감성과학, Vol.15, No.3, pp.367-380, September 2012

정보 검색 과제별 동적 검색 랭킹 모델 구현 및 검증: 사용자 중심 적합성 판단 모형 평가를 중심으로

Implementation and Verification of Dynamic Search Ranking Model for Information Search Tasks: The Evaluation of Users' Relevance Judgement Model

박정아*, 손영우**†
Jung-ah Park*, Young Woo Sohn**†

다음커뮤니케이션 검색본부*
Search Service Unit, Daum Communications*

연세대학교 심리학과**
Department of Psychology, Yonsei University**

Abstract

The purpose of this research was to implement and verify an information retrieval(IR) system based on users' relevance criteria for information search tasks. For this purpose, we implemented an IR system with a dynamic ranking model using users' relevance criteria varying with the types of information search task and evaluated this system through user experiment. 45 participants performed three information search tasks on both IR systems with a static and a dynamic ranking model. Three Information search tasks are fact finding search task, problem solving search task and decision making search task. Participants evaluated top five search results on 7 likert scales of relevance. We observed that the IR system with a dynamic ranking model provided more relevant search results compared to the system with a static ranking model. This research has significance in designing IR system for information search tasks, in testing the validity of user-oriented relevance judgement model by implementing an IR system for actual information search tasks and in relating user research to the improvement of an IR system.

Keywords : Relevance, User-centered Relevance, Relevance Judgement, Relevance Evaluation, Information Search Task, Search Ranking Model

요 약

본 연구는 정보 검색 과제별 주요 적합성 판단 기준을 실제 정보 검색 시스템으로 구현해 보고 사용자 평가를 통해 그 효과를 검증해 보고자 하였다. 이를 위해, 사용자 적합성 판단 기준들을 정보 검색 시스템에서 적합성을 결정하는 검색 랭킹 모델의 랭킹 요소들로 적용하였다. 그리고 정보 검색 과제별 차이가 있는 동적 검색 랭킹 모델과 차이가 없는 정적 검색 랭킹 모델을 시스템으로 구현하였고, 이에 대한 사용자 평가를 진행하여 비교해 보았다. 총 45명의 참가자가 실험에 참여하였고, 정보 검색 과제별 차이가 있는 동적 검색 랭킹 모델과 차이가 없

† 교신저자 : 손영우(연세대학교 심리학과)
E-mail : ysohn@yonsei.ac.kr
TEL : 02-2123-2444
FAX : 02-365-4354

는 정적 검색 랭킹 모델이 적용된 각각의 검색 시스템에서 3개의 검색 과제를 수행하였다. 3개의 정보 검색 과제로는 사실 검색 과제, 문제 해결 검색 과제, 의사 결정 검색 과제가 사용되었다. 각 참가자는 검색 결과 첫 페이지 상위 5 개의 검색 결과에 대해 적합성 정도를 7 점 척도로 평가하였다. 그 결과, 사용자는 전반적으로 모든 검색어에 동일하게 반응하는 정적 검색 랭킹 모델을 적용한 시스템보다 정보 검색 과제별로 사용자 적합성 판단 기준의 변화에 따라 랭킹 요소 가중치를 달리한 동적 검색 랭킹 모델을 더 높이 평가하는 것을 확인할 수 있었다. 본 연구는 이를 통해, 정보 검색 과제를 고려한 정보 검색 시스템 디자인의 필요성과 함께, 사용자 중심 적합성 판단 모형 연구 결과를 실제 정보 검색 시스템으로 구현하여 평가함으로써 사용자 중심 적합성 연구 결과의 타당성을 검증하였다는 점, 그리고 사용자 연구 접목을 통한 시스템 개선의 중요성을 강조하였다는 점에서 의의를 가진다.

주제어 : 적합성, 사용자 중심 적합성, 적합성 판단, 적합성 평가, 정보 검색 과제, 검색 랭킹 모델

1. 서론

정보 과학 연구 분야는 정보 검색 알고리즘에 관한 시스템 중심 연구와 사용자 행동에 관한 사용자 중심 연구로 분류될 수 있는데(Saracevic, 1999), 이들 간의 시너지는 아직까지 충분히 이루어진다고 보기 어렵다. 이에 대해 Saracevic(1999)은 “안타깝게도 대부분의 인간 중심 연구는 구체적인 해결책보다는 제안에 그치는 경우가 대부분이었다. 반면 시스템 중심 연구에서는 사람, 즉 사용자 측면을 거의 고려하지 못하였다”(p. 1057)고 하였다. 실제로 사용자 중심 접근과 시스템 중심 접근 간의 통합과 시너지 발생은 정보 과학 연구 분야의 도전 과제로 남아있다(Ingwersen & Järvelin, 2005).

정보검색의 주요 목적은 적합한 문서를 찾아주는 것으로, 적합성은 정보 검색의 핵심 개념이다. 정보 검색 시스템에서 적합성을 결정짓는 것은 검색 랭킹 모델이다. 검색 랭킹 모델에는 불린 모델(Baeza-Yates & Ribeiro-Neto, 1999), 벡터 스페이스 모델(Salton, 1971; Salton & Lesk, 1968), BM25(Robertson, 1997)등 다양한 모델들이 존재한다. 최근에는 기계 학습(machine learning) 방법도 많이 사용되고 있다(Joachims, 2002; Agichtein, Brill, & Dumais, 2006). 검색 랭킹 모델은 문서를 랭킹 함수에 의해 구해진 점수에 기반하여 내림차순으로 정렬된 목록(ranked list)으로 결과를 생성한다. 대부분의 검색 랭킹 모델에서는 모든 검색어에 하나의 랭킹 함수가 사용된다(Geng et al., 2008). 최근 Geng 등(2008)의 연구에서는 검색어별 다른 랭킹(Query-dependent ranking) 모델 방식의 필요성을 주장한 바 있다. 이처럼 정보 검색 분야에서 검색어 별 다른 랭킹의 필요성이 제기되고 있긴 하지만, 대부분 검색어 분류에 관한 연구들에 집중되어 있고, 실제 랭킹 모델을 제안한 경우는 드물다(Geng et al., 2008).

