文献の非主題的特徴と情報検索におけるその意義

○鄒 永利
相良 佳弘*

NON-SEMANTIC ATTRIBUTES OF DOCUMENTS AND THEIR IMPLICATION TO IR

○Yongli Zou
Yoshihiro Sagara**

Abstract: Indexing and searching mechanism based mainly upon semantic attributes of documents has been showing certain limitations, and the function of the non-semantic attributes and their implication to information retrieval and system design are not yet fully explored. By summarizing the non-semantic characteristics of documents and by analyzing them in terms of information needs, information seeking and use as well, the authors attempt to explore their potentialities to information retrieval and to information system design as well.

１．問題意識－主題検索の限界

従来の情報検索に関する研究は、主に主題検索を対象としてきた。主題検索では、検索したい概念を示す主題的キーワードやそれらキーワードの組み合わせを検索式として用いてきた。しかし、この主題的キーワードを中心とするアプローチには、いくつかの問題点が指摘されている。まず、文献の主題を表すキーワードと利用者の情報要求の主題を表す検索質問との照合の問題である。これはある問題状況に際して、二つの異なる知識状態（情報の発信者と問題を抱える情報利用者）を照合することによる問題である。さらに、表現上の問題点が二つある。文献の主題とそれを表現する索引語との間の意味的距離に関する問題と、利用者の情報要求とそれを表す検索語および検索式との間の意味的距離の問題である。これらの問題から以下の六つの点が指摘できる。

1)文献の意味あるいはアバウトネスが客観的な存在であるとしても、その理解、表現あるいは抽出を客観的に行うことはできない。
2)文献中の用語を情報を表現する手段とするのは、必ずしも妥当ではない。
3)主題的アプローチによって形成された索引概念は時代に特有なもので、永続的なものではない。
4)言語の曖昧さと豊富さのため、同じ語を持っている文献であっても、同じ概念を必ずしも表さない。
5)検索語あるいはそれらの組み合わせだけで、利用者の具体的な検索質問を完全に表現できるとはいえない。
6)利用者は、探している文献の属する主題カテゴリーを知っていること、あるいは情報要求を表す検索語を選べることを前提としている。

つまり、現在の情報検索は主に領域知識に基づき、単純な主題的類似性（semantic similarities）を用いているため、問題状況を完全に把握または表現できない利用者への対応には不十分だといえる。これまで情報検索研究において、文献に関する主題以外の特徴やデータのような主題的キーワ

* 慶應義塾大学大学院文学研究科図書館情報学専攻博士課程
** Graduate School or Library and Information Science, Keio University

ード以外の検索項目は、キーワードによる主題検索を単に補完するものとされてきた。そのため、この非主題的情報についての研究は十分ではなかった。主題以外の特徴やデータといった情報を活用することによって、情報検索における上述した問題の解決に貢献できると考えられる。

2. 文献の非主題的特徴

文献には主題的特徴や性質だけでなく、非主題的特徴もある。文献の非主題的特徴とは、文献の主題と直接に関連していない、つまり索引や検索の際に主題的キーワードで表現されない特徴である。こうした特徴は、外部からある程度客観的に観察できる特徴とも言える。図1は、文献の非主題的特徴を一部示したものである。それらの特徴は、直接/内的と間接/外的といった二つのカテゴリに分けること

