を精度によって評価する．

正解率=正解数/属したカテゴリ数
未分類率=未分類のコンテンツ数/全体のコンテンツ数
精度=2/(1/正解率+1/(1-未分類率))

2.3 概念体系生成による分類精度の向上

2.2 節の(2-3)について，極端に所属が少ないカテゴリはノイズとして排除することで属するカテゴリの分類精度の向上を図る．横断性については再現率，妥当性については適合率で評価し，分類精度を F 値で評価する．

適合率=検索された中の正解数/検索数
再現率=検索された中の正解数/全体の正解数
F 値=2/(1/適合率+1/再現率)

3. 評価実験

3.1 評価の対象

比較対象として人手で作成した正解を用意した． ターゲットサイトを岩手県私立幼稚園ポータルサイトと設定し，コンテンツは配信された外部公開のおたより 50 件を用いた．カテゴリ候補はどんな情報が求められるかを考察し決定した．具体的には「子供の様子」→「スポーツ」「遊び」，「教育方針」→「教育」「社会」「はやりの病気」→「医療」など 40 のカテゴリ候補を用意した．各カテゴリ候補の辞書は一般的な文書から作成した．上位カテゴリは幼稚園とし下位カテゴリは属した数を基に決定し，概念体系は**表 1** の様になった．未分類は下位カテゴリのどれにも属さなかった文書の数となって

Construction of Classification System in Category Retrieval for Regional Portalsite

Takuya KOBAYASHI†, Takeo TAKENO†, Azuma OKAMOTO†, Mitsuyoshi HORIKAWA†, Mitsumasa SUGAWARA†

Faculty of Software and Information Science, Iwate Prefectural University

表 1　作成した概念体系

上位カテゴリ	下位カテゴリ	属した数	下位カテゴリ	属した数
幼稚園	教育	32	環境問題	4
	社会	15	遊び	4
	自然	9	天気	4
	食	5	暮らし	4
	イベント	5	学習	4
	医療	4	未分類	2

表 2　(2-2)の属する上限と閾値のパラメータ

比較方法	類似度の絶対値	類似度の差分	属するカテゴリの上限
閾値	0.02	0.005	3
	0.04	0.01	5
	0.06	0.015	10

表 3　(2-3)の閾値のパラメータ

比較方法	属したカテゴリ数	類似度累計の構成比
閾値	5	0.0256

る．この概念体系と提案手法で生成した概念体系を比較する．

3.2　評価の手順

2.2 節の手法に基づいて実験を行い，そのなかで精度と F 値を求め，最も値が高くなるパラメータを明らかにする．ここで各パラメータを**表 2, 3** のように用意した．

(3-1)　2.2 節の(2-3)について，概念体系の精度を求める．

(3-2)　(3-1)と同様に概念体系を作成した場合としなかった場合を比較し，作成した場合の F 値の上昇率を求める．

(3-3)　F 値の上昇率が一定以上の基準を満たすものの中で精度が最も高い各パラメータを明らかにする．

設定する各パラメータ全ての組み合わせで実験を行いその時の精度と F 値を求めた．

4.　実験結果

全てのパラメータの組み合わせを行った結果は概念体系の妥当性が平均で約 75%となった．F 値については概念体系を作成した場合，分類精度が平均で約 43%上昇したことを確認した．しかし，カテゴリ毎にバラつきが有り，F 値が 0 のものも存在した．**図 1** は F 値の上昇率が一定以上の基準の満たすもので概念体系の精度の結果の一部あり，値が最も高くなった(2-3)の閾値が類似度累計の構成比平均のグラフである．x 軸は(2-2)の閾値であり類似度の差分を基にした値，y 軸は概念体系の精度，折れ線は(2-2)の所属する上限数となっている．パラメータは上限：5，閾値：0.01 が精度 82%で最も高くなった，最も低

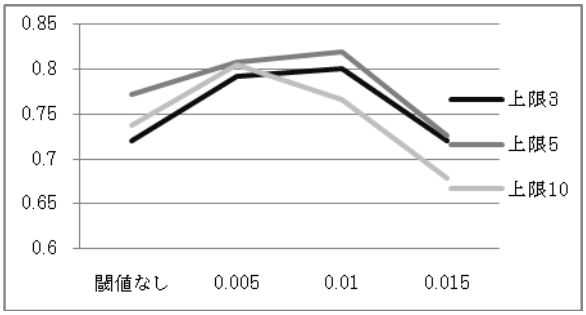

図 1　概念体系の精度，
((2-3)の閾値：類似度累計の構成比平均　0.0256
F 値の上昇率の基準：43%)

かったパラメータと約 40%の差があり，パラメータ設定によって精度に大きく影響を与えた．

5.　システム構築

実験結果からパラメータを決定し，分類手法を元にした機能を実装し地域ポータルサイトのプロトタイプの構築を行った．機能の位置づけは**図 2** のようになり，運営者によって登録された Web サイトを用いて情報の収集を行う．

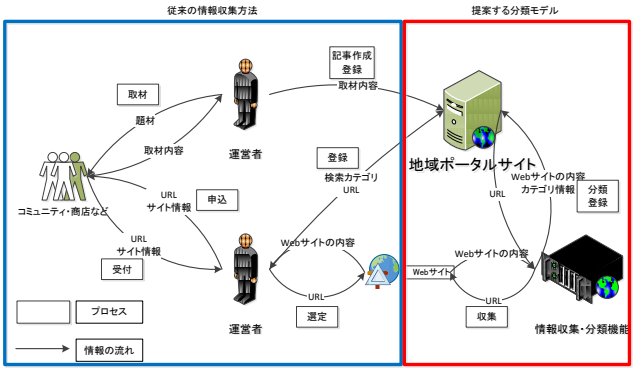

図 2　分類手法の位置づけ

6.　おわりに

本研究では，横断性や妥当性と言った問題点に対し文書分類に基づいた，階層的なカテゴリ構造を示す概念体系の自動生成・更新を行う手法の提案を行った．概念体系を生成することによる F 値の上昇，概念体系の妥当性，提案手法で用いる各パラメータを定量的な評価基準を用いて明らかにした．

今後，ターゲットサイトや上位カテゴリを増加させたときの概念体系の精度の調査を行っていく．また，地域ポータルサイトの運用を開始し，検索機能の有効性の評価を行う．

参考文献

1) 石田栄美：テキスト自動分類の概要，情報の科学と技術　56 巻 10 号(2006)

2) 中村幸雄：情報検索理論の基礎　共立出版株式会社　1998 年 9 月 1 日　初版