2P-02 適合可能性の示唆による効率的なブラウジング支援

曲 艶華　佐藤 慶三　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

現在，専門家のみならず一般の利用者まで，インターネットや電子図書館から，様々なサーチエンジンやブラウジングシステムを通じて，情報を入手するのが容易になってきている．一方，情報量の急速な増加と情報要求の多様化につれて，効率を重視して要求に適合する文献を取り出せる方策が望まれている．

ここでは，電子図書館構築に関連して，効率的なブラウジング支援法を提案する．本方法は適合判断済み文献との類似度を基に求めた，未読文献の適合可能性を，ユーザに提示することにより検索効率の上昇を目指す．この有効性については，検索用テストコレクション CACM[2]及び Medlars[3]から得た文献を対象に調べた．

2 ブラウジング支援の考え方

電子図書館構築では，文献集合は仮想空間上に配置され視覚的に表現される．関連性の高い文献が近くになるような文献集合の配置方法が数多く提案されている[1][4]．ユーザは視覚化された空間をブラウジングし，その過程で各文献に対し適合判断を下していく．

このブラウジングの過程を支援する方法を考察する．ブラウジングの過程は動的でどのようなものか前もって分からないが，結果的に見ると線形順で文献がたどられた形になっている．このブラウジング形態を考慮に入れ，適合文献・不適合文献との類似度を基に，次に参照すべき文献，すなわち適合可能性のある文献を，ユーザに示唆するのである．具体的には，未参照文献について，質問との類似度あるいは各適合文献との類似度の最大値が，各不適合文献との類似度の最大値以上ならば，この文献は適合可能性があるとする．ユーザはブラウジング過程中，システムから未読文献の適合可能性の示唆を受けて，これらの文献だけについて適合判断すればよい．結果的に検索効率の上昇と伴にコストの低減を図ることができる．

適合可能性を示唆する流れを図 1 に示す．ここでしきい値 β は，低い類似度をもとに誤った適合可能性を示唆してしまうのを避けるために導入したパラメータである．

適合文献集合・不適合文献集合，特に適合文献集合からの情報が文献の適合可能性を求める上で重要となる．このための処理として，ブラウジングの初期に，まず質問との類似度の高い少数の文献及びこの文献との類似度の高い文献に対して適合判断する．そして，未読文献を配置での並び順に適合判断して行くようにする．

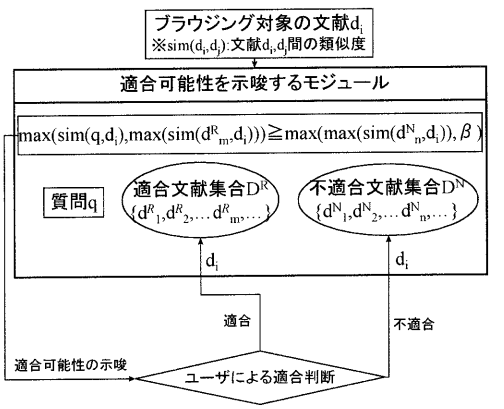

図 1 適合可能性を示唆する流れ

3 ブラウジング法

提案したブラウジング法を WEI (Worth-Examining Information based browsing) 法と名づけて，以下にまとめる：

(S1) 質問との類似度の高い少数の文献 $\{d_i, d_j, \cdots\}$ 及びそれらとの類似度の高い文献について，適合可能性を示唆する流れに従い処理する．次いで，残りの文献について，それらの配置順に適合可能性の示唆する流れに従い処理する．

(S2) S1 で適合可能性の示唆が成されなかった各文献について，質問との類似度の高い順に適合可能性を示唆する流れに従い処理する．

(S3) S2 でも適合可能性の示唆が成されなかった各文献について，質問との類似度の高い順に適合判断する．

Efficient Browsing under the Guidance of Worth-Examining Information
Yanhua QU, Keizo SATO, Makoto NAKASHIMA and Tetsuro ITO
Oita University, 700 Dannoharu Oita 870-1192, Japan

4　評価実験

検索用テストコレクションCACM[2]から1905編の文献を取り出し，これらの文献間の類似度を基にMSTによって類似度の高い文献同士が近くになるように配置した[1]．ブラウジングのための質問として44個を取り出した．