본 연구는 사용자가 적합성을 판단하는 기준이 정보 검색 과제별로 차이가 있음을 확인한 선행 연구 결과를 정보 검색 과제별로 다른 검색 랭킹 모델을 적용함으로써 그 효과를 검증해 보고자 하였다. 이를 위해 본 연구는 사용자 적합성 판단 기준과 연결될 수 있는 요인들을 랭킹 요소로 검색 랭킹 모델에 반영하였다. 그리고 정보 검색 과제에 따라 랭킹 요소 가중치가 다르게 반영되는 “동적 검색 랭킹 모델”과 과제와 상관없이 동일한 가중치가 반영되는 “정적 검색 랭킹 모델”을 각각 적용한 정보 검색 시스템의 사용자 평가를 비교해 보았다. 본 연구는 사용자 중심 적합성 판단 모형에 관한 선행 연구 즉, 정보 검색 과제별 적합성 판단 기준과 적합성 유형의 관계에 관한 연구 결과를 실제 정보 검색 시스템으로 구현해 보았다는 점, 이를 통해 사용자 연구 결과의 타당성을 검증해 볼 수 있다는 점, 결과적으로 사용자 연구 접목을 통한 시스템 개선의 중요성을 알린다는 점에서 의의를 가진다.

2. 선행연구

적합성은 정보 검색의 주요 연구 분야이다(Borlund, 2003; Mizzaro, 1997; Schamber, 1994). 적합성은 정보 검색 시스템의 기능과 평가에 있어 기본적이면서도 중요한 것으로 알려져 있다(Borlund, 2003). 기존 연구들에 의하면 적합성은 복합적이고 다차원적이면서 또한 상황에 따라 달라지는 동적인 개념으로, 인지적, 정서적, 사회 문화적 요인들에 영향을 받는다고 알려져 왔다(Schamber, 1994). 정보 이용자들은 실제 검색과정에

서 단지 '적합하다'/'적합하지 않다'가 아닌, 주관적이고 역동적으로 적합성을 판단한다(Kekäläinen & Järvelin, 2002). 이러한 의미에서 적합성은 다차원적이며 어떤 하나의 적합성만으로 정의될 수 없다(Cuadra & Katter, 1967; Barry, 1994; Saracevic, 1975, 1997; Schamber, 1994).

초기의 적합성은 정확률(precision)과 재현율(recall) 같은 시스템 중심의 개념으로 논의되었다. 그러나 시스템 중심의 적합성만으로는 적합성 개념을 충분히 설명하기 부족하다는 주장이 제기되면서 적합성 개념에 대한 논의가 재기되었다(Saracevic, 1975; Schamber, Eisenberg, & Nilan, 1990). 그리고 사용자 중심 관점의 연구가 활발해 지기 시작했다(Cosijn & Ingwersen, 2000; Borlund, 2003; Schamber et al., 1990; Saracevic, 1970). 적합성에 관한 시스템 중심 접근만으로는 실질적인 사용자 요구를 고려하기 어렵고, 적합성 판단에 영향을 주는 상당수의 주관적이고 상황적인 변수들을 고려하지 못하는 한계를 가진다(Barry, 1994; Fruend, 2008). 사용자 관점에서 적합성은 "다차원의 인지적 개념으로써 사용자의 정보 인식과 정보 이용자의 정보 요구 상황에 상당 부분 의존한다"고 정의될 수 있다(Borlund 2003, p913).

적합성에 대한 중요한 연구 주제 중 하나는 사용자의 적합성 판단 기준이다. 즉 사용자가 문서가 적합한지 아닌지를 판단하는 속성들이다. 기존 많은 연구들은 사용자가 어떤 기준으로 정보를 적합하다고 판단하는지, 즉 사용자 적합성 판단에 영향을 미치는 다양한 기준들을 연구해 왔다(예: Cuadra & Katter, 1967; Rees & Schultz, 1967; Schamber, 1991; Park, 1993; Barry, 1994; Barry & Schamber, 1998; Bateman, 1998; Choi & Rasmussen, 2002; Savolainen & Kari, 2006). Schamber, Eisenberg 와 Nilan(1990, p.773)은 "사용자가 정보를 평가하는데 사용하는 기준 등을 연구함으로써 적합성에 대한 더 구체적인 이해를 할 수 있게 됨과 동시에 시스템 디자인에도 도움을 줄 수 있을 것이다" 라고 하여 사용자 중심 적합성 연구의 중요성을 강조하기도 하였다.

최근 많은 연구에서 검색 행동, 특별히 적합성 평가에 있어서 정보 과제의 다양성으로 인한 효과를 연구하였다(Limberg, 1999; Kelly et al., 2002; Tombros, Ruthven & Jose, 2005). Limberg(1999)는 사실을 찾거나 명백한 질문에 대한 답을 찾는 사실 검색, 의사 결정에 필요한 충분한 정보를 찾는 의사 결정 검색, 내용을 이해하는데 필요한 자료를 찾는 내용 이해 검색과 같이 정보 이용 목적이 다른 경우의 학생들의 연구 행동을 비교하여 연구하였다. 그 결과 적합성, 성향, 정보량과 권위 등의 많은 요인들에서 다른 방식으로 정보를 다루고 평가하는 것을 알 수 있었다. Kelly 등(2002)의 연구에서는 미리 정해진 문서 집합을 가지고 검색 과제 유형(사실 검색과 과정 검색)에 따라 적합성 평가를 해 본 결과 각각 문서의 다른 요소들이 사용된다는 것을 밝혔다. Tombros, Ruthven과 Jose(2005) 또한 과제 유형에 따른 적합성 판단 기준을 연구하였다. 3가지의 통제된 검색결과, 배경 지식 검색(인터넷 사용자 인구 정보), 의사 결정 검색(좋은 하이파이 스피커 결정), 자료 수집 검색(일본 교토에서 열리는 학회에 참가해서 주말에 할 리스트 작성)을 사용한 이 연구에서는 과제에 따라 적합성을 평가하는데 사용된 판단 기준이나 특징에 차이가 있다는 것을 알려주었다.

