レレバンス判定に対する文献提示順と判定尺度の影響

安形輝（亜細亜大学） agata@asia-u.ac.jp

【抄録】検索結果の提示順と判定尺度の選択のそれぞれがレレバンス判定に影響することは 1980 年代の既往研究において指摘されていたが、その後の十分な検証が行われていない。本研究ではウェブ上のシステムを用いてインターネット環境下で実験を行い、提示順と判定尺度が個別にレレバンス判定結果に与える影響について検討した。また、判定尺度と提示順を関連付けた分析を行った。

1. はじめに

検索実験ではレレバンス判定に基づき性能評価が行われており、レレバンス判定は欠かすことができない重要な構成要素の一つといえる。レレバンス判定に影響する要素として判定尺度の選択や文献提示順を指摘した研究が 1980 年代にいくつか行われた。

1.1 判定尺度のレレバンス判定への影響

判定尺度のレレバンス判定への影響についての研究の多くは、「レレバント」「非レレバント」の2段階あるいはそれを細かく分けた3／5／7／9段階に分ける伝統的な段階尺度に対する批判として行われてきた。

Eisenberg[1]は、段階尺度とマグニチュード推定法（上限、下限を定めずに与えられた刺激を任意の数値等で表現させる尺度）を用いたレレバンス判定実験を行い、比較した。被験者 12 人が情報要求と書誌情報の間のレレバンスを判定する実験を行い、レレバンス判定尺度には段階尺度よりも自由度の高いマグニチュード推定法が適当であるとする結果を出している。また、Bruce[2]も情報探索過程におけるレレバンス評価の推移を調査するためにマグニチュード推定法を用いた。この研究でも、Eisenberg と同様に、マグニチュード推定法を支持する結果となった。

1.2 文献提示順のレレバンス判定への影響

人が何らかの判断を行うときに、その(刺激の)提示順によって順序効果があることは、さまざまな分野[3]で指摘されてきた。

情報検索の領域においても、検索結果の文献提示順がレレバンス判定に影響を与えることが、すでに 1988 年に Eisenberg ら[4]によって指摘されている。レレバンス評価の高い文献から低い文献へと並べた順序と、その逆順の2通りで提示されたレレバンス判定を比較している。調査の結果は、レレバンス評価の高いものから低いものへという順序で提示すると、被験者はレレバンスの高い文献を過小評価し、逆順の場合、レレバンスの低い文献を過大評価するとなっている。そして、文献は無作為に提示した方がよく、判定尺度はマグニチュード推定法を用いた方が提示順の影響を受けにくいと結論付けた。また、Parker ら[5]によっても同様の研究が行われている。彼らは判定する文献数が 15 件以下であれば、文献提示順はレレバンス判定に影響を与えないとしている。

1.3 本研究の目的

以上のようにレレバンス判定に関する研究では判定尺度や文献提示順が判定に影響するとされてきた。しかし、1990 年代以降レレバンス判定に関する研究が質的なもの[6]へ移行するにつれ、それ以降の検証は十分には行われていない。また、インターネットの普及に伴い人々の情報利用行動が変わったとするならば、インターネット環境下での検索を前提としたレレバンス判定に関して、判定尺度の選択や文献提示順からの検討が必要と考えられる。

そこで、本研究ではウェブ上の実験システムを用い、インターネットでの検索を模したレレバンス判定実験を行い、そのような環境下での判定尺度の選択や文献提示順がレレバンス判定に与える影響について分析を行うことを目的とする。

また、判定尺度については従来、段階尺度とマ

グニチュード推定法の二つの比較がなされてきたが、ここではウェブ上のシステムの利点を生かし、スライダーを用いた尺度を追加し、三つの尺度での比較を行った。このスライダーはより詳細な段階尺度とも、制限つきのマグニチュード推定法とも考えられる。さらには、従来あまり行われてこなかった判定尺度と提示順を関連付けた分析も行った。

