

文書の目次構造を用いた概念情報検索

5 N-5

伊藤 達雄

(株)リコー 情報通信研究所

1. はじめに

マニュアルなどの文書からあるトピックに関連した内容を探す場合に索引や目次を利用することが多いが、索引からでは単語によるキーワード検索と同じようにトピックの内容を検索要求として十分に指定できないため、意図に反した結果を得がちである。それに比べて目次には、各節に書かれた内容を概念的に表現した題目だけでなく、章の題目として節を補足する情報も含まれているので意図を反映しやすい。[1]では文献を検索する際に目次の情報が人が付与するキーワード以上に検索に有用であることが示されている。

我々は重要キーワードを直接文書から抽出する方法[2]も検討しているが、本稿では目次自体が持つ情報に着目し、章や節の題目から得られる概念と、トピックの概念との類似度による検索手段を提案すると共にその効果を検討する。

2. トピックの概念表現と類似度

目次の中の題目に限らず文章は複数の単語で構成されるが、単語自体の意味が文脈や状況によって異なることがあるため、文章が意味するトピックを正確に表現することは難しい。そこでトピックの大まかな内容を概念と呼び、トピックを適切に表現するような単語の集合によって表現されるものとする。

トピック T_i の概念 ⇔

$$C_i = \{W_m \mid T_i \text{ を適切に表現する単語} \}$$

またトピック T_i の概念 C_i に上位概念 C_u が定義されているとき、要素の和集合 C_i^u を新たにトピック

ク T_i の概念とする。

$$C_i^u = C_i \cup C_u$$

あるトピック T_i と別のトピック T_j の間の類似度は、それぞれの概念 C_i 、 C_j の間の類似度で表わされる。概念は単語から構成されているので、まず単語間の類似度を定義する。日本語の場合、単語の多くは表意文字である漢字であるため、単語文字列自体が意味を持つものと考え、文字列間の類似度を意味的な類似度と捉えている。¹

単語 W_m における W_n との類似度 ⇔

$$SW_{W_m}(W_n) = OW_{W_m}(W_n) / LW(W_m)$$

$OW_{W_m}(W_n)$: W_m の中に含まれる W_n との重複文字列長
 $LW(W_m)$: W_m の文字列の長さ

概念は単語の集合で表現されるので、概念間の類似度 $SIM(C_i, C_j)$ は、 C_i の要素(単語)中に含まれる重複文字列の度合いと、 C_j の要素中に含まれる重複文字列の度合いを加味したものとする。

$$SIM(C_i, C_j) = SC_{C_i}(C_j) + SC_{C_j}(C_i)$$

$$SC_{C_i}(C_j) = OC_{C_i}(C_j) / LC(C_i)$$

$$OC_{C_i}(C_j) = OW \sum_{W_m \in C_i} W_m (\sum_{W_n \in C_j} W_n)$$

$$LC(C_i) = \sum_{W_m \in C_i} LW(W_m)$$

ただし、 $\sum_{W_m \in C_i} W_m$ は C_i の全ての要素を結合した文字列

3. 目次の概念化と概念情報検索

検索単位となる節の内容をトピックとするとトピックの概念は適切な単語の集合で表現されなければならない。ここでは節の題目が内容を適切に表現していると考え、題目からキーワード抽出により得られる単語の集合を節の概念とする。

Concept-based information retrieval using the table of contents

Tatsuo Ito (tatsuo@ic.rdc.ricoh.co.jp)

Information and Communication Research and Development Center
 RICOH COMPANY, LTD.

¹[2]では文字列の類似度により文章から重要文を抽出している。

ただし、節の目次階層的な上位に節や章がある場合は、上位の題目の概念をその節の上位概念と捉え、各節において上位概念の融合を行なったものを改めて節の概念とする。

ここで、検索要求自体もトピックとして概念で表現し、ある概念に対し類似度が高い概念は意味的な関連度が高いものとする事で、検索要求に関連した情報を検索することができる。以下に目次の概念化及び概念情報検索のステップを示す。

- 1: 対象文書の目次部または文書構造を解析し、検索単位を設定する。
- 2: 検索単位における章や節の題目からそれぞれの概念を作成し、目次構造に合わせて上位概念を融合し、検索単位毎の概念とする。
- 3: 2で得られた概念を検索対象とする。
- 4: 検索要求トピックから概念を作成する。上位概念があれば融合する。
- 5: 要求概念と検索対象内の概念を比較し、類似度を測定する。
- 6: 類似度の高い概念の検索単位を提示する。
- 7: 必要に応じて、検索結果を検索対象とし、新たな検索要求を与え、4へ戻る。

4. 実験

我々は故障診断にマニュアル検索機能を融合した情報提供的な診断システム FIXIT[3]を開発している。診断時に現象、原因、質問事象に関連したマニュアル情報を提示することで、情報ツールとして知識と関連情報を提供できることを特徴としている。このシステムの中で、今回提案した方法により、知識内の事象をトピックと捉え、概念的に類似したマニュアル情報の検索を行った。

あるファクシミリの診断知識(事象の数: 201)とその操作説明書(題目数: 309)を対象とし、提示されたトピックが診断において適切であるかどうかを事象毎に検証し、評価を行った。ここで事象の概念は説明文や質問文から構築し、上位概念として事象名やそのカテゴリ名を用いた。ただし、これらの事象は検索することを意図せずに作

成されている。

全事象中、マニュアル内に関連情報が存在する172件を有効対象とし、複数の関連情報を得るため、類似度上位5位までを検索結果とした。

	目次階層	題目のみ
上位3位以内に該当題目	91.9%	90.1%
上位5位以内に該当題目	94.8%	93.0%
占有率2/5以上に該当題目	80.4%	60.3%

例えば事象「電話回線の接続確認」では、

- 1) 加入電話回線と接続する、
- 2) 2台目の電話を接続する、
- 3) ISDNと接続する、

等の題目が順に類似度上位に現れ、いずれの題目も診断状況によっては、関連性のあるトピックと考えられる。

目次階層を使わない場合は、占有率に差が生じ、検索結果の中に本来関連のないものが目立った。

5. まとめと今後の課題

目次の構造を概念的に捉え、検索要求の概念との類似度による検索方法を提案した。概念間の類似度を文字列の類似度に置き換えることで、意味辞書を用いることなく関連した情報検索が行えること、文書内のトピック検索に題目が使えること、目次構造が複数の関連情報を見つけるのに役立つことを示した。現状では直接関係する概念がない場合でも誤った類似概念を列挙するため、今後は類似度の適切な閾値の設定等を検討したい。

謝辞 本研究においてツール及び有益な助言を頂いた亀田雅之研究員に感謝する。

参考文献

- [1] 長尾真: 目次情報などを利用した図書・文献検索方式, 情報の科学と技術, Vol42, No8, p.711-718(1992).
- [2] 亀田雅之: 擬似キーワード相関法による重要キーワードと重要文の抽出, 言語処理学会 第2回年次大会, 1996.
- [3] Hart, P.E. and J. Graham: Query-free information retrieval, Second International Conference on Cooperative Information Systems(CoopIS), 1994.