보다 구체적으로 정보 검색 과제에 따라 사용자가 적합성을 판단하는 기준과 적합성 간의 관계를 알아본 연구도 있다. Park과 Sohn(2009)은 사용자 중심 적합성 판단 모형을 통해, 사실 검색에서는 최신이면서 다양하고 흥미 있는 내용의 문서가, 문제 해결 검색에서는 이해하기 쉬우며 찾는 내용의 범위나 초점을 잘 맞춘 문서가, 의사 결정 검색에서는 최신이고, 이해하기 쉬우며, 다양하고 흥미 있는 내용, 그리고 찾는 내용의 범위나 초점을 잘 맞춘 문서가 적합성을 판단하는데 긍정적인 영향을 주는 것을 알 수 있었다. 또한 이 연구는 사용자 적합성에 대해, 새로운 정보를 알게 된 인지 적합성, 문서가 과제 해결에 유용하다고 판단하는 상황 적합성, 검색 목적을 달성함으로써 만족하는 정서 적합성이 존재하며 이들이 함께 "적합성"을 이루는 개념들이라는 것을 제시함으로써 적합성 유형들 간의 관계를 실증적으로 밝히기도 하였다.

이러한 연구 결과들은 정보 검색 과제에 따라 적합성 평가에 사용되는 특징이나 문서 요소가 다르다는 것을 알려준다. 이처럼 정보 검색 과제가 검색 행동에 영향을 미친다는 것이 어느 정도 알려져 왔지만, 아직까지 검색 시스템 상에 존재하는 과제 기반의 영향력에 대한 실질적 적용 등은 아직 명확하게 규명되지 않은 상황이다.

3. 연구 시스템

본 연구는 정보 검색 과제별 주요 적합성 판단 기준을 실제 정보 검색 시스템에 적용하여 구현해 보고 사용자 평가를 통해 그 효과를 검증해 보고자 하였다. 이를 위해 사용자가 적합성을 판단하는 기준을 검색 시스템 상의 랭킹 모델에 반영하여, 정보 검색 과제별로 다른 검색 랭킹 모델을 적용함으로써 그 효과를 검증하는 방식으로 접근하였다. 정보 검색 시스템에서 적합성을 결정짓는 검색 랭킹 모델은 문서를 랭킹 함수에 의해 구해진 점수에 기반하여 내림차순으로 정렬된 목록(ranked list)으로 결과를 생성한다. 대부분의 검색 랭킹 모델에서는 모든 검색어에 하나의 랭킹 함수가 사용된다(Geng et al., 2008). 최근 Geng 등(2008)의 연구에서는 검색어별 다른 랭킹(Query-dependent ranking) 모델 방식의 필요성을 주장한 바 있다. 이처럼 정보 검색 분야에서 검색어 별 다른 랭킹의 필요성이 제기되고 있긴 하지만, 대부분 검색어 분류에 관한 연구들에 집중되어 있고, 실제 랭킹 모델을 제안한 경우는 드물다(Geng et al., 2008).

본 연구는 정보 검색 과제별 사용자 적합성 판단 기준을 연구한 선행 연구(Park et al., 2009) 결과를 실제 정보 검색 시스템의 검색 랭킹 모델 즉, 정보 검색 과제 별 다른 검색 랭킹 모델을 적용한 동적 검색 랭킹 모델과 정적 검색 랭킹 모델로 구현하여 사용자 평가를 통해 그 효과를 검증한 것으로써, 정보 검색 시스템에서 검색 랭킹 모델 부분에 초점을 맞추었다.

3.1. 검색 랭킹 모델 개요

본 연구에서 검색 랭킹 모델은 문서 별 랭킹 요소 점수를 가중치에 기반하여 선형적으로 합산하는 방식을 사용하였다(Kang & Kim, 2003; Desai & Spink, 2004). 문서의 랭킹 점수는 다음 식과 같이 계산되었다.

$$Score(Q,D) = \sum_{i=1}^{n} Wi \times Fi$$

n은 랭킹 요소의 개수를 나타내고, W는 각 랭킹 요소의 가중치를, F는 각 랭킹 요소에 대한 문서의 점수를 나타낸다. Q는 검색어, D는 문서를 가리킨다. 이렇게 계산된 점수가 검색어에 대한 문서의 검색 랭킹 점수가 된다. 각 랭킹 요소에 대한 가중치는 해당하는 랭킹 요소의 중요도를 대표한다(Cutler, Shih & Meng, 1997). 본 연구에서 각 랭킹 요소의 가중치는 계단식 접근으로 휴리스틱하게 설정되었다(Desai & Spink, 2004). 기본 가중치를 1로 보고 랭킹 요소의 상대적인 중요도에 따라 그 값을 조정하였다. 가중치는 3인의 정보 검색 전문가의 피드백을 받아 최종 확정하는 방식으로 결정하였다.

본 연구에서는 사용자 적합성 판단 모형에 관한 선행연구(Park et al., 2009)에서 밝혀진 사용자 적합성 판단과 연결될 수 있는 문서의 특징들을 정량화하여 랭킹 요소로 사용하였다. 실제 검색 랭킹 모델에 반영한 적합성 판단 기준은 주제성, 신선성, 이해가능성, 구체성의 4가지이고, 신뢰성, 흥미성, 특수성은 정량화로 인한 구현 제약으로 적용 범위에서 제외되었다. 각 적합성 판단 기준은 다음과 같이 랭킹 요소로 구현하였다.

1) 주제성

주제성은 문서의 내용과 검색어의 유사도를 계산하는 방식으로 시스템에 반영하였다. 문서의 내용과 검색어의 유사도는 시스템에 따라 단어의 수와 빈도를 고려하거나 벡터 스페이스 모델에서의 코사인(cosine) 유사도와 같은 방식으로 계산될 수 있다. 본 연구에서는 정보 검색에서 널리 사용되는 TF, IDF를 이용하여 다음과 같이 계산하여 주제성 점수로 반영하였다(Kang & Kim, 2003).

$$TS = \sum_{t \in (Q \cap Dd)} TF_{d,t} \times IDF_t$$

$$TF_{d,t} = 0.4 + 0.6 \times \frac{tf_{d,t}}{tf_{d,t} + 0.5 + 1.5 \times \frac{doclen_d}{avg(doclen)}}$$

$$IDF_t = \log\left(\frac{N + 0.5}{df_t}\right) / \log(N + 1)$$

N은 전체 문서의 수, $tf_{d,t}$ 는 문서 d에 나타난 단어(term) t의 빈도수, df_t 는 해당 단어 t가 나타난 문서의 수, doclend 는 해당 문서의 길이를, avg(doclen)은 전체 문서의 평균 길이를 각각 나타낸다.