電子図書館の構築並びに，適合可能性の示唆において，文献間（質問も1つの文献とみなす）の類似度が必要となる．文献の内容は，通常キーワードやキーフレーズなどの羅列で表される．ここでは，文献のタイトルとアブストラクトから出現頻度の高い語や句を自動抽出してこれらを得た．文献間の類似度は，キーワード・キーフレーズの文献中の出現頻度にtf×idfで重みをつけ，コサイン関数で求めた[5]．

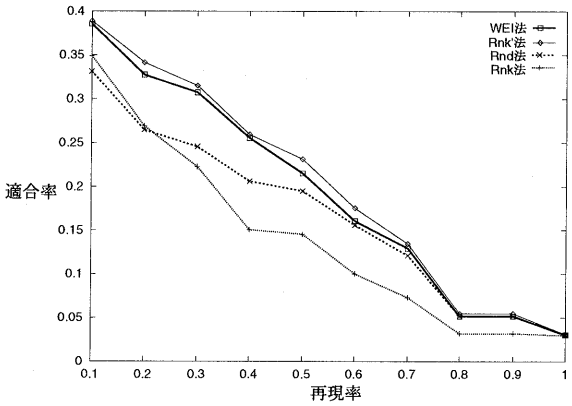

図2　再現率—適合率曲線

検索効率で見た結果を図2に示す（β=0.2）．従来から，検索結果の文献を質問との類似度の高い順にランキングしてブラウジングするRnk法がよく採用される[5]．そのため，先ず，Rnk法とWEI法を比較した．WEI法のS1では，質問との類似度の高い2つ文献を選んだ．再現率が0.1～0.9の場合に，統計的に見てWEI法がRnk法より95%の信頼度で高い適合率を持つことが分かった．

次に，文献の配置順でブラウジングを進めることの妥当性を調べた．最もすぐれた結果を与えるのは，S1で質問との類似度が0より大きい各文献について，ランキング順で適合可能性を示唆する場合と考えられる．この方法をRnk'法と名づけて得た結果を図2に示す．Rnk'法の方が多少優れていたが，それほど大きな差はなかった．

Rnk'法でブラウジングすると，ユーザが仮想空間上で文献の間を大きく移動をしなければならない．一方，WEI法では，文献の配置順に調べればよい．ブラウジングの質と検索効率の両方から見ると，WEI法がよりすぐれたブラウジング法であると言える．

また，上の手続きで，質問との類似度や，文献の配置を考慮しない状況でのブラウジング(Rnd 法を呼ぶ)の結果を調べた．結果を見ると全体的に検索効率が悪くなった．2で述べたように，提案したブラウジング法は適合判断済み文献から適合判断情報を利用するため，ブラウジングの最初の段階で適合文献を探しておく方が検索効率の面で望ましい．そのための処理として，まず質問との類似度の高い文献及びこの文献の隣接配置の文献についてあらかじめ適合判断しておくことが有効であると言える．

同じ実験方法によりMedlars[3]からの1033編の医学文献，30個の質問について，実験を行った．結果はCACM[2]文献を用いた時の実験結果と同様のものとなった．

5　まとめ

電子図書館構築を目的として，視覚化された仮想空間のブラウジングのし易さを考慮に入れながら，適合可能性のある文献だけを示唆する効率的なブラウジング支援法を定式化した．

ここでの方法を利用者主導の連続的なインタラクション的な環境で利用するには，インタフェースの整備が重要になる．今後の課題は，インタフェースの設計，有効性の検証とWWW上での実用化である．

参考文献

[1] C. Chen: Visualizing semantic spaces and author co-citation networks in digital libraries, Information Processing and Management, vol.35, no.3, pp.401-420, 1999.

[2] E. Fox, et.al.: Coefficients of combining concept classes in a collection, Proc. SIGR'88, pp.291-308, 1988.

[3] ftp://ftp.cs.cornell.edu/pub/smart/med

[4] M. A. Hearst, et.al.: Cat-a-Cone: An interactive interface for specifying searches and viewing retrieval results using a large category hierarchy. Proc. of SIGIR'97, pp.246-255, 1997.

[5] G. Salton, et.al.: Introduction to Modern Information Retrieval, McGraw-hill, New York, 1983.