トピックマップデータベースへの問い合わせ最適化手法の検討

栗原　優樹†　　木村　昌臣‡

芝浦工業大学大学院工学研究科†　　芝浦工業大学工学部情報工学科‡

1. 研究背景と目的

膨大な量の情報が個々に独立して存在していることが情報の検索において大きな妨げとなっていることから，情報間の関係を管理することにより効果的に情報を見つけ出すトピックマップが重要な技術の1つとして着目されている．トピックマップは主に，概念を指す「トピック」，トピック間の関係を表す「関連」，そしてトピックとWeb等の情報リソースとの関係を表す「出現」から構成される[1]．また，トピックマップを構成するこれらの情報を格納し，操作するためのトピックマップデータベースが構築され，トピックマップ専用の問い合わせ言語[2]が提案されている．データベースからトピックマップの情報を取得する場合，特に規模が大きくなると検索時間が大きく増加するため，問い合わせの最適化が重要になるが，従来のデータベースでは考慮されていない．そこで，本研究ではトピックマップデータベースの問い合わせ最適化手法の検討を行い，筆者らが構築したトピックマップデータベース（以降，TOME[3]）に適用し，検証を行う．

2. 検索ルートを利用した最適化手法の検討

2.1. データモデルと検索ルートの検討

問い合わせが行われた時，より早く結果を見つけるため，入力された問い合わせの結果を検索するための検索ルートを洗い出し，その妥当性を検証する必要がある．そのため，本研究で対象とする問い合わせとしてトピック間の関係を取得することが重要であると考え，ユーザが指定したトピックと指定した関係を持つトピック群を検索するための構文に限定した．トピックマップにはデータモデル（以降，TMDM[4]）が定義されており，トピックや関連を含む7種の情報項目で構成されている．TOMEではこれをオブジェクトとして実装している．

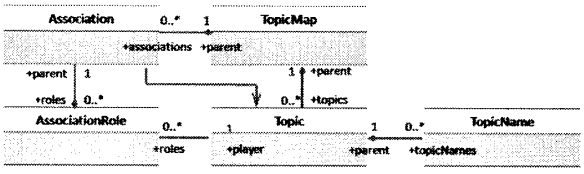

図 1：情報項目間の参照関係

Query optimization for the Topic Maps Database
†Yuuki Kuribara, Graduate School of Engineering, Shibaura Institute of Technology
‡Masaomi Kimura, Shibaura Institute of Technology

トピックマップはトピックと関連で構成され，関連役割によって関連とトピックが繋がっており，また，トピックに付随してトピック名が存在する．これらの関係を表したものが図1である．本研究の検索方法はこのTMDMに従って行い，前述した構文の検索結果を得るための検索ルートは図2，図3の2ルートが挙げられる．

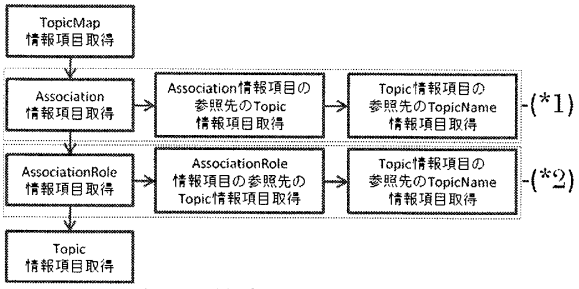

図 2：関連ルートの検索手順

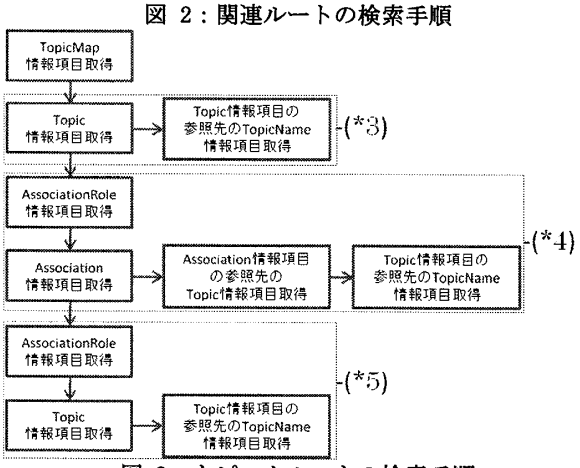

図 3：トピックルートの検索手順

3. 検索コストの見積もり

対象とするトピックマップによって最適なルートを選んで検索が行えるようにするため，各ルートで想定される検索コストを見積もるための式の定義を行う．なお，取得するオブジェクトのサイズは情報項目ごとに異なるが，本提案では全て統一して同じサイズとして扱っている．

3.1. 関連ルートの検索コストの見積もり

関連ルートの検索コストの見積もりは，指定された名称を持つ関連を取得する時間（図2-(*1)），指定された名称を持つ関連の関連役割を取得する時間（図2-(*2)）の合計とした．以下，全関連数N，全トピック数M，ユニークな関連数Qとする．