2) 신선성

Xu와 Chen(2006)은 기존 적합성 판단 기준 연구들에서 정의한 신선성, 최신성, 현재성, 시의성 등의 기준들을 신선성으로 묶어서 정의한 바 있다. 본 연구에서는 정량화 가능한 문서의 날짜 요소를 신선성으로 반영하였다. 문서의 날짜가 최근일수록 문서의 신선성 랭킹 요소 점수가 올라가는 방식이다. 최신성 점수 수식은 아래와 같이 계산되었고, 0～1의 값의 범위를 가진다.

$$NS = \frac{1.0}{x/7.0 \times 6.0 + 1.0}$$

x 는 현재 날짜와 해당 문서 날짜와의 차이를 나타낸다. 예를 들어 오늘 날짜 문서는 0, 어제 날짜 문서는 1, 일주일 전 날짜 문서는 7과 같은 식이다.

3) 이해가능성

이해가능성은 정보가 이해되기 쉬운 수준으로 정의된다(Savolainen & Kari, 2006). Xu와 Chen(2006)은 기존 연구에서의 이해가능성(Barry, 1994; Bateman, 1998; Spink et al., 1998), 명확성(Schamber, 1991; Barry, 1994; Hirsh, 1999), 가독성(Park, 1993), 지나치게 기술적인 내용(Park, 1993; Spink et al., 1998), 잘 씌어진 것(Bateman, 1998), 그림 등이 제공되는지(Bateman, 1998; Maglaughlin & Sonnewald, 2002), 언어(Spink et al., 1998; Tang & Solomon, 1998; Hirsh, 1999) 등을 이해가능성 기준에 포함시킨 바 있다. 본 연구에서는 이해가능성 개념 중에서 측정이 가능한 이미지 또는 동영상을 포함하고 있는지 여부로 검색 랭킹 모델에 반영하였다.

$$US = \begin{cases} 1 & if\ d \in A \\ 0 & if\ d \notin A \end{cases}$$

4) 구체성

구체성은 기존연구에서 내용이 충분한지(Park, 1993; Fitzgerald & Gallway, 2001; Spink et al., 1998), 길이가 적당한지(Bateman, 1998), 내용이 상세한지(Bateman 1998)등으로 정의되었다. 본 연구에서는 문서의 양, 즉 문서 길이를 구체성 요소로 검색 랭킹 모델에 반영하였다. 문서의 길이가 긴 문서일수록 점수를 더 받을 수 있도록 하였고, 400byte 단위를 기준으로 구간화 한 후 점수를 계산하였다. 문서 길이400 byte 미만은 0, 400byte ～ 800byte 사이는 1과 같은 식으로 구간화되었다. 다음 식에 의해 점수가 계산되었고 값의 범위는 0～1 사이이다.

$$LS = \frac{\log(Level)}{4}$$

$$Level = \frac{doclen}{400}$$

3.2. 실험시스템

실험을 위해 정보 검색 과제 별 다른 동적 검색 랭킹 모델과 정보 검색 과제에 따라 차이가 없는 정적 검색 랭킹 모델이 각각 적용된 정보 검색 시스템이 준비되었다.

3.2.1. 시스템 1: 정적 검색 랭킹 모델 시스템

정적 검색 랭킹 모델을 이용한 시스템은 대부분의 정보 검색 시스템처럼 모든 검색어에 하나의 랭킹 함수가 적용되는 방식으로, 각 랭킹요소의 가중치가 검색어와 상관없이 동일하게 적용되었다. 랭킹 요소의 가중치는 주제성을 대표하는 검색어와 문서의 유사도 가중치는 1, 최신성을 대표하는 날짜 가중치는 0.5, 이해가능성을 대표하는 이미지/동영상 여부는 0.1, 구체성과 관련된 문서길이는 0.05 로 설정하였다. 이 가중치는 3인의 정보 검색 전문가의 피드백을 받아 최종 확정하는 방식으로 결정하였다.

3.2.2. 시스템 2: 동적 검색 랭킹 모델 시스템

동적 검색 랭킹 모델을 이용한 시스템은 기본적으로 설정된 각 랭킹요소의 가중치가 정보 검색 과제에 따라 다르게 적용되었다. 정보 검색 과제는 앞선 연구에서 구분한 것과 동일하게 사실 검색 과제, 문제 해결 검색 과제, 의사 결정 검색 과제로 구분하였다. Park과 Sohn(2009)의 연구 결과에 따르면 사실 검색 과제에서는 주제성, 신뢰성, 신선성, 구체성, 흥미성이, 문제 해결 검색 과제에서는 주제성, 신뢰성, 이해가능성, 특수성이, 의사 결정 검색 과제에서는 주제성, 신뢰성, 신선성, 이해가능성, 구체성, 특수성, 흥미성

이 적합성을 판단하는 주요 기준으로 밝혀졌다. 이에 따라 사실 검색 과제에서는 기본 검색 랭킹에서 신선성, 구체성 가중치를 높이고, 문제 해결 검색 과제에서는 이해가능성 가중치를, 의사 결정 검색 과제에서는 신선성과 이해가능성 가중치를 높게 반영하였다. 과제별 가중치는 3인의 정보 검색 전문가의 피드백을 받아 최종 확정하는 방식으로 결정하였고, 표 1과 같이 적용되었다.

