1U－5　概念階層を介した索引づけと用語間の概念的関係を考慮した測度による文献のランキング

中島 誠　伊藤 哲郎

大分大学工学部 知能情報システム工学科

1 はじめに

近年，様々な分野で概念階層としてのシソーラスや分類体系（以下まとめて概念階層と呼ぶ）の整備がなされている．これに伴い，質問や文献の主題が，概念階層を通じて関連づけられた索引用語によって表わされているとし，それらの間の概念的関係を質問との類似性判断に反映させて文献をランキングする方法を提案した [1]．この方法を用いれば質問や文献の表現上の違いに注意しなくとも望んだ文献を取り出しやすくなる．

しかしながら，このような索引用語が付加された文献の数は少ないのが現状である．文献中に現れる語を手がかりに索引用語をうまく選び出す方法が定式化できれば上記の方法を利用する助けとなるが，文献中で用いられる語は通常複数の意義を持ち，文献によってその使われ方を限定する必要がある．ここでは，このような語の多義性を解消しながら索引づけを行なう方法を提案し，その有効性を調べる．

2 文献のランキング

質問との概念的な類似性に従った文献のランキングがどのようになされるかを [1] に従ってまとめておく．質問や文献の主題を表す索引用語間の関連性は，概念的により一般的なものがより特定的なものより下のレベルになるよう配置された概念階層によって捉えられる．このとき，質問 - 文献間の類似性判断は，1) 質問と各文献の表現から，両者の概念的関連性を表すための被覆を求め，2) 各被覆について，その特定度を計算して指標とすることでなされる．

ここで，質問と文献の表現の被覆は，両者に含まれる索引用語をまとめて言及する一般的な索引用語の組である．このような索引用語を求めることを索引用語の一般化という．概念階層を通じて見たとき，関連の深い索引用語についての一般化結果は，より特定的になる．このことに注目し，被覆を質問側の個々の索引用語と文献側で最も関連深いものとの一般化結果と，これとは逆に文献側の個々の索引用語と質問側で最も関連深いものとの一般化結果を集めることで得る．被覆の特定度は，概念階層のレベルに応じた値を被覆中の索引用語に与えて（最上位レベルにあるものに最も大きな値が与えられる），その平均値とする．質問と各文献との被覆が求まると，被覆の特定度に従った文献のランキングが得られる．

Indexing based on Word Sense Disambiguation and Conceptual Document Ranking
Makoto NAKASHIMA Tetsuro ITO
Oita University, 700 Dannoharu Oita 870-1192, Japan

3 索引づけ

提案する索引づけの方法は，基本的には，出現頻度の高い語や句をキーワードあるいはキーフレーズとして抽出し，これらを部分文字列とする索引用語を選んで文献に付加するようにする．ただし，通常キーワード・キーフレーズが多義性を有することも原因して，複数の索引用語が選び出されてしまう．ここでは，同一文献中から抽出されたキーワード・キーフレーズは互いに概念的に深い関連があり，その使われ方は概念的に矛盾がないものと捉えて，索引用語を選び出すようにする．

文献より抽出したキーワード・キーフレーズの集まり $\{w_i | i \geq 1\}$ を W とする．w_i を部分文字列とする索引用語の集まりを $\{t_{i_n} | n \geq 1\}$ として T_i と書く．w_i について $t_{i_n} \in T_i$ を考える．$w_j (\neq w_i)$ を部分文字列とする索引用語の集まり T_j の中で，概念階層中で見て t_{i_n} と最も関連の深い索引用語 t_{j_m} を t_{i_n} の支持語と呼び，これらの組を (t_{i_n}, t_{j_m}) と書く．ある索引用語の集まり $E \subseteq \bigcup_i T_i$ について，その中の任意の索引用語の支持語が同様に E に含まれるとき，E を文献の表現の候補と呼ぶ．索引づけの手続きを以下にまとめる．

(S1) 文献よりキーワード・キーフレーズの集合 W を求め，各 w_i について T_i を求める．

(S2) 各 T_i に含まれる索引用語とその支持語の組を求め，その重みを計算する．

(S3) 文献の表現の候補で評価値が最大となる E を見つけて，文献の表現とする．

索引用語とその支持語の組 (t_{i_n}, t_{j_m}) の重み $p(t_{i_n}, t_{j_m})$ はキーワード・キーフレーズの文献中での出現頻度と，文献中で特定的な概念についてはより一般的な概念を用いた説明がなされていることを考慮して，$(\beta \cdot tf_i) \cdot tf_j$ とする．$tf_i (tf_j)$ は，$w_i (w_j)$ の文献中での出現頻度で，β には t_{i_n} より特定的あるいは一般的な索引用語が $T_j (\neq T_i)$ の中にあれば 1 より大きな値を与える．E の評価値は次式により求める．

$$\frac{\sum_{w_i \in W} av(E \cap T_i)}{|W|}$$

ここで,

$$av(E \cap T_i) = \frac{\sum_{t_{i_n} \in E \cap T_i} av(t_{i_n})}{|E \cap T_i|},$$

$$av(t_{i_n}) = \frac{\sum_{t_{j_m} \in E_{i_n}} p((t_{i_n}, t_{j_m})) - \overline{p}}{|E_{i_n}|}.$$

E_{i_n} は E に含まれる索引用語 t_{i_n} の支持語の集合, $\overline{p}$ は全ての索引用語と支持語の組の重みの平均である.

