レレバンス評価と情報検索の成功

安形 輝(亜細亜大学)

情報検索分野における主要な評価の枠組みであるクランフィールド型検索実験では主題的なレレバンス判定に基づき再現率・精度による評価を行ってきた。前提条件を設定した上で、ある文献のレレバントな度合い、特に、ある文献はどこまでレレバントになりうるかについて考察する。ある文献が支配的なほど強くレレバントになりうるならば、そのような文献を検索で漏らすことは検索の致命的な失敗となる。

1. 検索実験に対する従来の批判

従来、クランフィールド型検索実験に対する批判は数多く行われてきたが [1]、それらは主として以下の二つの点から行われてきた。

(1) 主題的なレレバンス

クランフィールド型検索実験における主題的なレレバンスは便宜的に現実の情報検索過程では評価の判断主体である利用者を無視している。そこで、認知的アプローチをとる情報検索研究者は、主題的なレレバンスが利用者志向のレレバンスから乖離していることを数多くの研究において指摘してきた [2]。しかし、認知的なアプローチでの研究成果は、検索実験という枠組みに応用が難しく、結果としてクランフィールド型検索実験の枠組みに与えてきた影響は非常に少ない。

(2) 再現率・精度

再現率と精度に対する批判はさらに主として二点から行われてきた。

①再現率を正確に算出するためにはテスト集合中すべての文献について各検索質問に対するレレバンス判定を行っておく必要がある

②再現率と精度は2値のレレバントに基づくため順位付け出力を適切に評価できない

①があるために、大規模な集合、物理的制約から数千件を超えるテスト集合は検索実験において扱うことが現実的ではなかった。しかし、TREC などの大規模集合を使った検索実験プロジェクト [3]では、参加システムの検索結果をマージしたものを全数と見なすプーリングという手法によって再現率の算出を行う形で検索実験の枠組みを拡張している。

本研究では、TREC のような大規模検索実験プロジェクトで用いられているクランフィールド型検索実験の枠組みに内在する危険性を明らかにすることを目的とする。そして、可能な範囲でクランフィールド型検索実験の枠組みのなかで、検討を行っていく。

2. 前提条件

本研究での考察の前提として以下のような条件を設定する。

(a) レレバンス評価は文献単位で行われる

情報検索が扱う対象にはさまざまなものがあるが、ここではその中でも文献のみを扱う。なおかつ、複数の文献から構成される集合に対するレレバンス評価も指摘されているが [4]、便宜上、一つの文献単位で扱うものとする。

(b) レレバンス評価は主題的なレレバンス概念に基づく

これは、実験環境下でコントロールが難しい利用者を除きたいクランフィールド型検索実験では最初に設定される条件である。

主題的なレレバンス概念は認知的なアプローチの研究からは数多く批判されてきた。

しかし、その定義について明確に扱ったものは少ない。初期の索引実験においてCuadra と Katter が行った定義は「レレバンスは情報要求の記述と論文との間の文脈上の対応である。すなわち、その論文が情報要求の記述に対して適切な素材を取り扱っている程度である。」[5]となっている。しかし、実際の実験においてはより具体的な扱いが規定されることもある。例えば、TREC では「判定者がレポートを記述するさいにある文献中に含まれる情報を少しでも利用するならばその文献はレレバントと見なす」となっている。

ここでは、主題的なレレバンスは「検索質問と文献の間の主題的な関連性」と定義し、レレバンスの評価は「文献に含まれる検索質問で表現された主題に関連する情報の量」と評価を行う上での主題の定義と範囲を定める「その主題を扱うため情報」から行われるものとする。

(c) レレバンス評価は検索質問に文献がどのくらいレレバントかで行われる

これは、条件(b)から検索質問情報を主題的に関連性が高いものほど、よりレレバントであることを示すものである。クランフィールド型検索実験の要となる再現率と精度は、レレバントであるかレレバントでないかに基づき算出されるが、レレバンスに関する多くの研究では「レレバンスは多値的であり、その程度が問題であって、単純にはい／いいえで決定されるものではない」[2]とされている。

(d) レレバンス評価のさいに各文献は独立したものとして扱う

クランフィールド型検索実験では、各文献を独立したものとして扱い、ある文献のレレバンス評価を行う場合には、他の文献は何ら影響を与えないものとする。しかし、Goffman はレレバンスが伝達される情報の尺度として定義されるならば、文献だけでなく、文献間の相互関係に関しても定義される必要があるとしている[6]。また、Eisenberg による文献の提示順序がレレバンス評価に影響を与えているとする研究もある[7]。本研究では、まずこの条件の範囲内で考察した場合、この条件をはずした場合の両面からの考察を行う。