Table 1. Weights of search ranking factors

ranking factor	static search ranking model	dynamic search ranking model		
		fact finding search task	problem solving search task	decision making search task
topicality (similarity of between search query and document)	1	1	1	1
novelty (document date)	0.5	1	0	1
understandibility (inclusion of image/video)	0.1	0.1	0.3	0.3
volume (document length)	0.05	0.1	0.05	0.1

4. 연구방법

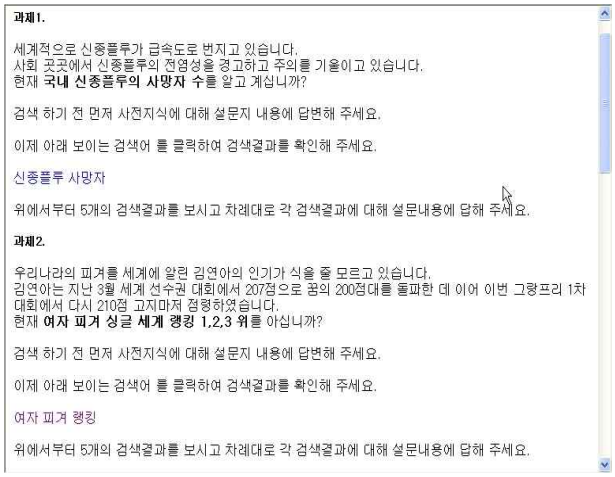
본 연구는 실험 참가자가 정적 검색 랭킹 모델이 적용된 시스템과 동적 검색 랭킹 모델이 적용된 시스템 각 2 개의 시스템에서 3 개의 정보 검색 과제를 수행하는 참가자 내 실험으로 설계되었다. 정보 검색 과제 유형별로 검색 랭킹 모델의 차이를 구현하기 위하여, 정보 검색 과제에 해당하는 검색어를 제한하였다. 제한된 검색어들로 검색 랭킹 만족도를 비교하는 것은 기존 검색 랭킹 모델 관련 연구들에서 사용되어 온 방법이다(예: Vaughan, 2004; Drori, 2002). 학습 효과를 고려하여 각 3개의 정보 검색 과제 유형별로 2 개의 다른 검색어가 지정되었고, 이는 순서 효과를 고려하여 라틴방형으로 설계되어 제시되었다.

4.1. 참가자

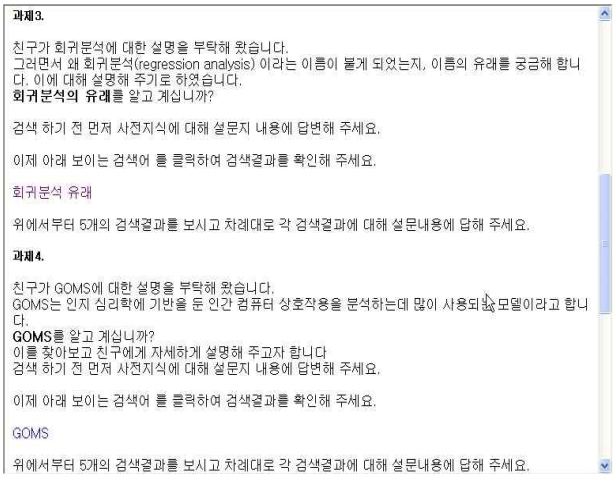
총 45명의 대학생들과 일반인이 실험에 참가하였다. 실험은 실험실에서 진행되었다. 45명의 실험 참가자 중 남자는 31명(68.9%), 여자는 14명(31.1%)이었고, 평균 24.7세였다. 이들은 대부분 10년 이상 컴퓨터를 사용해 왔고(84.4%), 매일 1시간 이상 컴퓨터를 사용하였다(75.6%). 대부분 네이버에서 인터넷(62.2%) 및 검색(68.9%)을 사용하였다. 이들은 스스로 인터넷 검색에 대해 잘 알거나(37.8%) 보통 수준으로 알고 있다(48.9%)고 생각하였다.

4.2. 정보 검색 과제

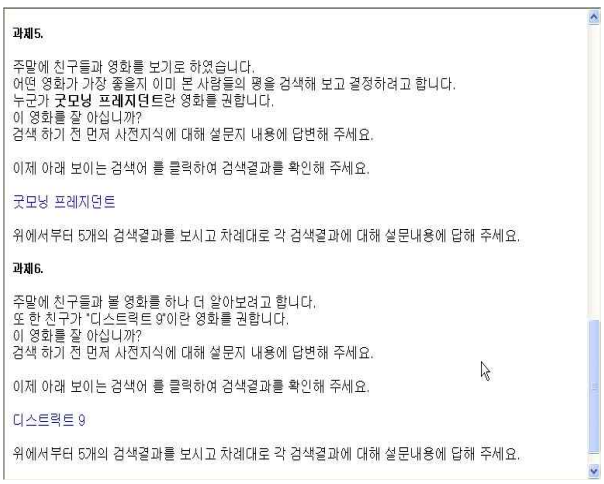
정보 검색 과제는 Park과 Sohn(2009)의 연구에서 사용된 것과 마찬가지로 사실 검색 과제, 의사 결정 관련 검색과제, 문제 해결 검색 과제의 3가지를 대표하는 검색 과제로 구분하였다. 참가자 내 실험이었기 때문에 학습효과를 고려하여 각 과제 유형별로 2개의 다른 검색 과제에 해당하는 검색어가 제시되었다. 사실 검색 과제로는 신종플루의 사망자 수, 그리고 현재 여자 피겨 세계 랭킹 1, 2, 3 위를 찾는 과제가 지정되었고, 각각 검색어는 “신종플루 사망자”, “여자 피겨 세계 랭킹”이 지정되었다. 문제 해결 검색 과제로는 회귀 분석의 유래를 알아보는 과제와 GOMS에 대해 설명하는 과제를 지정하였고, 각각 검색어는 “회귀 분석 유래”, “GOMS”가 설정되었다. 의사 결정 검색 과제로는 주말에 볼 영화를 검색을 통해 결정하는 과제를 지정하였고, “굿모닝 프레지던트”와 “디스트릭트 9”의 두 가지 검색어가 설정되었다. 검색어를 미리 정의하기는 했지만 참가자에게 검색 과제에 관한 충분한 맥락 정보를 제공할 수 있도록 Borlund(2003)의 모의 정보 검색 과제 상황 지침에 따라 과제에 대한 충분한 배경 설명과 함께 검색어를 제시하였다. 정보 검색 과제 제시 화면 예는 그림 1에 나타나 있다. 검색하기 전 각 검색 과제에 대한 사전 지식을 미리 측정하였다.



