3F2-04

類似関係にある複数カテゴリの組み合わせの発見

Discovering Combinations of Similar Categories

市瀬 龍太郎[*1]　武田 英明[*2]
ICHISE Ryutaro　TAKEDA Hideaki

[*1]国立情報学研究所知能システム研究系
Intelligent Systems Research Division, National Institute of Informatics

[*2]国立情報学研究所実証研究センター
Research Center for Testbeds and Prototyping, National Institute of Informatics

In the present paper, we propose a method to discover combinations of similar categories in two different concept hierarchies. We conducted an experiment to show the ability of our method using real-world Internet directories. The results of this experiment show that the proposed method can find similar category combinations on the Internet directories.

1. はじめに

多量の情報の中から必要なものを探すために，情報の階層的分類はしばしば使われる手法である．その手法では，あらかじめ，情報を分類しておき，階層をたどることにより，必要な情報に容易にたどりつくことができる．これらが，使われている例として，インターネットディレクトリ，図書目録分類，オントロジーなど，さまざまなものをあげることができる．しかし，これらで使われる概念体系は，通常，一つのものだけが使用されており，複数の概念体系を使う事はできない．このため，異なる概念体系を用いた複数の分類階層があった時に，同じものに対する分類を行っていたとしても，その分類情報をお互いに利用する事は難しい．

筆者らは，このような問題に対して，情報の分類の類似性に着目をして，異なる概念体系の間で類似するカテゴリを発見する手法を提案してきた [Ichise et al. 03]．類似したカテゴリの対応が分かると，分類情報を別の概念体系の類似したカテゴリでも利用できる．しかし，従来の研究では，1 対 1 の対応に限られていた．異なる概念体系においては，常に 1 対 1 でカテゴリ間のマッピングを取ることができるとは限らない．本論文では，1 対多のカテゴリ間で，類似性を発見する手法を提案する．このような関係を発見することができれば，より的確な概念間のマッピングを取れるのみならず，異なる概念体系間で，概念体系の再編を行ったりすることに利用できると考えられる．

2. 概念階層のモデル

本研究で取り扱う概念階層は，分類カテゴリが階層状になっているものを対象とし，それぞれの分類カテゴリにインスタンスを割り当てることができるものとする．図で表すと，図 1 のようになる．ここで，黒丸で表されるのが，カテゴリであり，カテゴリは，階層構造を構成している．そして，各カテゴリには，そのカテゴリに属するインスタンスが割り当てられる．階層の中間にあるカテゴリに対しても，インスタンスの割り当てができるものとする．

本研究では，インスタンスの分類が似ているならば，類似するカテゴリとして取り扱う．例えば，文学作品の Web ページをインスタンスとして分類する時に，ある概念階層では，タイトルにより分類を行い，別の概念階層では，著者によって分類を行う場合を考えよう．その時に，源氏物語とした分類と紫式部とした分類は，源氏物語が紫式部の代表作であるため，Web ページの分類は，ほぼ一致すると考えられる．この時，2 つの分類が似ているため，2 つの概念階層の分類基準が異なるにもかかわらず，類似カテゴリとして考えることができる．

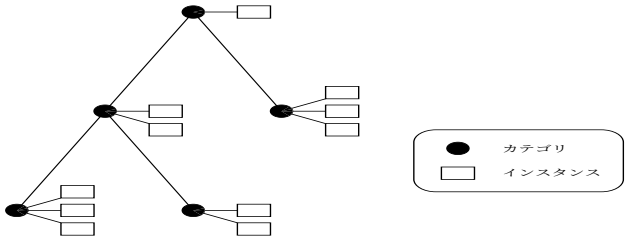

図 1: 概念階層のモデル

この考え方は，同様に複数の概念に対しても拡張をして考えることができる．例えば，ある概念階層で使われるインターネットクライアントのページという分類概念は，別の概念階層で使われているインターネットエクスプローラーと Mozilla の分類概念を組合わせたものである場合などが考えられる．本研究では，このような 1 対多の概念同士の類似関係を発見することを目標とする．

3. 類似概念発見手法

複数のカテゴリの分類で類似するものを単純に探すだけだと，意味的に全く関係のない概念同士の組合せが出て来る可能性がある．そこで，本研究では，カテゴリの組合せとして，同じ分類階層のカテゴリのみを組み合わせるものとする．つまり，カテゴリの組合せとして，兄弟となるもののみの組合せを発見することとする．この関係を図 1 と同様な形式で描くと，図 2 となる．この図においては，カテゴリ A の分類と類似するものは，カテゴリ B と C の組合せが対応するものとなる．

また，含まれるインスタンスの個数が少ないカテゴリ同士の組合せを見付けるのを避けるために，ミニマムサポートの概念を導入する．ミニマムサポートは，割合で指定され，全体のインスタンスの中で，ミニマムサポートの割合以上のインスタンスを含むカテゴリが発見の対象となる．

連絡先: 市瀬龍太郎，国立情報学研究所知能システム研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，ichise@nii.ac.jp