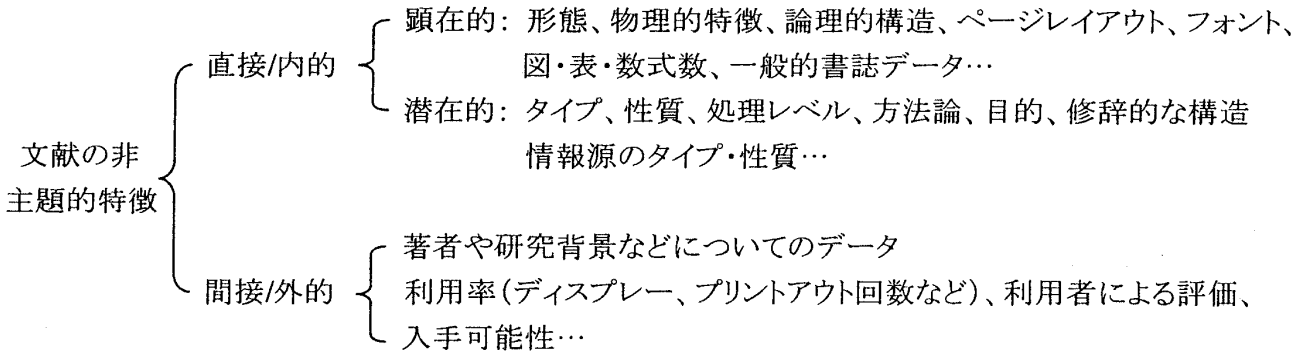

図1. 文献の非主題的特徴

ができる。直接あるいは内的特徴は文献自体が持っている特徴であり、間接あるいは外的特徴は文献自体ではなく外部から与えられる特徴である。さらに直接/内的特徴には文献の形態(媒体、類型等)、物理的特徴(大きさ、長さ、規格、放送時間等)、論理的構成(序、先行研究、方法論、結論等)のような顕在的な特徴とともに、文献のタイプ(レビュー、論文、試験報告、教科書など)、性質(既存理論・方法の評価・証明・応用、あるいは新たな理論・方法の提出など)、方法論(定性的あるいは定性的)など潜在的に存在する特徴もある。

非主題的特徴は、主に以下の四つの過程において著者、編集/出版者、図書館・情報専門家および利用者によって文献に付与されるものである。

①文献の創作過程:論理及び修辞的構成、文献のタイプ、図表など
②編集/出版過程:フォント、形態、ページレイアウト、出版データなど
③図書館/情報サービスでの加工、処理過程:主に一般的書誌データなど
④文献の利用過程:利用率、評価など

①と②の過程によるものは 文献が出版される段階で与えられるものである。③と④の過程からの特徴は、特定の図書館・情報サービスおよび情報システムの方針、利用者の利用によって生まれる。このように一つの文献に対して、様々な非主題的情報が存在していることになる。これら文献の非主題的情報は、文献の主題の表現、理解あるいは文献へのアクセス、あるいは利用のために付与されたものであるので、主題とは潜在的に関連と考えられる。つまり主題の視点から文献に近づく際に、文献における非主題的特徴も有用であると言える。これらの特徴が、従来の情報検索研究においてどのように扱われてきたか、また実際にどのような役割を果たしているかを次章で検討する。

3. 非主題的特徴の意義

これまでの情報要求や情報探索行動及び情報検索に関する研究においても、文献の非主題的特徴の意義をある程度明らかにしている。これらの先行研究をまとめると、利用者にとって非主題的特徴は、情報要求の理解・把握、適合性判定、情報アクセスポイント、特定情報の同定といった四つの面において重要であるといえる。

①情報要求の理解・把握:情報要求という概念に対する解釈は多様であるが、知識・認知構造における欠陥(世界像における不完全さ(Mackay)、認知的ギャップ(Devin)、知識の変則状態(Belkin))として取り扱うことは広く受け入れられている。主題的次元から見ると情報要求は不完全、曖昧であると一般的にいわれている。一方情報要求には特定的な側面もある。それは非主題的次元である。これを示したのが図 2 である。つまり、一種の知識構造として情報要求には、主題的次元と非主題的次元がある。 主題的次元は必要となっている情報の主題、専門領域、分野についての知識であるが、利

情報要求
- 主題的次元: 必要となる情報の主題と領域 ➔ 知識・認知的欠陥(不完全さ、認知ギャップ、ASK) ➔ 利用者が情報要求を明瞭に適切な言葉で表現できない ➔ 曖昧的
- 非主題的次元: 問題状況の解消に役立ちかつ利用者が理解できる情報のタイプ、性質、レベル、形態など ➔ 特定、具体、個別的