上記手続きの S3 で全ての可能な文献の表現の候補を調べることは現実的でない. これには, 遺伝的アルゴリズム (GA) を用いて対処する. GA では, 解候補を表す個体の集合に遺伝的操作を繰り返して, 効率的に問題を解く. ここでは, 文献の表現の候補 E を個体とし, $|W|$ 個のビット文字列のリストで表した遺伝子型に対応させる. 遺伝子型の i 番目の文字列は T_i に対応し, 文字列中の n 番目の"1" のビットは, t_{i_n} が E に含まれていることを表す.

4 実験

実験では ACM の 5 つの論文誌に掲載された文献 280 編を対象に GA を用いた索引づけの手続き (以下, GDA と記す) により索引用語を選び, その有効性を見た. 各文献には, ACM の CR 分類体系から人手により選ばれた索引用語が付加されていた. T_i のサイズが大きな w_i は, それだけ広範な概念を表しており, 主題の表現のために重要でないと言える. このことを考慮し, 手続きの S1 において, T_i のサイズが $\alpha(\geq 1)$ 以下の w_i のみを文献の概要より抽出した. GA での個体数は 50, 繰り返し数は 100 とした.

まず, GDA によりどれだけ正確な索引用語が選ばれているかを, 選ばれた索引用語と人手で付加された索引用語の数の和に対するこれらに共通するものの数の割合で見た [2]. ここで, GDA により選ばれた (あるいは, 人手で付加された) 索引用語の中で, 他により特定的なものがあるものは削除した.

評価結果を 280 編の文献に対する平均で表 1に示す. 表中の数値は, 最良の (S1 で文献に付加された索引用語のみよりキーワード・キーフレーズを抽出した) 場合の結果に対する割合である. 比較のために, 各 T_i 中の索引用語すべてを選んだ場合 (NDA と記す) と, w_i について, $T_j(\neq T_i)$ の中により一般的な索引用語がある $t_{i_n} \in T_i$ を選んだ場合 (SDA) の結果も示した.

表 1: 正しく選ばれた索引用語の割合.

	GDA			NDA	SDA
α	β=5	β=10	β=20		
5	.41 (2.5)	.42 (2.3)	.40 (2.2)	.28 (11.1)	.33 (1.5)
10	.44 (2.6)	.47 (2.6)	.45 (2.6)	.23 (17.2)	.40 (3.7)
20	.40 (3.2)	.41 (3.7)	.41 (3.1)	.18 (23.3)	.39 (5.2)

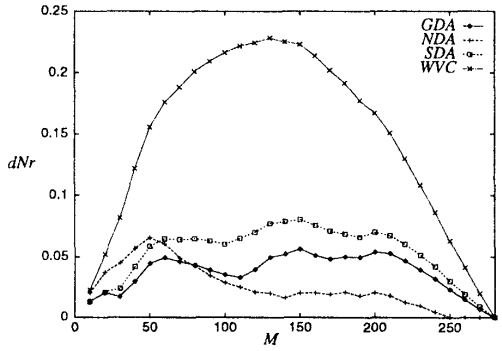

図 1: 検索効率.

括弧内の数字は選ばれた索引用語の数の平均である. GDA ではより少数の索引用語で良い結果が得られていることがわかる.

次に, 検索効率の観点より索引づけの評価を行なった. これには, 任意の文献を質問とし, 2 節で示した方法で 280 編の文献をランキングして見た. 検索効率は, ランキング上位 M 編の文献の中の適合文献の数を見ることで調べられる. 各論文誌にはそれぞれ主題を同じくする研究分野の文献が掲載されていることから, 質問と同じ論文誌に掲載されていれば, 適合文献として判断した. M が小さい段階でより多くの適合文献が得られる程良いといえる. 評価のため, 人手で付加された索引用語を文献の表現に用いた場合と, ここでの索引づけ方法により選ばれた索引用語を用いた場合との適合文献の数の差 dNr を調べた.

図 1に α=10, β=10 の場合の GDA, NDA, SDA の結果を示す (dNr は M に対する適合文献の数の差を質問と同じ論文誌に掲載された文献の数で正規化してある). 横軸に近い結果程良い. M が小さい段階 ($M \leq 70$) で GDA が最も優れていることがわかる. 広く用いられているキーワード・キーフレーズのベクトルを用いて文献を表現した場合の結果も調べたが (図中 WVC), 他のいずれよりも劣っていた.

5 おわりに

質問との概念的関連を考慮した文献ランキングのために, 文献から抽出したキーワード・キーフレーズをもとに, 複数の索引用語の中から主題表現のために重要と思われるものを選び出し, 索引づけを行なう方法を定式化した. 今後, 実際の検索システムの構築に適用していく予定である.

なおこの研究の一部は平成 9 年度文部省科学研究費基盤研究 (C)09680402 による.

参考文献

[1] 中島 誠, 金子 雄一知, 伊藤 哲郎: 用語間の概念的関係を考慮した測度による文献のランキング, 電子情報通信学会論文誌 D-I (掲載予定).

[2] Salton, G. and McGill, M. J. "Introduction to Modern Information Retrieval," McGraw-hill, New York (1983).