(e) 検索質問、文献ともに言語的に表現される

言語の表現能力が完全でないとするならば、検索質問や文献を表現する時点で言語によるずれが生じることになる。つまり、この条件下では、最初から完全なる情報検索は実現が難しいことになる。

3. レレバンス評価に関する検討

3.1 文献はどこまでレレバントになりえるか

多くの情報検索研究におけるレレバンス評価では「非常にレレバント（Highly Relevant）」などの最大値を設定している[8]。ここでは最大の値を設定せずに、任意の一文献がどこまでレレバントになりうるかについて、検討を行う。

① 絶対的なレレバンス評価

上記の条件(d)から絶対的にレレバンス評価が行われる場合にも、主題的なレレバンス評価が(b)で決めたように文献中に含まれる関連する情報の量から決まるならば、情報の量の上限が決まらない限り、レレバンス評価の上限も決まらないことになる。

② 相対的なレレバンス評価

条件(d)をはずし、文献間の相互関係によってレレバンス評価の値、そして上限が決まるならば、任意の二つの文献があった場合に、その比較からどちらかの文献のレレバンス評価が決まることになる。つまり、どの文献があったとしても、ある文献よりも関連する情報量が多い文献を想定すればよりレレバンス評価の値が大きな文献をいくらでも想定可能である。

また、同様に条件(d)をはずした環境下に

おいて、「主題を扱うための情報」に影響を与える文献があれば、その存在は検索質問や文献の主題自体に変化をもたらし、結果として他の文献のレレバンス評価を変化させることを想定できる。

3.2 二つのレベル

以上の検討を踏まえて、ある文献のレレバンス評価が他の文献のレレバンス評価に対して致命的な二つのレベルを考える。

(i) 支配的なレベル

支配的なレベルでレレバントな文献とは、「レレバントな度合いが、他のレレバントな文献のレレバントな度合いの合計以上であるほどレレバントである文献」である。

例えば、検索質問「夏目漱石の書いた小説にはどのようなものがあるか？」があった場合、夏目漱石が書いた個々の小説のレレバントな度合いよりも、個人書誌のレレバントな度合いは非常に大きいはずである。

(ii) 破壊的なレベル

破壊的なレベルでレレバントな文献とは「他のレレバントな文献をレレバントでないものにしてしまう文献」である。

例えば、検索質問「中田英寿選手の出場する 10 月期の試合日程はどうなっているか？」があった場合に、個々の試合日程それぞれはレレバントな文献であるが、「中田英寿が足の故障のため引退」という文献があった時点で他の文献はレレバントでなくなってしまう。

4. 検索実験に内在する危険性

4.1 古典的な検索評価

情報検索の目的がよりレレバントな文献をより多く検索することだけだとすれば、それに応えることは、文献をすべて出力すること(検索しないこと)で用意に実現できる。この全体集合に含まれるレレバントな文献をすべてチェックすることは以前の検索実験で行われていた全数チェックと同じである。実際の情報検索では利用者が効率的に情報を利用できるようにさまざまなアルゴリズムを駆使し、よりレレバントらしい文献を取捨選択することで全体集合よりも非常に数の少ない検索結果を出力する。古典的な検索評価は、全数チェックによるレレバントな文献と検索結果中のレレバントな文献を比較することで行われてきた。このような環境下では、任意の文献のレレバントな度合いが支配的なレベルにあろうと破壊的なレベルにあろうと必ずチェックされるため問題にはならない。

4.2 プーリング手法とその危険性

TREC などの現代的な検索実験プロジェクトでは、クランフィールド型検索実験の枠組みのなかで、大規模テスト集合を扱うために、全数チェックを行わずプーリング手法を用いている。しかし、プーリング手法ではどの検索システムでも検索されなかった部分に含まれる文献は自動的にレレバントでない文献と見なされ、チェックの対象から外される。そして、検索されなかった部分にはどのような文献があるかがわからないという古くからの暗黒物質問題を再び抱えてしまうことになった。プーリングされた集合は実際には現在の検索技術を反映したものである。それのみに依存する評価は、現在の検索技術の枠組み内での局部的な最適化につながる危険性をはらんでいる。TREC の関係者もその危険性については認識しており、プーリング手法の妥当性の検証や手法の改善に関する研究が行われてきた。Voorhees による研究では複数回にわたるデータを比較し、妥当性の検証を行っており、Zobel による研究は TREC の単純なプーリング手法では多くのレレバントな文献が見逃されていることが示唆し、その改善の提案がされている。