図2. 情報要求の二面性

用者の知識・認知的欠陥の部分でもあるので、曖昧な認識、表現にならざるを得ない。一方非主題的次元は、情報要求を生み出す客観かつ具体的な課題に必要となる情報のタイプ、質、レベル、形態などについての知識であるので、個別的かつ具体的なことになる。利用者が認識可能か否かは別として、何らかの情報が必要であるという状態を解消する適合情報は、主題だけでなく主題以外の他の条件をもそろっている具体的なものだと考えられる。したがって、情報あるいは文献の非主題的特徴から適合情報に近づき、情報要求の主題的次元を理解、把握していくのは有効な手法であると考えられる。

②適合性判定:図1に示す文献の非主題的特徴のうち、文献のタイプ、性質、処理レベル、形態、言語、新鮮さ、あるいは著者の立場、採用した方法論、アプローチなどの要素は、適合性判定において重要な役割を演じている。適合性判定の基準に関する Barry[1]、Park[2]、Cool[3]、Howard[4]、Ellis[5]、Wang[6]等の研究で示された適合性判定に使用される非主題的特徴の例として、採用した方法論、文献のレベル、質、タイプ、研究の背景、発行元、文献の長さ、媒体、著者の知名度、地位、職業、学位、所属、文献で示されている態度あるいは価値判断と分野におけるパラダイムの受容性、具体的な理論、方法を提出したか、信頼度、執筆者の目的、入手可能性などがある。適合度の順位付けに関する研究においても、利用者が検索結果から文献の適合度を判定する際に、文献の非主題的特徴が何らかの影響を及ぼしていることが示されている[7]。

③情報アクセスポイント:文献の非主題的特徴は、情報へのアクセスポイントとして重要な役割を演じている。Ellis がまとめた情報探索の総合的モデルにおいて、情報探索行動は検索開始、連鎖、ブラウジング、フィルタリング、追跡・監視(monitoring)、抽出と終了から構成されている。その中に文献の非主題的特徴をアクセスポイントとする探索行動が極めて多い。検索開始は、特定な著者あるいは文献のタイトルを手がかりに行われることも多く、連鎖、ブラウジング、追跡・監視、抽出は、既知の文献、著

者、情報源(雑誌、目録、コラムなど)を前提として行われる[8][9]。つまり基本的な情報探索行動における利用者は、情報要求を主題的に定義して情報を探し始めるより、むしろ文献の非主題的特徴から情報を探しながら情報要求を明確にしていくとしている。利用者の情報へのアクセス形態を 6 種類にまとめた Bates の'いちご摘み'(Berry picking)モデルにも、非主題的なアクセス形態が主であると示されている[10]。Batesは、我々が情報にアクセスする際のアプローチには、以下のものがあると述べている。文献の引用文献に着目する、引用索引を利用する、中心的な雑誌に収録されている論文に着目する、特定の文献と場所的に近接する他の文献群に着目する、書誌や索引・抄録を利用する、特定の著者の文献に着目するといった六つのアプローチである。これらのほとんどのアクセス形態は、非主題的特徴に基づくものである。

④特定情報の同定:非主題的特徴は、文献内で必要な情報の所在を特定するツールとしても利用可能である。利用者は自分の情報要求や興味に応じて、文献全体を利用するのではなく、文献の特定部分しか利用しない。この特定部分を選択する際に非主題的特徴が重要な役割を演じている。特に学術情報の利用者が一つの文献を初めから終わりまで読む例はほとんどなく、文献の論理的構成やレイアウトなどから、文献の非常に限定された特定部分しか利用しないことを Kircs の調査が示している[11]。また、情報要求を問題解決の視点から調査した Allen の研究も、利用者は異なる段階(問題識別、解決策の確認、解決策の選択)での情報要求が、文献の異なった特定部分に含まれている情報を必要とすることを示している[12]。例えば、解決策の確認段階では、文献の先行研究部分に含まれている情報、解決策の選択段階では評価や結論の部分の情報が、もっとも役立つとされている。つまり利用者が必要とする特定の情報を見つけたり特定したりする際には、主題からというよりもむしろ論理的構造やレイアウトのような文献の非主題的特徴の方を、よく利用していると言える。