2. 実験環境

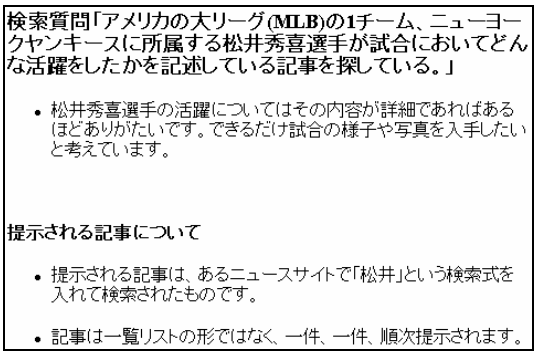
レレバンス判定の実験は 2004 年 9 月と 2005 年 9 月の二回、行われた。二つの実験は、被験者、検索質問数、提示する文献数以外には基本的に同じ条件下で行われた(以下、前者を実験1、後者を実験2とする)。

2.1 被験者

被験者は亜細亜大学夏期司書講習の情報検索演習において「インターネット検索」の回を受講済みの受講生から募集した。実験1は 2004 年度、実験2は 2005 年度の受講生を対象とした。

2.2 検索質問と文献集合

検索質問には利用者の欲しい文献、その背景、検索に使われたキーワードが記述されている。


検索質問「アメリカの大リーグ(MLB)の1チーム、ニューヨークヤンキースに所属する松井秀喜選手が試合においてどんな活躍をしたかを記述している記事を探している。」

- 松井秀喜選手の活躍についてはその内容が詳細であればあるほどありがたいです。できるだけ試合の様子や写真を入手したいと考えています。

提示される記事について

- 提示される記事は、あるニュースサイトで「松井」という検索式を入れて検索されたものです。
- 記事は一覧リストの形ではなく、一件、一件、順次提示されます。

図1 検索質問例

レレバンス判定を行った文献集合は、インターネット上で入手できる新聞記事から構成されている。これらは、実験の前日に Yahoo!ニュース検索[7]、Google News Beta[8]からキーワードで実際に検索されたものを使用している。一部の記事には写真も含まれている。なお、記事内容は改変していないが、含まれるキーワードが判別しやすいよう、黄色く色づけし表示するよう加工した。

実験1は検索質問1問と提示する文献 10 件、実験2は検索質問3問と提示する文献 18 件(6 件×3 問)から構成されている。

2.3 判定尺度

判定尺度としては、以下の三尺度を用意し、検索質問ごとに機械的に割り当てた。

①5段階尺度

多くの検索実験において伝統的に用いられてきたカテゴリの尺度であり、ここでは、レレバンスの低い方から「全く適合せず」「あまり適合せず」「どちらともいえない」「だいたい適合」「非常に適合」までの5段階としたラジオボタンを用いた。

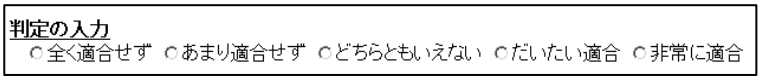

判定の入力
○全く適合せず ○あまり適合せず ○どちらともいえない ○だいたい適合 ○非常に適合

図2 5段階尺度

②スライダー

マウスを使ってハンドルを移動することで、レレバンス判定の値として 0 から 100 までの値を表現することができる。

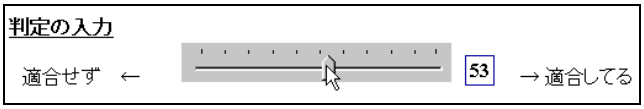

判定の入力
適合せず ← 53 →適合してる

図3 スライダー

③マグニチュード推定法

範囲を決めない数値を記入させるため、制限の少ないテキストボックスを用いた。

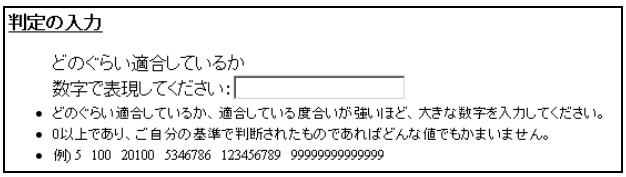

判定の入力
どのぐらい適合しているか
数字で表現してください:
- どのぐらい適合しているか、適合している度合いが強いほど、大きな数字を入力してください。
- 0以上であり、ご自分の基準で判断されたものであればどんな値でもかまいません。
- 例) 5 100 20100 5346786 123456789 99999999999999

図4 マグニチュード推定法

2.4 提示順序

各検索質問について、調査者があらかじめ判定した「レレバント」から「非レレバント」の順に文献の順序を並べ替えておいた。この順序に基づき、被験者の検索質問ごとに、「レレバント」から「非レレバント」、「非レレバント」から「レレバント」、「混合」という三つの提示順序を機械的に割当てた。

