5Y－5

キーワード間の概念的関係を考慮した質問と文献の類似性判断

金子 雄一知　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

電子図書館の普及に伴い大量の文献情報を扱えるようになったことから，質問 - 文献間の類似性判断が容易な検索システムの整備が必要となっている．検索結果を質問との類似度の降順にランキングしてユーザに提示する方法は索引語がうまく選ばれており，類似測度が妥当なものであれば，検索効率の面からみて最も優れた方法と考えられる [4]．ただし，質問や文献を表現するキーワード間の概念的関連をうまく扱えなければ，これらの表記上の違いによって必要な文献を取り出すことが困難になる [2]．

ここでは，質問や文献を表現するキーワードの概念的関連を考慮しながら，質問と文献間の類似性判断を行なう方法を概念学習での一般化記述を参考に得た被覆の考えを用いて定式化する．また，質問 - 文献間の類似測度としての被覆の特定度を導入して，その妥当性を文献のランキング結果より調べる．

2 類似性判断

質問および文献はこれらの内容を表すキーワードの組によって表現されているとする．キーワード間の関連性は，概念的により一般的なキーワードがより特定的なものより下のレベルになるよう階層的に配置されたシソーラスやカテゴリ体系等 (まとめて概念階層と呼ぶ) によって捉えられる．このとき，キーワードの概念的関連を反映した質問 - 文献間の類似性判断を以下のような 2 段階の手続きによって行なう．

S1: 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る．

S2: 各被覆について，その特定度を計算し質問と文献との類似性の指標とする．

ここで，質問と文献の表現の被覆は，両者に含まれるキーワードをまとめて言及する一般的なキーワードの組である．このような一般的なキーワードを求めることをキーワードの一般化という．概念階層を通じて見たとき，概念的に関連の深いキーワードについての一般化結果は，より特定的になる [1, 3]．このことに注目し，被覆を質問側の個々のキーワードと文献中で最も関連深いキーワードとの一般化結果と，これとは逆に文献側の個々のキーワードと質問中で最も関連深いキーワードとの一般化結果を集めることで得る．

質問と各文献との被覆が得られるとその特定さを数量化した特定度に従って文献をランキングできる．問題は被覆の特定度をどのように求めるかである．

3 特定度

被覆の特定度として，絶対的特定度を与える s_a，相対的特定度を与える s_{gr} と s_{lr} の計 3 つを導入する．

s_a は，概念階層のレベルに従って値を与える [3]．ここでは最下位レベル h のキーワードに 1，root に 0，それ以外のレベル l のキーワードには l/h を与える．そして各被覆のキーワードの値の平均値をその被覆の特定度とする．この測度には概念階層中に新しい語が追加されるとランキングが変ってしまうという欠点がある．

s_{gr} は，被覆中のあるキーワード k_i について，すべての被覆のキーワードの中で，k_i と概念階層上で上下関係や同じキーワードであるといった関係のあるキーワードと比較してどの程度特定的であるかを，次式で与えられる点数 P_i で測る．

$$P_i = \frac{A_i + B_i - C_i}{2A_i}$$

ここで，A_i は k_i と関係のあるキーワードの数，B_i は k_i よりも一般的なキーワードの数，C_i は k_i よりも特定的なキーワードの数とする．そして，この点数を各被覆の個々のキーワードについて求め，平均値をその被覆の特定度とする．

s_{lr} は，すべての被覆の中から 2 つの被覆を取り出しどちらが特定的であるかを判定する．このことをすべての被覆の組み合わせで行ない，各被覆の特定的だと判定された頻度をその被覆の特定度とする．ここで，2 つの被覆で一方のすべてのキーワードが他方のどれかより特定的であり，後者のすべてのキーワードが前者のどれかより一般的であれば前者は後

Relevance Judgements of Query-Document Pairs by Reflecting Conceptual Relevancy between Keywords
Yuuichi KANEKO Makoto NAKASHIMA Tetsuro ITO
Oita University, 700 Dannoharu Oita 870-11, Japan

者より特定的であると判定する．判定できない場合，s_{gr} での点数付けをすべての被覆で行なうのではなく，比較している二つの被覆だけで行ない，平均値の高い方が特定的であると判定する．