図2-(*1)について，TMDMにおいて名称が同じ

関連は複数存在可能であり，全ての関連に対してオブジェクトを取得して検索を行う必要があることから，コストは全関連数に比例する．また，検索は同時に関連及びその関連のトピックとそのトピック名が取得されるため，各関連について3種類のオブジェクトを取得することになる．

図2-(*2)について，指定された名称を持つ関連の関連役割のみを取得すればよいため，コストは指定された関連名を持つ関連の数に比例する．また，同時に関連役割とその役割を持つトピックとトピック名が取得される．各関連は2つの関連役割を持つため，6種類のオブジェクトを取得する．

以上から関連ルートのコストは

$$3 \times N + 6 \times \frac{N}{Q}$$

と定義する．

3.2. トピックルートの検索コストの見積もり

トピックルートの検索コストの見積もりは同様に，指定されたトピックを取得する時間（図3-(*3)），指定されたトピックに定義された関連を取得する時間（図3-(*4)），指定されたトピックに定義された指定された名称の関連先のトピックを取得する時間（図3-(*5)）の合計とした．

図3-(*3)について，TMDMにおいてトピックは重複を許さないため，見つかり次第終了となることから，コストは1つのトピックを順次に探していった場合の平均時間に比例する．また，同時にトピックとそのトピック名が取得されるため，2種類のオブジェクトを取得することになる．

図3-(*4)について，指定されたトピックの持つ関連の数のみを取得すればよいので，コストは各トピックの持つ関連の数の平均に比例する．また，同時に関連役割とその役割を持つ関連，その関連を表すトピックとトピック名が取得されるため，4種類のオブジェクトを取得することになる．

図3-(*5)について，図3-(*4)で取得された関連のうち，指定された関連名を持つ関連の数のみを取得すればよいため，コストはこれに比例する．また，同時に関連役割とその役割を演じるトピックとトピック名が取得されるため，3種類のオブジェクトを取得することになる．

以上からトピックルートのコストは

$$2 \times \frac{N}{2} + 4 \times \frac{N}{M} + 3 \times \frac{N}{MQ}$$

と定義する．

4. 実験と考察

提案した手法をTOMEに適用し，2種類のトピックマップを用意し，検証を行った．トピックマップはポケモンとそのタイプ，型をトピックとし，各ポケモンの進化とタイプを関連としたポケモントピックマップ（トピック数：174個，関連数：432個）と，江戸川乱歩とその著作物，出身地をトピックマップとした江戸川乱歩トピックマップ（トピック数：29個，関連数15個）の2種類を作成した．実験は検索ルートに着目した検索効率化の有用性と提案した見積もり式の妥当性を調べるため，それぞれのトピックマップに対し，両検索ルートで検索を行い，実行時間を10回ずつ計測して平均を算出し，また，定義した計算式で両ルートの検索コストを計算し比較した．（表1）

表1：実験結果

トピックマップ名 トピック数 関連数	検索の条件 (トピック/関連名)	関連ルート	コスト（関連ルート）	トピックルート	コスト（トピックルート）
乱歩トピックマップ トピック:39 関連:15	江戸川乱歩/生まれる	31	75	157	42.8
	江戸川乱歩/著す	47		187	
	明智小五郎/登場する	78		203	
ポケモン トピックマップ トピック:174 関連:432	ピカチュウ/進化	250	2160	47	198.8
	ノーマル/属する	297		156	

実験結果から，検索ルート毎に比較するとどれも3～5倍程度の差が出ていることがわかり，また，トピックマップが変わるとより早い検索ルートも変化していることから検索ルートを選択した検索の有用性が示されている．また，事前に見積もったコストに対する実行時間の比率を見てみると，ポケモントピックマップに関しては大よそ期待通りの結果になっており，有効に機能している．しかし，乱歩トピックマップの場合，実際の検索時間が関連ルートの方が早いのに対して見積もった時間はトピックルートの方が早いと示している．これは両結果から関連ルートの見積もり方が大きいことが原因と考えられる．また，取得するデータの違いによるデータの大きさの違いも考慮にいれる必要があると考えられる．

5. まとめと今後の課題

本研究では，トピックマップデータベースへの最適化手法の提案と，TOMEへの提案した手法の適用を行い，実験を通して本手法の妥当性の検証を行った．今後の課題として，コスト計算の精度の向上，インデックスなどによる情報検索の効率化が挙げられる．また，今回は検証のため小規模な2種類のトピックマップで行ったが，今後はより大規模なトピックマップを用意し検証を行う．

参考文献

[1]内藤求：トピックマップ入門，東京電機大学出版局(2006).
[2]Ontopia：tolog Language tutorial (2007). http://www.ontopia.net/
[3]栗原優樹，細谷岳志，木村昌臣：拡張したトピックマップデータベースの構築（2009）
[4] ISO/IEC JTC1/SC34：Topic Map – Data Model(2008). http://www.isotopicmaps.org/sam/sam-model/