2.5 実験システム

本実験のために、ウェブ上からアクセス可能な

実験システムを構築した。このシステムにより、以下のようなことが可能となった。

・同時に多数の被験者に対する実験

・3 つの尺度、3 つの提示順の機械的な割当て

・レレバンス判定にかかる時間の記録

2.6 実験手順

レレバンス判定実験は以下の手順で行われた。

① 被験者がウェブ上の実験システムへアクセス

② レレバンス判定実験の説明

③ 年代、性別等の基本的な属性の入力

④ 検索質問の説明

⑤ 文献ごとにレレバンス判定。実験1では 10 文献、実験2では 6 文献を判定

⑥ 規定の検索質問を終わっていなければ④へ、終わっていれば⑦へ

⑦ 実験への協力の謝辞の表示

3. 実験結果

3.1 基本的な属性と判定尺度と提示順の割当て

被験者の基本的な属性を表1に示した。

表1 被験者の属性

		実験1			実験2	
		性別			性別	
		女性	男性	不明	女性	男性
年代	20	72	26	0	64	26
	30	27	5	1	15	6
	40	14	7	0	8	3
	50	7	4	0	3	1
小計		120	42	1	90	36
総計		163			126	

単位:人数

表2 提示順、尺度別査被験者数

実験1		提示順			合計
		レ⇒非	非⇒レ	混合	
尺度	5段階	18	17	18	53
	スライダー	17	19	19	55
	マグニチュード	18	18	19	55
合計		53	54	56	163

単位:人数

実験2		提示順			合計
		レ⇒非	非⇒レ	混合	
尺度	5段階	45	44	42	131
	スライダー	43	40	41	124
	マグニチュード	40	41	39	120
合計		128	125	122	375

単位:質問ごとの延べ人数

各被験者に検索質問と文献集合が提示されるときに、実験システムは無作為に提示順、尺度を割り当てる。その分布は表2のようになった。この表で、「レ⇒非」「非⇒レ」は文献の提示順を示しており、前者はレレバントな文献から非レレバントな文献を提示した場合、後者はその逆順に提示した場合となっている。実験2は検索質問と文献集合が3組あり、被験者数は質問ごとに集計したため、延べ人数となっている。

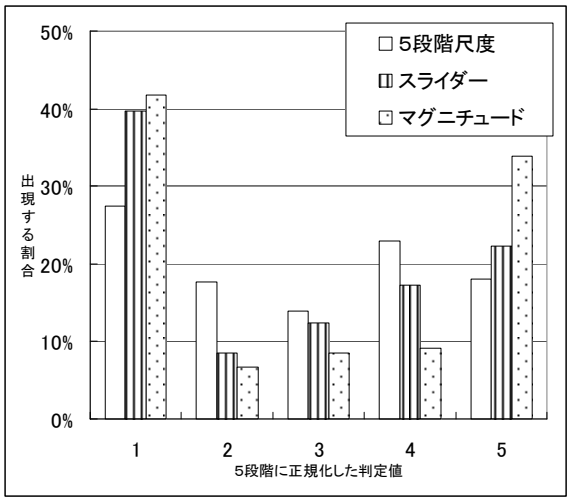


図5 判定尺度ごとの判定値の分布

3.2 判定尺度による影響

判定尺度同士の比較をするために、最大値と最小値を使い、各尺度の判定値を、0(レレバンス評価＝低)から 1(レレバンス評価＝高)の間に入るような正規化を行った。参考値として、マグニチュード推定法の判定値も変換して示した。

判定値の出現分布を 5 段階尺度にあわせ 5 ブロックに分けて示したのが図5のグラフである。

表3 実験1:正規化した判定値の平均

	5段階	スライダー	マグニチュード
文献1	0.187	0.158	0.110
文献2	0.099	0.090	0.042
文献3	0.170	0.124	0.110
文献4	0.142	0.127	0.141
文献5	0.136	0.151	0.130
文献6	0.445	0.496	0.410
文献7	0.786	0.801	0.755
文献8	0.788	0.801	0.824
文献9	0.612	0.663	0.623
文献10	0.811	0.870	0.878
最大値	0.811	0.870	0.878
最小値	0.099	0.090	0.042