4 実験

ここでは文献としてACMの論文220編を用い，概念階層としてACMのCRカテゴリを用いた．CRカテゴリのカテゴリ名を以降CRと呼ぶ．ACMの各論文にはCRカテゴリ中のどのカテゴリに属する論文であるかが複数個のCRで示されているので，これをその論文を表すキーワードであると考え実験を行なった．また，150個の論文を任意に取り出し質問として用いた．

もし各CRの間に概念的な関係がない場合に，ここでの方法が現在広く使われている類似測度(コサイン関数等)による結果と大きく異なっていれば，その代りとしてここでの方法を用いることには疑問が生じる．そこで各CRの間に概念的な関係がない場合の結果をコサイン関数による結果と比較してみた．

コサイン関数による文献のランキング結果と他の測度によるものとがどれだけ違うかを $\sum_{i=1}^{m} |i-r_i| \times w_i / \sum_{i=1}^{m} w_i$ で求めた．この値が小さいほどコサイン関数と同様のランキング結果が得られていることになる．ここで，m はランキングされた文献の総数，i はコサイン関数によるランキング結果でのある文献の順位，j は同じ文献の比較する測度による順位である．w_i は順位 i の重みで $1-\frac{i-1}{m}$ とした．結果を150の質問についての平均で表1に示す．

表1: コサイン関数との違い(±標準誤差)

測度	m=30	m=50
s_a	0.237±0.028	0.223±0.027
s_{gr}	0.617±0.043	0.647±0.045
s_{lr}	0.308±0.032	0.294±0.030

この結果より3つの特定度は類似測度としてコサイン関数と大きな違いはないと言え，各CRの間に概念的な関係があればより良い結果が期待できる．

次に N 個の文献の順位づけに必要なコストについて調べた．s_{gr} と s_{lr} では特定度を求める際にキーワードのペアごとに概念階層を参照する必要があり，この操作に大きなコストを要する．それゆえ，全体的なコストは概念階層を参照する回数で計った．その結果 s_{gr} では N，s_{lr} では N^2 に比例したコストがかかった．

最後に質的な評価として，用意した質問について採用した特定度によってランキング結果が大きく異なるものについて調べた．表2に文献の例と D_1 を質問としたときの3種の特定度での各文献の順位を示した．ここで例えば，H.5.1はH.5.1-Videoの一般的なカテゴリを表す．s_a では，質問と同じCRを含んでいる文献でも，それが一般的な場合低い順位となり(D_3)，特定的な質問のCRと同じものを含んでいればそれだけで高い順位となる(D_2)．s_{gr} では D_4 のように質問中のCR(I.2.10-Motion)と関連のあるCRがない文献は，他に関連のあるCRを含んでいても順位が低くなる．総合的に s_{lr} ではこれらの欠点が緩和された順位づけとなっている．扱う文献数が大きくなると s_{lr} は非常に大きなコストが必要になるが，これには，まず，s_a を用いて必要な数より多めに文献を取り出し，これらを s_{lr} で再ランキングすることが考えられる．

表2: 文献例と各特定度での順位

	文献の表現	文献の順位 s_a	s_{gr}	s_{lr}
D_1	H.3.2, H3.3, I.2.10-Motion, H.5.1-Video	1	1	1
D_2	H.3.0, H.5.1-Video	4	21	24
D_3	H.3.2,H.5.2-Interactionstyles, H.3.6-Largetextarchives,I.4.5	18	4	9
D_4	H.3.2, H.3.3-Retrival, H.5.1-Audio,H.5.1-Video, H.5.1-Animations	2	9	2

5 おわりに

質問-文献間の概念的関連を考慮した類似性判断のための方法を被覆の考えとその特定度を導入して定式化した．今後，実際の検索システムに適用してその有効性を調べていく．

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] Dietterich, T. G., and Michalski, R. S.: A comparative review of selected methods for learning from examples, R.S. Michalski, J.G. Carbonell and Y.M. Mitchell (eds.), Machine Learning, Morgan Kaufmann, Los Altos, CA (1983).

[2] Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T.: The vocabulary problem in human-system communication, Communication of the ACM, Vol.30, No.11, pp.964-971.

[3] Kolodner, J. L.: Case-Based Reasoning, Morgan Kaufmann, San Mateo, CA (1993).

[4] Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley, Reading, MA, 1989.