以上のことから情報検索システムが利用者が望む情報を提供するためには、従来の主題的キーワードによる検索だけでは不十分であり、非主題的特徴を検索に反映していくことが必要であろう。

4. 情報検索における非主題的特徴の利用の可能性

今日の主題志向の情報検索システムは、利用者の情報要求を十分に反映させた検索結果を出力するには至っていない。そこで、現在検索の場では利用されていない文献の非主題的特徴を利用することによって、より利用者の情報要求を反映させた検索を行うことが可能になると考えられる。非主題的特徴を情報検索に用いるにあたって、以下のような利用が考えられる。

①文献には、序、先行研究、仮説、方法、研究デザイン、試験、推論、結論、注釈、引用といった論理的構造が存在するが、これを検索に利用することが可能である。前述のように利用者は文献全体ではなく、特定の部分を利用する場合が多い。そこで文献の論理的構造を、利用者が必要としている特定部分に着目する手段として用いることができる。また、文献の扱っている分野や研究領域、研究手法によっては、文献が特定の論理的構造を持つことがある。この論理的構造から、特定の分野の研究または研究手法を用いた文献を検索することも可能となるだろう。

文献の論理的構造はキーワード検索の性能向上の手段としても利用できる。キーワードは、文献中での出現位置によって文献主題との関連性が異なると考えられる。複数のキーワードを用いて検索を行う場合、それらキーワード群の出現分布からキーワード間の関係を推測することが可能になる。そのため、文献の論理的構造を把握し、各キーワードがどの構造部分に出現しているかを重み付け等に利用することによって、検索効率の向上を図ることが可能になる。このように、キーワードが文献中のどの位置に、またどのように分布で出現したかを、検索に応用することは効果があると考えられる。M.A.Spasser の引用関係に関する研究では、被引用文献の引用文献での出現位置が両文献間の主題の関連性の強さや、引用文献の被引用文献に対する評価に影響することを示された[13]。

②文献における特定の修辞的な構造、つまり文章構造を検索に利用することも考えられる。文献の典型的な論理的構造は、すべての文献に適用可能なものではない。また、個々の論理ラベルのもとに、文献中のそのタイプの情報がすべて存在するとは限らない。しかし修辞的構造(rhetorical structure)は潜在的な構造として、どんな文献にも存在する。特に学術文献の修辞的構造は、論理的構造と同様に相当標準化されてきており、情報検索に利用可能である。学術論文における修辞的構造に関する研究は、論文のテキスト(句あるいは段落単位)を論理的にカテゴライズ(仮説、試験環境、従来の理論・方法、著者の提出した理論・方法、結論など)できると示している。もしこうしたカテゴリに基づいて文献を索引することが可能になれば、特定種類の情報の検索は可能になるだろう。例えば、「著者の提出した理論・方法」というカテゴリのところで「our solution to the problem」、「.....advocated here」、「contrary to conventional wisdom」、「in this paper.....different approach」のような修辞的構造と関わっているすべてのテキストが検索されることとなる。Kircz は、学術論文における修辞的構造をカテゴリ化した上で、残された問題は各々のカテゴリに属する様々な具体的な表現を同定することであると述べている(14)。

③文献の非主題的特徴はフィルターとして文献の性質やタイプについての判断に用いることもできる。現在の商用情報検索システムにおいて、文献の性質·タイプの判断に利用されているフィルターは、情報源(雑誌のタイトル)、言語、レビューといった非主題的特徴のうちの一部にすぎない。著者や文献に関する非主題的特徴は、文献の性質·タイプについての判断に十分には利用されているとは言えない。例えば、検索された文献集合を更に限定するために、分析された理論/方法に対する立場·価値判断(肯定/否定)、ある理論/方法についての証明、応用、評価、一つの新たな理論/方法の提出、研究手法(定量的/定性的)、目的(理論志向/実践志向)、研究レベルなどの性質は、有効なフィルターとなり得る。文献のこのような性質は、索引作業者の判断によって決定することも、また図、表、数式数、引用文献数、利用率、および著者に関する情報(団体、職業、職位、最後学歴)などの非主題的特徴を使用して機械的に決定することも可能である。