グラフからは、5 段階尺度と比較し、スライダー

(とマグニチュード推定法)は1ないし5という極端な値が多い傾向が見てとれる。

表3に示したように、ほぼ同様の傾向は最大値、最小値に現れている。

3.2 文献提示順の影響

先行研究 [4]と同様に分散分析を行い、レレバンス判定が提示順から影響を受けているか否かの統計的検証を行った。実験1、2の5段階尺度、スライダー尺度共にp値は1%水準でも有意な差が見られた。Eisenberg らの研究を補強するものと言える。一方で10件あるいは6件という少ない文献から構成される集合においても提示順による影響が見られたのは Parker らの研究 [5]とは異なる結果であり、より詳細な分析が必要である。

表4 実験1:文献提示順の影響

	5段階		スライダー	
	非⇒レ (N=17)	レ⇒非 (N=17)	非⇒レ (N=17)	レ⇒非 (N=17)
文献1	2.12	1.47	18.18	12.71
文献2	1.76	1.18	8.94	11.41
文献3	2.00	1.29	10.59	12.82
文献4	1.82	1.41	9.47	17.53
文献5	2.00	1.18	13.41	20.29
文献6	3.24	2.06	56.29	37.82
文献7	4.12	4.18	87.12	69.53
文献8	4.06	4.00	92.18	68.47
文献9	3.82	2.82	81.76	48.76
文献10	4.24	3.94	95.00	77.35
最大値	4.24	4.18	95.00	77.35
最小値	1.76	1.18	8.94	11.41

実験1の文献提示順ごとの判定平均値を示したのが表4である。この表において、各尺度の判定平均値の最大値は「レ⇒非」の順に提示した場合の方が、より低くなる傾向が見られた。被験者が提示順の早い文献を過小評価していると考えられ、既往研究の結果を裏付ける結果と言える。しかし、「非⇒レ」の提示順で過大評価する傾向は5段階尺度の最小値では見られたが、スライダーでは見られなかった。

3.4 判定尺度と文献提示順の関係

尺度と提示順の関係を分析する視点の一つとして、最大/最小値の出現頻度を判定尺度と提示順別に算出した結果を表5に示した。この表から5段階尺度は「レ⇒非」において、約半分が最小値と判定されてしまったことがわかる。

表5 実験1:最大値と最小値の出現頻度

最大値頻度		判定尺度		
		5段階	スライダー	マグニチュード
提示順	非⇒レ	2.65	2.42	1.89
	レ⇒非	1.89	1.00	1.28
	混合	2.33	1.58	1.42

最小値頻度		判定尺度		
		5段階	スライダー	マグニチュード
提示順	非⇒レ	3.53	3.00	2.44
	レ⇒非	4.83	2.82	2.94
	混合	3.56	2.32	2.95

ウェブ上の実験システムを用いた二つの実験において、レレバンス判定への判定尺度や文献提示順の判定結果への影響が示された。今後は順位相関や判定時間との関係からの分析を引き続き行う予定である。

■ 謝辞

本実験にご協力くださった亜細亜大学司書講習受講生と関係者の皆様に深く感謝致します。

【注・引用文献】

1) Eisenberg, Michael B. "Measuring Relevance Judgment". Information Processing and Management. Vol.24, No.4, 1988, p.373-389.

2) Bruce, H. W. "Cognitive View of the Situational Dynamism of User-Centered Relevance Estimation". Journal of the American Society for Information Science. Vol.45, No.3, 1994, p.142-148.

3) 例えば、最高裁判所裁判官国民審査でも順序効果があることが、平松貞実. 世論調査で社会が読めるか:事例による社会調査入門. 東京:新曜社, 1998, 250p.で指摘されている

4) Eisenberg, Michael B.; Barry, Carol. "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.39, No.5, 1988, p.293-300.

5) Parker, Purgailis; Lorraine, M.; Johnson, Robert E. "Does Order of Presentation Affect Users' Judgement of Documents". Journal of the American Society for Information Science. Vol.41, 1990, p.493-494.

6) 例えば、Rong, Tang; Solomon, P. "Use of Relevance Criteria across Stages of Document Evaluation: On the Complementarily of Experimental and Naturalistic Studies." Journal of the American Society for Information Science and Technology. Vol.52, No.4, 2001, p.676-685.

7) http://nsearch.yahoo.co.jp/bin/search

8) http://news.google.co.jp/