(a) Fact-finding search task(사실 검색 과제)



(b) Problem-solving search task(문제 해결 검색 과제)



(c) Decision-making search task(의사 결정 검색 과제)

Figure 1. Screenshots of search tasks

4.3. 실험 도구

검색 데이터는 한국 검색 포털 기업인 다음에서 실제로 블로그 검색 서비스에 사용되고 있는 약 7천만 건의 블로그 데이터를 대상으로 하였다. 이 데이터는 국내 주요 블로그 – 네이버, 다음, 싸이월드, 티스토리, 이글루스 등의 데이터들이 포함되어 있다. 정적 검색 랭킹 모델을 적용한 정보 검색 시스템(시스템 1)과 동적 검색 랭킹 모델을 적용한 정보 검색 시스템(시스템 2)은 검색 데이터와 인터페이스 등 모든 조건이 동일하였다. 두 시스템간의 유일한 차이는 검색어에 따른 랭킹 요소의 가중치로, 이는 앞서 표 1에 제시되었다. 시스템 1은 검색어와 상관없이 각 랭킹 요소의 가중치가 변하지 않은 정적 검색 랭킹 모델이 적용되었고, 시스템 2는 사실 검색, 문제 해결 검색, 의사 결정 검색 과제별로 각 랭킹 요소의 가중치가 달라지는 동적 검색 랭킹 모델이 사용되었다. 그림 2는 제시된 검색 결과 화면의 한 예이다.

(a) System 1 with static search ranking model

(b) System 2 with dynamic search ranking model

Figure 2. Screenshots of search results("swine flu deaths")

4.4. 실험 절차

참가자는 먼저 실험 참가 동의서를 읽고 서명한 후, 실험에 대한 대략적인 설명을 들었다. 이후 사전 설문(나이, 성별, 직업, 인터넷 사용 경험 및 지식, 검색 사용 경험 및 지식 등)에 답하였다. 그리고 정적 검색 랭킹 모델이 적용된 시스템 1과 동적 검색 랭킹 모델이 적용된 시스템 2에서 각각 3개, 즉 총 6개의 정보 검색 과제를 수행하였다. 시스템과 정보 검색 과제는 순서 효과를 고려하여 라틴방형으로 설계되어 참가자에게 제시되었다. 그림 1과 같이 제시된 화면에서 차례대로 검색 과제에 대한 설명을 읽고, 검색 전에 먼저 검색 과제에 대한 사전 지식에 대해 7 점 척도로 답하였다. 그리고 나서 화면에 제시된 검색어의 링크를 클릭하여 검색결과를 확인하고 평가하였다(그림 2 참고). 참가자는 검색 결과 첫 페이지에서 10 개의 검색 결과 중 상위 5 개를 평가하였다(참고: Su, 2003; Spink & Jansen, 2004).

결과적으로 모든 참가자들은 정적 검색 랭킹 모델과 동적 검색 랭킹 모델이 적용된 시스템 별로 3 개의 정보 검색 과제, 즉 총 6 개의 검색 과제를 수행하였고, 각 과제에 대해서 다음과 같은 항목으로 평가하였다. 먼저 검색 결과 중 상위 5 개 각 문서에 대하여 적합성 유형별로 평가하였다. 즉 정보성으로 알려진 인지 적합성은 “이 문서는 새로운 정보를 알게 해준다(정보성)”의 항목으로, 유용성으로 알려진 상황 적합성은 “이 문서는 유용하다(유용성)”, 만족도로 알려진 정서 적합성은 “이 문서는 만족스럽다(만족도)”의 항목으로 측정하였다. 그리고 사용자에게 직접적으로 적합성 여부를 물어보는 항목으로 “이 문서는 검색결과로써 적합하다(적합성)”를 별도로 측정하였다.

검색 결과 상위 5개에 대한 각 문서 단위 평가 외에도, 과제에 대한 전반적인 검색 결과 평가를 위해 검색 결과에 대한 통합 점수를 평가하도록 지시되었다. 이는 “원하는 내용을 알게 되었다(지식습득)”, “검색이 성공적이었다(검색성공)”, “검색 목적을 달성했다(목적달성)”, “검색결과에 만족한다(검색만족)”의 항목을 통해 측정하였다. 모든 항목은 7점 척도로 측정되었다(전혀 그렇지 않다 1, 보통이다 4, 매우 그렇다 7).

5. 연구결과

5.1. 시스템 효과 분석

상위 5개 검색 결과에 대한 각 사용자 평가 점수는 일반 누적 점수(CG: Cumulative Gain)와 검색 결과 순서를 고려한 누적 점수(DCG: Discounted Cumulative Gain)로 산출되었다. 이는 Järvelin과 Kekäläinen(2002)가 제안한 검색 결과 평가 방법으로 검색 랭킹 모델 비교나 검색 성능 평가 연구에서 많이 사용되고 있는 방식이다(Clarke et al., 2008; Agichtein, Brill & Dumais, 2006; Geng et al., 2008). 일반 누적 점수 CG는 다음 수식에 의해 계산되었다.

$$CG = \sum_{i=1}^{n} rel_i / N$$

즉, 각 문서에 대한 사용자 평가 점수 5개를 합산하여 평균을 내는 방식으로 사용하였다. Järvelin 과 Kekäläinen(2002)은 일반 누적 점수 외에도 검색 결과 순위를 고려한 누적 점수인 DCG를 제안하였다. 일반 누적 점수인 CG는 문서가 검색 결과 순위를 고려하지 않는 반면 DCG는 이를 반영한 점수 산출 방식이다. DCG는 다음 수식에 의해 계산되었다.

$$DCG = (rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}) / N$$

이처럼 검색 결과 상위 5개 각 문서별로 사용자가 평가한 정보성, 유용성, 만족도, 적합성 점수 각각은 순위를 고려하지 않은 누적 점수 CG와 순위를 고려한 누적 점수 DCG로 산출되었다. 이렇게 산출된 일반 누적 점수 CG와 순위를 고려한 DCG 점수, 그리고 전반적인 검색 결과 평가 점수를 시스템 비교에 사용하였다. 표 2 에 각 시스템에 대한 평가 항목별 사용자 평가 점수의 평균과 표준편차, 두 시스템 간 사용자 평가 점수 차이를 비교한 시스템 효과와 정보 검색 과제별 시스템 효과가 제시되어 있다.

