社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－179 (10)
2007－SLP－ 66 (10)
2007／5／25

階層構造におけるカテゴリの統合と類似文書抽出への適用

佐野 司　　市岡 健一　　福本 文代
山梨大学大学院医学工学総合研究部
E-mail: {g07mk011, g07mk001, fukumoto}@yamanashi.ac.jp

本稿では, 日英のカテゴリ階層構造に注目し, カテゴリ同士の類似性を推定し, カテゴリを統合する手法を提案する. 本手法では一方の階層構造に分類されている文書を, 他方の階層構造に分類することにより, 異なる階層間におけるカテゴリ同士の類似性を推定する. 文書を分類するための手法として機械学習 SVMs を用いた. また, 類似性を求めるために, χ^2statistics を用いた. さらに, 得られたカテゴリの組に対しアプリオリアルゴリズムを用いることで, 類似したカテゴリ集合を抽出した. 統合したカテゴリの有効性を検証するため, 類似文書の抽出を行った結果, 適合率は 0.328 であり階層構造を利用しない手法 (0.088) と比較して精度の向上がみられた.

Integrating Cross-Language Hierarchies and its Application to Relevant Document Extraction

SANO Tsukasa　　ICHIOKA Kenichi　　FUKUMOTO Fumiyo†
Interdisciplinary Graduate School of Medicine and Engineering
University of Yamanashi
E-mail: {g07mk011, g07mk001, fukumoto}@yamanashi.ac.jp

This paper presents a method for integrating cross-language category hierarchies, i.e. Reuters'96 and UDC code hierarchy of Japanese by estimating category similarities. First, we classify documents from one hierarchy into categories with another hierarchy using cross-language text classification technique, and extract category pairs of two hierarchies. Next, we apply χ^2 statistics to these pairs in order to obtain similar category pairs, and finally we apply the generating function of Apriori to the result of category pairs, and find sets of similar categories. Moreover, we examined whether or not integrating hierarchies helps to support retrieval of relevant documents. The retrieval result of 0.328 precision showed improvement over the baselines.

1 はじめに

インターネット上には膨大な数の文書が存在している．そのため文書の効率的な管理や利用が重要となる．その方法の一つに階層構造を利用した分類がある．文書を予め階層構造の各ノードに分類しておくことで，高速かつ信頼性の高い検索を行うことができる．しかし，階層構造はそれぞれによってカテゴリの個数や構造，1カテゴリあたりに分類される文書の数が異なるため，カテゴリ数が少ない場合には，ユーザにとって不必要な文書まで提示してしまうという問題が生じる．この問題を解決するための手法の一つに階層構造の統合がある．市瀬らは，Web上の情報を統合するために HICAL と呼ばれるシステムを開発した [1]．HICAL は，Yahoo! や Google など，Web上で提供されているインターネットディレクトリ (階層構造) において2つの階層構造に分類されている共有文書の数を利用し，カテゴリ間の類似性を求めることでカテゴリの統合を行うというものである．基本となるアイデアは，互いに類似しているカテゴリには，それぞれ類似した情報が分類されているということである．一方の階層構造を元ディレクトリ，他方を目標ディレクトリとし，元ディレクトリに存在しているカテゴリと目標ディレクトリのカテゴリとの類似度を求める．そして，類似度が閾値以上のカテゴリの組を見つけ出し，元ディレクトリのカテゴリを目標ディレクトリに追加することでカテゴリ同士を統合する．しかしこの手法は，共有文書を持たないカテゴリに対するカテゴリ間の類似性を求めることができないという問題を持つ．本研究ではこの問題に対処するため，文書分類に基づいてカテゴリ間の類似性を推定し，カテゴリを統合する手法を提案する．本手法では文書を分類するための手法として機械学習 SVMs[2] を用いた．また，χ^2statistics を用いてカテゴリ間の類似性を推定した．さらに得られたカテゴリの組に対し，Agrawal らによって提案されたアプリオリアルゴリズム [3] [4] を用いることで，類似したカテゴリ集合を生成した．また，統合したカテゴリの有効性を検証するため，統合結果を類似文書抽出に適用した．