しかし、このような検証や改善は 2 値のレレバンス評価やレレバンス評価にはある一定の上限が設定されていることを前提としている。上述のようにある文献のレレバ

ントな度合いが支配的なレベルあるいは破壊的なレベルを考えられるとするならば、その一文献を見つけるか否かは検索の成功に大きく影響を与えることになる。

テスト集合がより大規模になっていく現在、1 文献も漏らさないことを統計的に検証することは困難である。このような条件を設定した下では、現在のプーリング手法による最適化が局部的な最適化につながる恐れがないとはいえない。

5. 危険性回避の提案

(1) レレバンス評価は 2 値で行う

ある文献がレレバントであるかないかという 2 値でレレバンス評価を行うならば、ある一文献が検索において致命的な存在になることはない。しかし、①情報検索過程におけるレレバンス評価がどのように行われるか、②情報検索システムの順位付け出力との関係、を考慮した場合現実からより乖離する方向に戻ることが妥当かは疑問である。

(2) レレバンスでない文献に基づく評価

レレバンス評価に対する実証的な研究 [8] の結果をみるとレレバントな文献よりレレバントでない文献に対しての方が評価者間の一致度が高いという傾向が見られる。これからレレバントでない文献に基づく評価の方がより信頼性が高くなることが考えられる。

また、利用者を想定すれば、あるレレバントでない文献が与える損失は、その文献の吟味に対して使われる労力、時間として見なすことができる。そのようなコストがある一定の範囲に収まると仮定するならば、やはり、現在のようにレレバントな文献に基づく評価よりもレレバントでない文献に基づく評価のほうが、より信頼性が高い選択肢と考えられる。

実際に古典的な尺度の一つである Cooper による平均検索長（Expected Search Length）[11]はそのような考えに基づく尺度であり、順位付け出力に関する修正を行うことで大規模な検索実験に応用可能であると考えられる。

【引用文献】

1) Ellis, David. "The Dilemma of Measurement in Information Retrieval Research". Journal of the American Society for Information Science, Vol.47, No.1, p.23-36(1996)
2) Saracevic, T. “Relevance: A review of and a framework for the thinking on the notion in information science”. Journal of the American Society for Information Science. Vol.26, No.6, p.321-343(1975)
3) Voorhees, E.; Harman, D. “Overview of the Ninth Text REtrieval Conference (TREC-9)”. TREC-9 Proceedings. National Institute of Standards and Technology Special Publication. p.1-23(2000)
4) Tiamiyu, M.A.; Ajiferuke, I.Y. “A Total relevance and document interaction effects models for evaluation of information retrieval process”. Information Processing and Management. Vol. 24, No.5, p.391-404(1988)
5) Cuadra, C.A.; Katter, R.V. “Opening the black box of relevance”. Journal of Documentation. Vol.23, No.4, p.291-303(1967)
[日本語訳は情報学基本論文集 II. p.138 より]
6) Goffman, W. “On relevance as a measure”. Information Storage and Retrieval. Vol.2, No.3, p.201-203(1964) [ゴフマン,W.著. 岸田和明訳 “5. レレバンス尺度について” 情報学基本論文集 II. p.111-115(1998)]
7) Eisenberg, M., Barry, C. “Order effects: A study of the possible influence of presentation order on user judgment of document relevance”. Journal of the American Society for Information Science Vol.52, No.2, p.293-300(1988)
8) Spink, A.; Greisdorf, H. “Regions and levels : Measuring and mapping user’s relevance judgments”. Journal of the American Society for Information Science and Technology. Vol.52, No.2, p.161-173(2001)
9) Voorhees, E M. “Variations in relevance judgments and the measurement of retrieval effectiveness”. Information Processing and Management, No.36, p.697-716(2001)
10) Keenan, S.; Smeaton, A.F.; Keogh, G. “The effect of pool depth on system evaluation in TREC”. Journal of the American Society for Information Science and Technology. Vol.52, No.7, p.570-574(2001)
11) Cooper, William S. “Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems”. American Documentation, Vol.19, No.1, p.30-41(1968)