시스템 1, 2에 대한 각 평가항목에서 차이가 있는지를 알아보는 시스템 효과는, 반복측정 이원 분산 분석(2:시스템 X 3:과제-사실/문제해결/의사결정)을 통해 분석되었다. 정보 검색 과제 유형별로 시스템에 있어

차이가 있는지를 알아보는 과제별 시스템 효과는 반복측정 일원 분산 분석을 통해 분석되었다. 분석 과정에서 참가자들의 과제에 대한 사전 지식수준은 통제하였다. 표 2에 제시된 두 시스템간의 평균 점수를 비교해 보면, 대부분의 평가항목에서 시스템 2, 즉 정보검색 과제별로 랭킹 요소의 가중치를 달리한 동적 검색 랭킹 모델에서의 모든 사용자 평가 점수들-정보성, 유용성, 만족도, 적합성-이 높은 것을 알 수 있다. 이는 상위 5개 각각을 평가한 일반 누적 점수와 순위를 고려한 누적 점수에서 모두 마찬가지였고, 과제에 대한 통합 점수도 마찬가지였다.

다만 의사 결정 검색 과제에서는 몇몇 항목들 - 적합성 CG 점수, 적합성 DCG 점수, 정보성 DCG 점수, 유용성 DCG 점수 - 에서 정적 검색 랭킹 모델이 적용된 시스템 1의 점수가 약간씩 높은 현상을 보였다.

전체적으로 두 시스템 간 점수 차이가 유의미한지를 나타내는 시스템 주 효과를 보면, 표 16에서 보여지는 것처럼 검색결과 상위 5개 각각의 정보성, 유용성, 만족도, 적합성을 측정한 점수를 누적한 CG 점수, 그리고 순위를 고려하여 누적한 DCG 점수 모두 두 시스템간의 차이가 통계적으로 유의미하게 나타났다. 문서의 누적 점수 외에 전반적인 검색 결과에 대한 평가 즉 “원하는 내용을 알게 되었다”, “검색이 성공적이었다”, “검색 목적을 달성했다”, “검색결과에 만족한다” 또한 모두 시스템 주 효과가 통계적으로 유의미한 것으로 나타났다.

정보 검색 과제별 시스템 주 효과를 보면, 사실 검색 과제에서 시스템 주 효과는 대부분의 평가 항목에서 유의미하였다. 다만, 순위를 고려하지 않은 유용성 CG 점수만이 근소한 차이(p = .05)로 통계적으로 유의미하지 않았다. 문제 해결 검색 과제에서는 정보성, 유용성, 만족도, 적합성 평가 점수를 일반 누적한 CG 점수와 순위를 고려하여 누적한 DCG 점수 모두 통계적으로 유의미한 것으로 나타났다. 그러나 전반적인 검색 결과 평가 점수인 통합 점수에서는 두 시스템간의 차이가 통계적으로 유의미하지 않았다. 의사 결정 검색 과제에서는 대부분의 평가항목에서 두 시스템간의 차이가 통계적으로 유의미하지 않았다. 다만 유용성 DCG 점수만이 통계적으로 유의미한 것으로 나타났다(p < .05). 만족도 DCG 점수는 근소한 차이(p = .05)로 통계적으로 유의미하지 않았다.

과제별 시스템 주 효과를 종합해 보면, 상위 5개 각 문서의 사용자 평가를 누적한 CG 점수는 사실 검색 과제, 문제 해결 검색 과제에서는 모두 동적 검색 검색 랭킹 모델이 적용된 시스템의 사용자 평가가 높은 것으로 나타났고 결과적으로 시스템 주 효과를 확인할 수 있었다. 의사 결정 검색 과제에서는 시스템 주 효과가 유의미하지는 않았지만 직접적인 적합성 평가를 제외한 정보성, 유용성, 만족도에서 시스템 2의 사용자 평가 점수가 모두 높은 것을 확인할 수 있었다. 직접적인 적합성 평가에서는 정적 검색 랭킹 모델을 적용한 시스템 1에 대한 평가 점수(4.71)가 동적 검색 랭킹 모델을 적용한 시스템 2에 대한 평가 점수(4.66)보다 약간 높은 것으로 나타났다. 순위를 고려한 상위 5개 문서의 사용자 평가 누적 DCG 점수는 CG 점수와 마찬가지로 사실 검색 과제, 문제 해결 검색 과제에서 모두 동적 검색 랭킹 모델을 적용한 시스템 2에 대한 평가 점수가 높은 것을 확인할 수 있었고 이 차이는 유의미하였다. 의사 결정 검색 과제의 경우, 만족도를 제외한 정보성, 유용성, 적합성에서 정적 검색 랭킹 모델을 적용한 시스템 1에 대한 평가 점수가 높았다. 만족도는 동적 검색 랭킹 모델의 평가 점수가 높은 것으로 나타났다. 그러나 검색 결과 전반적인 평가 점수는 사실 검색 과제에서만 시스템 주 효과가 통계적으로 유의미한 것으로 나타났다. 그러나 통계적으로 유의미하지는 않았지만 문제 해결 검색, 의사 결정 검색 과제에서도 시스템 2의 사용자 평가 점수가 더 높은 것을 확인할 수 있었다.

Table 2. Descriptive statistics and analysis results of system effect(기술 통계 및 시스템 효과 분석 결과)

score type	evaluation item	search task	System 1 (static search ranking)	System 2 (dynamic search ranking)	System effect	
					per-task	total
			M (*SD*)	*M* (*SD*)	*F*	*F*
Cumula-tive Gain (CG)	informative-ness	fact-finding	4.30 (1.19)	4.76 (0.96)	4.931*	7.012*
		problem-solving	3.55 (0.93)	4.44 (0.80)	11.144**	
		decision-making	4.70 (1.06)	4.77 (0.92)	1.117	

