検索システムの性能評価と利用者実験との比較

齋藤ひとみ * 寺井仁 † 高久雅生 ‡ 江草由佳 §
* 愛知教育大学 情報教育講座 (hsaito@auecc.aichi-edu.ac.jp)
† 名古屋大学附属図書館 研究開発室 (terai@nul.nagoya-u.ac.jp)
‡ 情報・システム研究機構 新領域融合研究センター (masao@nii.ac.jp)
§ 国立教育政策研究所 (yuka@nier.go.jp)

1. はじめに

様々な検索システムがサービスとして提供・利用されている現在において，情報検索システムの性能評価は非常に重要である．検索性能の評価実験は，1950 年代中期の Cranfield 実験から始まったとされ，その後アメリカを中心とした Text Retrieval Conferences(TREC) や，日本語データを対象とした NII-NACSIS Test Collection for IR Systems(NTCIR) に代表される大規模テストコレクションの利用による客観的な評価実験 (システム評価) の実現へと発展している．

しかし Hersh et. al.(2001) や Turpin and Hersh (2001) は，TREC 7-9 の Interactive Track において，システム評価がユーザ実験の結果と一致しなかったことを報告している．これらの知見は，従来のシステム実験での性能指標の結果が，利用者実験による認知特性や主観評価による結果と必ずしも合致しないことを示唆している．しかし，研究データが少なく，なぜ合致しないのかという点についてはまだ十分に検討されていない．さらに，先行研究は TREC のデータを使用したものが多く，その他の大規模テストコレクションにおいてはほとんど検討されていない．

以上を踏まえ本研究ではユーザ実験を実施し，検索システムの性能と満足度などの被験者の主観評価と NTCIR-5 Web task(Oyama et.al., 2005) におけるシステム評価との比較を行う．NTCIR-5 Web task では，Web から約 1 億ページのデータを収集し，269 の課題に対して 63 の検索システムが評価に参加した．ユーザ実験では，このうち 3 つの課題およびシステムをそれぞれ使用した．

2. 方法

実験計画

19〜36 歳の男女 31 名 (男性 21 名，女性 10 名) が実験に参加した．被験者のインターネット平均利用時間は 1 日あたり 2.98 時間 (SD=2.43) であった．

実験は，3 × 3 の混合計画で実施された．第一要因は 3 つの課題 (被験者内要因) で，第 2 要因は 3 つのシステム (被験者内要因) であった．表 1 に示すとおり，被験者は課題 (Movie, Shopping, Restaurant) とシステム (High, Middle, Low) を組み合わせた 3 パターン (Sa, Sb, Sc) に 10 名または 11 名がランダムに割り振られた．

表 1: 課題とシステムの組み合わせ

	High	Middle	Low
Movie	Sa	Sc	Sb
Shopping	Sb	Sa	Sc
Restaurant	Sc	Sb	Sa

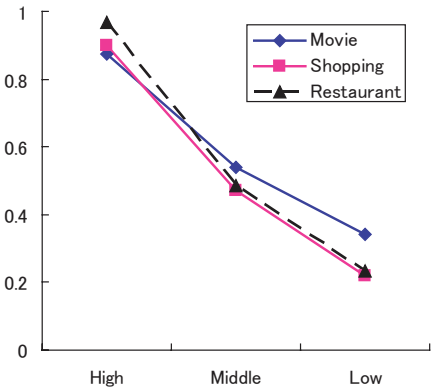

図 1: 各システムの課題ごとの nDCG 値

材料

課題とシステムは，検索性能指標である nDCG (normarized Discounted Cumulative Gain) と MRR(Mean Reciprocal Rank) に基づいて 3 つずつ選定した．nDCG とは，適合文書の適合度を点数に置き換えて検索順位の重みをかけた指標であり，MRR 値とは，最初に出現した適合文書の順位の逆数を全課題で平均した指標である．各指標が高いほど性能は高い．

システムは，検索性能指標 nDCG 値と MRR 値が高い，中程度，低いものを NTCIR-5 Web 参加システムから 3 つ選び，High, Middle, Low とした．課題は，課題に対する nDCG 値が 1 つのシステム内で同程度のものを 3 つ選び，課題内容からそれぞれ Movie, Shopping, Restaurant とした．課題ごとのシステムの nDCG 値を図 1 に示す．なお，課題実施中は，被験者は異なる検索システムを利用していることを知らされていなかった．