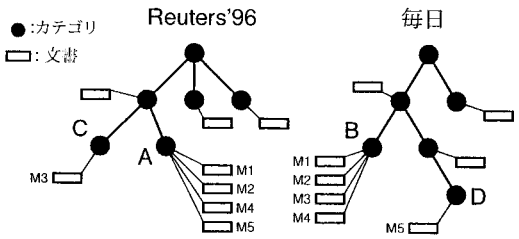

図 1: 文書分類結果の例

2 カテゴリの統合

一般に，2つの異なる階層構造の統合とは，2つの階層構造から全く新しい階層構造を1つ作り上げることを意味する．本研究では，カテゴリの統合を異なる階層構造間で類似したカテゴリ同士を1つに纏め，カテゴリ集合を生成することと定義する．本研究におけるカテゴリの統合手法は大きく分けて2つの処理からなる．処理1は機械学習 Support Vector Machines (SVMs) を用いて文書分類を行い [5], [6]，分類結果に対して χ^2statistics を適用することで，異なる階層構造間のカテゴリ同士の類似性を求める．処理2は，処理1で生成した類似カテゴリペアに対して Agrawal らによって提案されたアプリオリアルゴリズムを適用し，3カテゴリ以上の類似カテゴリを作成する処理である．

2.1 類似カテゴリペアの生成

カテゴリを統合するための1つ目の処理は，異なる階層間において類似したカテゴリの組を生成することである．本稿では，類似した2つのカテゴリを類似カテゴリペアと呼び，3つ以上のカテゴリを類似カテゴリ集合と呼ぶ．カテゴリの類似性を判定するために，一方の階層構造に分類されている文書をテストデータとし，これらを他方の階層構造に分類する．我々は，文書分類のためのコーパスとして，Reuters'96 と RWCP コーパス (毎日新聞'94) を用いた [7]．Reuters'96 を訓練データとし，日英機械翻訳ソフトウェアを用い，RWCP コーパスを英語に翻訳した結果をテストデータとして分類を行なった．文書分類には SVMs を用いた．図1は，文書分類結果の例を示す．図1において，例えば，毎日新聞 $M1, M2, M4$ 及び $M5$ は Reuters'96 のカテゴリ A に分類されていることを表す．

表 1: 文書数のクロス表

	A	$\neg A$
B	$freq(A,B)=a$	$freq(\neg A,B)=b$
$\neg B$	$freq(A,\neg B)=c$	$freq(\neg A,\neg B)=d$

次に, 文書分類結果を用いてカテゴリ同士の類似性を求める. 分類結果に式 (1) で示す χ^2statistics を適用し類似度を求める.

$$\chi^2(A,B)=\frac{N\times(ad-cb)^2}{(a+c)(b+d)(a+b)(c+d)} \quad (1)$$

式 (1) において N は分類すべき文書の総数を示す. また, a, b, c, 及び d を表 1 に示す. 表 1 において $freq(A,B)$ はカテゴリ A と B に共に出現する文書数を示す. 式 (1) により得られた χ^2 値を式 (2) により正規化し, その値をカテゴリ同士の類似度とする. 類似度が一定の閾値以上である場合, その組を類似カテゴリペアとする.

$$\chi^2_{new}=\frac{\chi^2_{old}-\chi^2_{min}}{\chi^2_{max}-\chi^2_{min}} \quad (2)$$