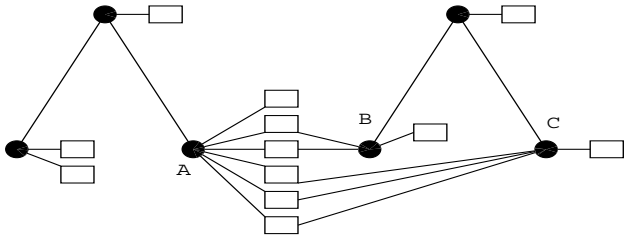

図 2: 複数の概念階層の例

類似したカテゴリを発見する手順は，次のようになる．

1. 2 つの概念階層 H_1, H_2 とした時に，H_1 からミニマムサポート以上のインスタンスを含むカテゴリを列挙.
2. 1 で見付けたそれぞれのカテゴリに対して，ミニマムサポートの制約を満たし，インスタンスを介してつながっている H_2 のカテゴリを列挙.
3. 2 で見付けたそれぞれに対して，階層内の兄弟の組合せを全て作成.
4. 3 で見付けたもの全てに対して，カテゴリに類似性があるかをテスト.
5. 4 のテストを通ったもの全てを出力.

カテゴリに類似性があるか否かの判定は，Hical [Ichise et al. 03] と同様に κ 統計量 [Fleiss 73] を用いる．

4. 実験

提案手法で，類似する概念の組合せが発見できるかどうかを調べるため，インターネットディレクトリのデータを用いて，実験を行った．Yahoo!と Google で使用されているインターネットディレクトリから，下記のカテゴリ以下のデータを取り出し，カテゴリに所属する Web ページをインスタンスとして類似カテゴリの発見を行った．

- Google : Arts / Movies
  Yahoo! : Entertainment / Movies_and_Film
- Google : Recreation / Outdoors
  Yahoo! : Recreation / Outdoors
- Google : Arts /Photography
  Yahoo! : Arts / Visual_Arts / Photography
- Google : Computers / Software
  Yahoo! : Computers_and_Internet / Software

これらのディレクトリが含んでいるデータ数は，表 1 のようになっている．リンク数は，各カテゴリの中に含まれる URL の総数を示しており，同じ URL でも異なる複数のカテゴリに含まれる場合には，それぞれ別個に数えている．共有リンク数は，両方の階層からリンクされている同じ URL の数を示している[*1]．

これらのデータを用いて，類似カテゴリを発見する実験を行った．実験の際には，κ 統計量の有意水準として 5%，カテゴリを検出するためのミニマムサポートとして，0.8%を用いた．また，Google のディレクトリに対して，Yahoo!で類似するカテゴリを求める実験を行った．その結果，表 2 が得られた．また，発見できたカテゴリの対応の例を以下に示す．

	Yahoo!		Google		共有リンク数
	カテゴリ数	リンク数	カテゴリ数	リンク数	
Movies	5209	22193	6818	31913	2960
Outdoors	2595	15074	1107	15935	799
Photography	566	5794	273	5317	727
Software	627	3822	2289	40868	910

表 1: 実験データ

	発見した数
Movies	12
Outdoors	332
Photography	1948
Software	216

表 2: 発見したカテゴリの組合せの数

- Google: Outdoors/Organizations
  Yahoo!: Outdoors/Institutes/
  Yahoo!: Outdoors/Clubs_and_Organizations/
- Google: Photo/Techniques_and_Styles/Documentary
  Yahoo!: Photo/Photojournalism/Photojournals/
  Yahoo!: Photo/Photojournalism/Photo_Essays_and_Exhibits/

実験結果より，提案手法で，類似しているカテゴリの組合せが発見できることが示された．発見できたカテゴリの組合せの数は，用いたカテゴリの種類によって大きく異なっている．これは，カテゴリ数の差によるものだと考えられる．カテゴリ数が多くなると，一つのカテゴリに含まれるインスタンスの数が少なくなってしまうため，サポートを満たす事ができなくなるカテゴリが多くなる．その結果，発見したカテゴリの組合せの数が少なくなってしまうと考えられる．一方，発見数の多いカテゴリでは，同じような組合せがたくさん出力されており，そのことが数の増加を招いていると考えられる．

5. おわりに

本研究では，複数の概念階層において，類似している複数のカテゴリの組合せを発見する手法を提案し，その手法を Google と Yahoo!のインターネットディレクトリを使って，検証を行った．実験結果より，提案手法により類似したカテゴリの組合せを発見できることが示された．今後は，似たような類似カテゴリの組合せが多数発見されてしまうのをどのように防ぐか，サポート値の最適な設定方法はどのように決めればいいか，兄弟以外で有用な組合せを発見する手法などについて研究をする．

参考文献

[Fleiss 73] Fleiss, J. L.: *Statistical Methods for Rates and Proportions*, John Wiley & Sons (1973), (邦訳 : 係数データの統計学, 佐久間 昭 訳, 東京大学出版会 (1975)).

[Ichise et al. 03] ICHISE Ryutaro, TAKEDA Hideaki, HONIDEN Shinichi: Integrating Multiple Internet Directories by Instance-based Learning, In Proceedings of the 18th International Joint Conference on Artificial Intelligence(IJCAI-03), (to appear).

*1 共有リンク数は，同一 URL を一度しか数えないのに対して，リンク数は，複数のカテゴリに出現する場合に複数回数えるので，一部のカテゴリでは，共有リンク数は，リンク数よりも非常に数が少ない．