手続き

最初に，インターネットやコンピュータの利用経験に関するアンケートを行った．次に実験の教

表 2: 遂行時間，パフォーマンスおよび主観評価の分析結果

	システム × 課題	システム	課題						
	p-value	p-value	p-value	F(2,84)	Main effect			Mse	p-value
適合度	0.193	0.133	0.000**	26.887	Shopping	>	Movie	0.152	0.000**
					Restaurant	>	Movie		0.000**
遂行時間	0.275	0.722	0.249	1.412		-		-	-
困難さ	0.276	0.364	0.057+	2.972	Restaurant	>	Shopping	1.238	0.023*
満足度	0.798	0.21	0.052+	3.703	Shopping	>	Movie	1.019	0.015*
確信度	0.742	0.771	0.021*	4.058	Shopping	>	Movie	0.961	0.001**
適正度	0.559	0.934	0.023*	3.961	Shopping	>	Restaurant	1.127	0.001**
性能	0.733	0.73	0.000**	8.894	Movie	>	Restaurant	0.481	0.000**
					Shopping	>	Restaurant		0.000**
難易度	0.843	0.105	0.013*	4.577	Movie	>	Shopping	0.523	0.003**

+: $p < 0.1$, *: $p < 0.05$, **: $p < 0.01$

示と練習課題を行った．その後，条件に応じて表 1 の課題がランダムな順序で与えられた．課題は Web ブラウザ上に提示された．課題を開始すると，検索目的（例：大和市役所について調べたい），背景（例：大和市役所の場所，電話番号等の情報を知りたい），適合条件 (例：大和市役所の公式ページを適合とする)，および検索画面へのリンクが提示された．被験者は好きなタイミングで検索画面へ移ることが出来た．検索画面では，システム評価で使用した検索語 (例：大和市役所) で検索した結果の一覧が提示された．被験者は検索結果一覧から課題文に適合すると思われるページを探した．ページが見つかった時点で検索を終了し，課題に対する評価を行った．全ての課題が終了した後でシステム間，課題間の評価を行った．

課題に対する評価では，(1) 課題の困難さ，(2) 結果に対する満足度，(3) 結果に対する確信度，(4) 課題の検索に対するシステムの性能，の 4 点に対して 5 段階の評価を求めた．一方システム間，課題間の評価では，3 つの課題で異なる検索システムが使用されていたことを説明した後に，(1) 3 つのシステムの性能，(2) 課題自体のわかりやすさ，の 2 点に対して 5 段階での評価を求めた．

3. 実験結果と考察

課題ごとの適合度，遂行時間および 6 つの主観評価について，課題とシステムの 2 要因被験者間分散分析を行った (表 2)．

適合度は，NTCIR-5 Web Task において公式判定者が行った適合判定結果を基に算出した．公式判定者は，システムが検索したページについて，適合・部分適合・不適合をそれぞれ判定した．本実験では，被験者が適合すると判断したページが適合である場合は 1，それ以外は 0 として適合度を算出した．

分析の結果，適合度については課題の主効果が見られ，Movie に比べ Shopping や Restrant の方が有意に高かった．平均遂行時間には有意な差は見られなかった．

次に主観評価について分析した結果，すべてにおいて課題の主効果が確認された．まず課題に対する困難度は Restrant が有意に高く，満足度と確信度は Shopping が Movie に比べて有意に高かった．課題に対するシステムの適正度は，Shopping が Restaurant に比べて有意に高く，システムの性能は他の課題に比べ Restraunt が有意に低かった．また課題の難易度は Shopping が Movie に比べて有意に高かった．

どの結果においてもシステム間に差がなかった．したがって先行研究と同様に NTCIR の評価データにおいても，システム評価によるシステムの性能とユーザ実験におけるユーザパフォーマンスの結果が一致しないことが確認された．また，課題間の差は適合度と 6 つの主観評価において見られた．したがって，システム性能の差に比べて，課題の違いの方がユーザ実験では顕著に現れることが示唆された．

4. おわりに

実験の結果，NTCIR-5 の Web Task におけるシステム評価とユーザ実験の結果は一致しなかった．これらの結果は，ユーザの評価により近い検索性能指標の開発を示唆する結果であったといえる．今後は，被験者のログデータなどの分析を行い，システム性能指標との関連性についてより詳細に検討を進める．

参考文献

Hersh, W. R., Turpin, A., Price, S., Kraemer, D., Olson, D., Chan, B., & Sacherek, L. (2001). Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations.. *Inf. Process. Manage.*, **37** (3), 383–402.

Oyama, K., Takaku, M., Ishikawa, H., Aizawa, A., & Yamana, H. (2005). Overview of the NTCIR-5 WEB navigational retrieval subtask 2 (Navi-2). *Proceedings of NTCIR-5 Workshop Meeting*, 242–222.

Turpin, A. & Hersh, W. (2001). Why batch and user evaluations do not give the same results. *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 225–231. New York, NY, USA:, ACM Press.