2.2 類似カテゴリ集合の生成

処理 2 は 3 つ以上から成る類似したカテゴリの集合を生成する処理である. 本研究では処理 1 で生成した類似カテゴリペアに対して Agrawal らによって提案されたアプリオリアルゴリズムを適用することで, 類似カテゴリ集合を生成した [3], [4]. アプリオリアルゴリズムは, 式 (2) で求めた類似度が最小サポート値を満たす要素数 k の集合から, 末尾の要素のみが異なる集合を探し出し, その要素を結合することにより要素数 $k+1$ の集合を作成するというものである. 例として, 要素数 2 のカテゴリ集合から要素数 3 の類似カテゴリ集合を生成する処理を説明する. 類似度の閾値を 0.1 とした場合, 図 2 で表す要素数 2 のカテゴリ集合の中で類似度 0.1 以上のカテゴリ集合の個数は 4 となる. これら 4 つの集合から, 要素数 3 の類似カテゴリ集合の候補を生成する. 要素数 2 の類似カテゴリ集合の先頭に注目し, この集合に対して末尾のカテゴリのみが異なる集合を探し出し, そのカテゴリを注目のカテゴリ集合に結合することにより, 要素数 3 の類似カテゴリ集合の候補を生成する. 以上の処理を 4

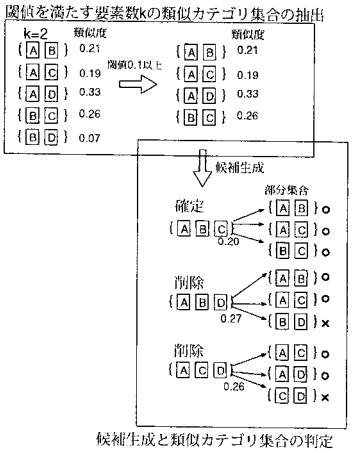

図 2: 類似カテゴリ集合の生成

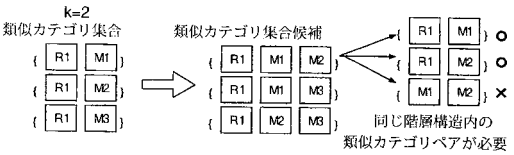

図 3: 階層構造内の類似カテゴリ集合の必要性

つの集合の各々に対して行ない, 候補を生成する. 次に, この候補が類似カテゴリ集合であるかどうかの判断を行う. 生成された候補の先頭に注目し, この候補から要素数が 1 だけ少ない部分集合の全てを作成し, 候補の生成元である要素数 2 の類似カテゴリ集合に存在するか否か, 判定する. 全ての部分集合が存在した場合にのみ, その候補は類似カテゴリ集合と判定される. この処理を繰り返すことにより要素数が 3 以上のカテゴリ集合を生成する. 得られた要素数 $k+1$ の類似カテゴリ集合における類似度は, 結合した 2 つの集合の平均とした.

類似したカテゴリ集合の生成において, 部分集合が候補生成元の類似カテゴリ集合に存在するかを確認する際に, 異なる階層構造のカテゴリ要素を持つ類似カテゴリペアだけでは, 要素数 3 以上の類似カテゴリ集合を生成することができない. 図 3 は, 要素 3 の類似集合が生成できない例を示す. そこで, 同じ階層内の親と子, 及び親と孫の関係を持つカテゴリ同士を無条件で類似カテゴリペアとし, 本手法により生成された類似カテゴリペアに追加することにより, 類似カテゴリ集合の生成を行った[1].

[1]同じ階層内の親と子, 及び親と孫の関係を持つカテゴリ同士の類似度は実験的に 0.5 と定めた

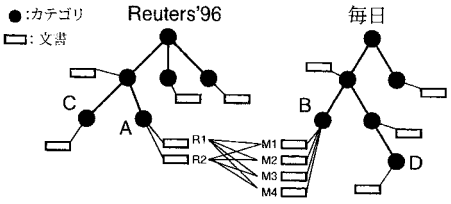

図 4: 類似文書抽出の例

3 類似文書抽出への適用

階層構造の統合結果の有効性を検証するため, 類似カテゴリペア, 及び類似カテゴリ集合の結果を用いて類似文書の抽出を行った. 図 4 は, 統合結果である類似カテゴリペアの例を示す. 図 4 において, Reuters'96 のカテゴリ A と毎日新聞のカテゴリ B が類似ペアとして抽出されたとする. カテゴリ A に属する Reuters'96 の文書, 及びカテゴリ B に属する RWCP の翻訳結果である文書をそれぞれベクトルで表現する. ベクトルの次元は文書に出現する名詞とし, ベクトルの要素は, TF*IDF[8] を用いて名詞に重み付けを行った結果とする. 余弦尺度を適用し, カテゴリ A, 及び B に属する任意の 2 つの文書の類似度を求める. 類似度が一定の閾値以上の 2 文書を類似した文書として抽出する.