最後に、情報要求を代表する典型的な適合文献と検索対象となる文献間の主題の類似性を比較する際に、非主題的特徴に見られる類似性を相互比較をすることも考えられる。Oddy の THOMAS システムのような典型的な文献を基準に、これと類似している文献を探すメカニズムでは、キーワードの比較、つまり、単なる主題的次元での比較しか行っていない(15)。もし、それに前述した文献の性質·タイプなどに関する非主題的特徴の比較を組み込むことができれば、THOMAS のようなシステムには、典型的な適合文献との類似性が高い文献を得ることができるだけでなく、利用者の情報要求への対応性もさらに高まると考えられる。

終わりに

ここまで、文献の非主題的特徴とその情報検索における意義を検討してきた。非主題的特徴の利用は、検索式と検索対象との主題的類似度によって検索を行うシステムにとってより効果が大きいと考えられる。検索式に「文献の性質」、「文献の利用率」、「情報源の信頼性」、「論理的構成」など利用者が主題的キーワードとして十分に表現できなかった非主題的要素を取り入れることにより、本来利用者が望んでいた情報により近い検索式とすることが可能になる。特に順位付け検索においては、このような非主題情報まで含めて順位付け判定を行わなければ、利用者の望む順位をシステムが再現することは困難であろう。今後、主題検索において主題分析を行うことの重要性は変わらないが、主題以外の要素にも着目し、文献の非主題的特徴を利用することによって検索性能を向上させていくことが、必要となるだろう。

References

(1) Barry, C. L. : User-defined relevance criteria: An exploratory study. Journal of the American Society for Information Science, 1994, 45(3):149-59

(2) Park, T. K. : The nature of relevance in information retrieval: An empirical study. Ph. D. thesis, Indiana University, 1992

(3) Cool, C. Belkin, N. Kantor, P., Frieder, O. : Characteristics of texts affecting relevance

(4) Howard, D. L. : Pertinence as reflected in personal constructs. Journal of the American Society for Information Science, 1994, 45(3):172-85

(5) Ellis, D. : A behavioral approach to information retrieval system design. Journal of Documentation, 1989, 45(3):171-212

(6) Wang, P. L., Soergel, D. : A cognitive model of document use during a research project: Study 1. Document selection. Journal of the American Society for Information Science, 1998, 49(2):115-33

(7) 相良 佳弘：適合度順検索における順位付けの問題点とその方向. 慶應大学大学院修士論文、1998、3

(8) Ellis, D. : Modelling the information-seeking patterns of academic researchers: A grounded theory approach. Library Quarterly, 1993(63), 469-86

(9) Ellis, D., Haugan, M. : Modelling the information seeking patterns of engineers and research scientists in an industrial environment. Journal of Documentation, 1997, 53(4), 384-403

(10) Bates, M. : The design of browsing and berry-picking techniques for the online search interface. Online Review, 1989, 13(5):407-24

(11) Kircz, J. G. : Rhetorical structure of scientific articles: The case for argumentational analysis in information retrieval. Journal of Documentation, 1991, 47(4): 354-72

(12) Allen , B. L. : Information Task: Toward A User-centered Approach to Information Systems. San Diego: Academic Press, 1996

(13) Spasser, M. A. :The enacted fate of Undiscovered Public Knowledge, Journal of the American Society for Information Science, 1997, 48(8):707-717

(14) Kircz, J. G. : Rhetorical structure of scientific articles: The case for argumentational analysis in information retrieval. Journal of Documentation, 1991, 47(4): 354-72

(15) Ellis, D. : Progress & Problems in Information Retrieval. London: Library Association Publishing, 1996