4 実験

実験では, Reuters'96 と RWCP コーパス (毎日新聞'94) に付与されている UDC コードの階層構造を用いた. Reuters'96 は, 806,791 記事から成り, 126 のカテゴリに分類されている. RWCP コーパスは, 27,555 記事から成り 9,902 のカテゴリに分類されている. 実験では, 1996 年 8 月 20 日から 1997 年 2 月 19 日までの 384,655 記事から成る Reuters'96 の記事 (3 記事以上を含むカテゴリ数 102), 及び RWCP コーパス (毎日新聞'94) の 27,755 記事 (カテゴリ数 2,013) を用いた. 機械翻訳は, 「インターネット翻訳の王様バイリンガル ver5 IBM」を用いた. Reuters'96 の訓練データは 1996 年 8 月 20 日～11 月 19 日 (3ヶ月分) の 197,683 記事であり, テストデータとして 1996 年 11 月 20 日～1997 年 2 月 19 日 (3ヶ月分) の 186,962 記事を用いた. RWCP コーパスの訓練データは 1994 年 1 月 1 日～6 月 30 日 (半年分) の 13,955 記事であり, テストデータには 1994 年 7 月 1 日～12 月 31 日 (半年分) の 14,222 記事を用いた. Reuters'96 は tagger[9] を用い, RWCP コーパスは Chasen[10] を用い, それぞれ 836,822, 及び 105,888 語の名詞単語を抽出した. 類似文書抽出に用いたデータは, 1997 年 6 月 23～29 日の Reuters'96 の 14,897 記事と, 1997 年 6 月 26 日の毎日新聞を翻訳した 387 記事である.

表 2: 文書分類の精度

文書分類	適合率	再現率	macro-F
Reuters'96	0.738	0.766	0.752
RWCP コーパス	0.383	0.240	0.295

表 3: 類似カテゴリペア生成の結果

上位のペア	生成ペア数	正解数	適合率
1～200	200	157	0.785
201～500	300	163	0.543
501～1,000	500	269	0.538
1,001～2,000	1,000	351	0.351

4.1 文書分類結果

類似カテゴリペアを生成するために用いる文書分類の精度を検証した. 結果を表 2 に示す. 表 2 において, '文書分類' は, 訓練データ, 及びテストデータの種類を示す. 表 2 より, Reuters'96 の精度は, macro-F 0.752 であり, RWCP コーパスの分類精度 (0.295) よりも高い精度が得られていることがわかる. そこで本稿では, Reuters'96 を訓練データとし, RWCP コーパスを分類した結果を用いて類似カテゴリペアを生成する.

4.2 類似カテゴリペアの生成結果

本手法により類似カテゴリペアを生成し, 類似度の高い順, 上位 2,000 を人手によって類似しているかどうか判定し, 適合率を求めた. 結果を表 3 に示す. '生成ペア数' とは, 本手法により生成された類似カテゴリペアの数を示す. '正解数' とは, 生成された類似カテゴリペアの中で, 人手によって類似していると判断されたペアの数を示し, '適合率' は生成された類似カテゴリペアの中の正解数の割合を示す. 表 3 より上位 200 までは, 0.78 以上の適合率が得られた. 表 4 は, 閾値の変化による類似カテゴリペアとその適合率を示す. 表 4 より, 閾値が 0.05 以上の場合には, 0.86 以上の適合率が得ら

表 5: 類似カテゴリペアの例

ロイター	毎日	類似度	ロイター	毎日	類似度
Sports	スポーツ	1.000	Labour Issues	労働	0.196
Crime	刑法	0.436	Employment/Labour	労働	0.186
International relations	国際法	0.233	Legal/Judicial	刑法	0.159
Inventories	鉄鉱石	0.232	Wholesale Prices	長いす	0.158
Inventories	銑金の製造	0.232	Health	病理学	0.156
Sports	野球	0.223	Soft Commodities	食品	0.155
Sports	サッカー	0.217	Government Finance	税の形式	0.152
Science and Technology	スペースシャトル	0.203	Forex Markets	貨幣相場	0.149
Weather	一般地質学	0.197	DOMESTIC POLITICS	政党	0.148
Expenditure/Revenue	税の形式	0.196	EXPENDITURE/REVENUE	売上高税	0.147

表 4: 閾値別類似カテゴリペア生成の結果

閾値	生成ペア数	正解数	適合率
0.50	1	1	1.000
0.10	48	45	0.937
0.05	153	132	0.863
0.02	652	413	0.633
0.01	2,003	952	0.475

表 6: 閾値別類似カテゴリ集合生成の適合率

閾値	生成集合数	正解数	適合率
0.10	3	3	1.000
0.05	31	31	1.000
0.01	558	490	0.878
0.005	1,696	1,300	0.767

れていることが分かる. 生成された類似カテゴリペアの例を表 5 に示す. 表 5 から, 上位 1 位から 3 位までの 'スポーツ' や '刑法', 及び '国際法' など, 実際に類似しているカテゴリが類似カテゴリペアとして生成されている. しかし, 上位 5 位で示される 'Inventories' と '鉄鉱石' など, 誤ったペアも生成されている. また, 1 位は類似度が 1.000 で 2 位は 0.436 であることから, 類似したカテゴリペアの多くは類似度の値が 0.436 以下の低い値となっていることがわかる. 本研究では類似度の値を求める際に χ^2statistics を用いた. 今後は, 情報利得や相互情報量を用いた場合との比較を行うことで χ^2statistics の有効性を検証する必要がある. また, 正解データを作成することで再現率についても検証する必要がある.

4.3 類似カテゴリ集合の生成結果

生成した類似カテゴリペアに対してアプリオリアルゴリズムを適用し要素数 3 以上の類似カテゴリ集合を生成した. 結果を表 6 に示す. 表 6 において '生成集合数' は本手法により生成された類似カテゴリ集合の個数を示す.

表 6 から閾値 0.005 の時, 適合率が 0.767 であり, 正解数は 1,300 である. これより, 類似カテゴリペアよりも多くの集合が正しく抽出できていることが分かる. また, 閾値が低くなるほど, 上位の類似カテゴリ集合は類似したカテゴリ集合で占められていた. このことから, 要素数 3 以上の類似カテゴリ集合は詳細な分類を行うカテゴリの生成を行っていると考えられる.

4.4 類似文書の抽出結果

本手法を用いて統合したカテゴリを利用し, 類似文書の抽出を行った. 本手法の有効性を検証するため以下で示す 2 つのベースラインとの比較を行った.

1. **階層構造を利用しない手法**

   Collier[8] らの手法と同様であり, Reuters'96 と RWCP の各文書に対して, 類似文書の判定を行なった.

2. **Reuters'96 の階層構造を利用した手法**

   Reuters'96 の階層構造に対して, 文書分類を用いて RWCP の各文書を分類した. Reuters'96 の各カテゴリに対して, 予め Reuters'96 のカテゴリが付与されている各文書と分類された RWCP の各文書に対して, 類似文書の判定を行なった.

類似文書の判定は本手法と同様, 余弦尺度を用い, 一定の閾値を越える 2 つの文書を類似文書として抽出した. 本手法における類似文書の抽出では, 上位 180 の類似カテゴリペア及び, 上位 250 の類似

表 7: 類似文書抽出の適合率

閾値	階層構造を利用しない手法			一方の階層構造を利用した手法			類似カテゴリペア			類似カテゴリ集合		
	抽出数	正解数	適合率	抽出数	正解数	適合率	抽出数	正解数	適合率	抽出数	正解数	適合率
0.5	142	14	0.099	76	8	0.105	12	8	0.667	113	8	0.071
0.4	385	51	0.132	161	31	0.193	63	30	0.476	191	30	0.157
0.3	1,613	232	0.144	778	140	0.180	460	156	0.339	651	169	0.260
0.2	7,683	412	0.054	4,125	287	0.070	2,433	324	0.133	3,001	337	0.112
0.1	61,881	605	0.010	32,012	421	0.013	19,654	500	0.025	22,957	512	0.022
平均	14,340.8	262.8	**0.088**	7,430.4	177.4	0.112	4,524.4	203.6	**0.328**	5,382.6	211.2	0.124

カテゴリ集合 (要素 3 以上) を統合カテゴリとして用いた. 実験結果を表 7 に示す.

表 7 において, '抽出数' とは, 各手法により類似文書であると判定された文書ペアの数である. 表 7 より, 閾値が 0.5 の時, 本手法によって生成した類似カテゴリペアを用いた手法の適合率が 0.667 であり, 最も高い精度が得られていることが分かる. 一方, 類似カテゴリ集合を利用した場合の適合率の平均は 0.124 であり, Reuters'96 の階層構造のみを利用した手法 (0.112) とほとんど差がない結果となっている. 類似カテゴリ集合を利用した手法は, 類似カテゴリペアを利用した手法よりも誤って類似していると判断したペアを多く抽出している. これは, 3 つ以上の類似カテゴリを利用したため, 各カテゴリに誤って分類された文書が集まってしまったためであると考えられる. また, 抽出精度は閾値に大きく依存していると考えられることから, 類似カテゴリペア, 集合共に, 最適な閾値を求める手法を検討する必要がある.

5 まとめ

本研究では, 文書分類によりカテゴリ間の類似性を推定しカテゴリの統合を行う手法を提案した. さらに, 統合したカテゴリを用いて類似文書の抽出を行い, 統合したカテゴリの有効性を検証した. 実験の結果, 類似文書の抽出では, 階層構造を利用しない手法と比較して適合率が 0.240 向上することを確認した. 今後の課題としては, (1) 正解データを作成し, 再現率による検証を行うこと. (2) 機械翻訳の代わりに自動的に抽出した対訳語を用いて統合を行うこと [11]. (3) Web 階層構造へ適用することが挙げられる.

参考文献

[1] R.Ichise, H.Takeda and S.Honiden: Integrating Multiple Internet Directories by Instance-based Learning, *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 22–28 (2003).

[2] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer, New York (1995).

[3] R.Agrawal and R.Srikant: Fast Algorithms for Mining Association Rules, *Proceedings of Very Large Databases Conference*, pp. 478–499 (1994).

[4] R.Agrawal and R.Srikant: On Integrating Catalogs, *Proceedings of the 10th International World Wide Web Conference(WWW-10)*, pp. 603–612 (2001).

[5] Y.Yang and Y.Lui: A re-examination of text categorization methods, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'99)*, pp. 42–49 (1999).

[6] S.Dumais and H.Chen: Hierarchical Classification of Web Content, *Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR'00)*, pp. 256–263 (2000).

[7] RWCP.: RWC Text Database (1998).

[8] N.Collier, H. and A.Kumano: Machine Translation vs. Dictionary Term Translation - a Comparison for English-Japanese News Article Alignment, *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 263–267 (1998).

[9] H.Schmid: Improvements in Part-of-Speech Tagging with an Application to German, *Proc. of the EACL SIGDAT Workshop* (1995).

[10] Y.Matsumoto: 形態素解析システム「茶筅」Version2.3.3 使用説明書, *NAIST Technical Report* (1997).

[11] T.Utsuro, T.Horiuchi, T. K. and T.Nakayama: Effect of Cross-Language IR in Bilingual Lexicon Acquisiton from Comparable Corpora, *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics(EACL'05)*, pp. 355–362 (2003).