	utility	fact-finding	4.06	(1.13)	4.34	(0.83)	4.007	
		problem-solving	3.20	(0.67)	4.00	(0.86)	9.825**	8.173**
		decision-making	4.48	(1.12)	4.56	(0.95)	3.037	
	satisfaction	fact-finding	3.82	(1.10)	4.13	(0.88)	4.160*	
		problem-solving	3.03	(0.62)	3.72	(0.92)	10.921**	11.113**
		decision-making	4.26	(1.12)	4.46	(0.94)	2.769	
	relevance	fact-finding	4.08	(1.07)	4.37	(0.85)	6.402*	
		problem-solving	3.29	(0.69)	3.99	(0.93)	8.087**	11.597**
		decision-making	4.71	(1.11)	4.66	(0.91)	1.382	
Discou-nted Cumula-tive Gain (DCG)	informative-ness	fact-finding	3.12	(0.90)	3.53	(0.66)	7.212*	
		problem-solving	2.70	(0.68)	3.45	(0.54)	14.817***	9.071**
		decision-making	3.45	(0.73)	3.40	(0.64)	1.553	
	utility	fact-finding	2.95	(0.86)	3.21	(0.57)	6.915*	
		problem-solving	2.49	(0.55)	3.15	(0.63)	10.191**	9.785**
		decision-making	3.28	(0.80)	3.26	(0.67)	4.188*	
	satisfaction	fact-finding	2.76	(0.83)	3.07	(0.62)	7.423**	
		problem-solving	2.36	(0.51)	2.99	(0.68)	12.763***	14.454***
		decision-making	3.10	(0.81)	3.23	(0.64)	3.757	
	relevance	fact-finding	2.95	(0.80)	3.27	(0.63)	8.282**	
		problem-solving	2.53	(0.55)	3.21	(0.65)	12.860***	13.246***
		decision-making	3.43	(0.75)	3.34	(0.62)	1.459	
entire score	knowledge gain	fact-finding	4.93	(1.74)	5.60	(1.18)	8.395**	
		problem-solving	5.27	(1.51)	5.67	(1.09)	3.467	11.882**
		decision-making	5.07	(1.45)	5.33	(1.30)	0.293	
	search sucess	fact-finding	4.47	(1.71)	5.36	(1.46)	19.615***	
		problem-solving	4.40	(1.67)	5.24	(1.43)	1.262	9.548***
		decision-making	4.82	(1.48)	5.18	(1.37)	0.296	
	goal achievement	fact-finding	4.64	(1.80)	5.64	(1.40)	15.747***	
		problem-solving	5.22	(1.59)	5.51	(1.29)	2.993	6.981*
		decision-making	4.84	(1.55)	5.22	(1.48)	2.277	
	search satisfaction	fact-finding	4.36	(1.67)	5.18	(1.42)	13.711***	
		problem-solving	4.09	(1.56)	4.78	(1.57)	1.677	8.488**
		decision-making	4.47	(1.67)	4.91	(1.38)	0.003	

(* p < .05, ** p < .01, *** p < .001)

5.2. 사용자 피드백

참가자가 실제로 검색결과를 어떻게 인지하고 평가했는지를 참가자가 답한 평가 이유를 통해 확인해 보았다. 사실 검색 과제인 신종 플루 사망자 수와 여자 피겨 랭킹에 대해서는 최신성 기준에 대한 언급을 다수 확인할 수 있었다. 최신성에 관한 언급은 다음과 같았다. “현황 나와 있지 않음 예전자료 뿐”, “너무 오래된 문서이다.”, “업데이트 되지 않아 1~2주전 현황만 알 수 있다.”, “보건부 통계자료를 들어 가장 최근의 사망자수를 명시하였고 전염성과 그 추이에 대한 원래의 물음에 대한 대답을 한 문서이기 때문이다”, “역시나 최신의 수치를 제공하고 있지만 가장 최근의 결과는 아니라서 보통 정도의 만족도이다.”, “최신의 수치라기에는 시간이 지난(보름) 문서이다.”, “가장 최근의 발표수치와 그 숫자가 어떻게 나오게 되는지 수

치의 발표는 어떤 의미를 가지는지 정보를 알게 되어 괜찮았다", "원하는 정보에 관련된 새로운 사실 알게 됨", "날짜가 지난 최신 정보 아니다", "시의성이 떨어진다.", "가장 최신의 사망자 현황을 알려주는 최신 문서이다.", "최근 내용 미 기재", "원하는 정보 습득 및 정보의 최신성이 만족스러움", "정보 습득 차원에선 매우 만족이지만 최신 측면에서 다소 불만족", "현황을 보고 싶은 건데 최신성이 많이 떨어짐.", "결과가 너무 오래 전 것임" "과거의 정보만을 알려 주는 거라서 새롭거나 유용한 정보는 아니다.", "문서 작성 일자가 10월 27일전 새로운 사실을 반영하지 못한다", "최근의 결과이다. 점수가 랭킹과 같이 나와 있다", "작성기간이 오래 되서 그랑프리 대회 결과가 반영되지 않아 신뢰성이 조금 떨어짐" 등 이었다.

사실 검색 과제에서는 최신성 외에도 내용의 다양성과 풍부함과 관련된 구체성에 관한 언급들도 확인할 수 있었다. "국내뿐 아니라 국제 현황과 함께 주변 정보들도 함께 얻을 수 있어서", "수치에 대한 정보뿐만 아니라 이러한 통계 정보가 어떤 경로로 산출되는지에 대한 정보를 주어서 좋았다", "신종플루의 사망자 현황뿐만 아니라 전염성의 내용도 추가로 알려주었다.", "알지 못했던 의외의 결과를 알 수 있다", "국내 + 국외 정보까지", "원하는 총 사망자수와 그 외 정보도 추출함", "자세한 통계, 구체적 수치", "대체로 만족. 기간별 통계 잘 나와있고 부가적인 코멘트나 설명 제시", "여러 자료(사망자 개인의 특징 초점, 전체적 사망자 현황, 세계적 추이, 부가 정보)" 등 이었다.

문제 해결 검색 과제와 의사 결정 검색 과제에서는 이해가능성과 관련된 이미지에 대한 언급을 확인할 수 있었다. 문제 해결 검색 과제에서는 "용어의 설명과 유래가 자세히 되어 있고 예시가 많아 이해가 쉬움", "회귀분석의 유래와 회귀분석에 대한 쉬운 예가 제시되어 있어서 좋았다", "잘 모르는 사람이 GOMS가 무엇인지 알게 해주고 거기에 대한 심화적 내용도 접근하기 쉬운 구성으로 되어 있다", "그림 자료 등을 검색할 수 있었으면 좋았을 것 같다", "용어의 설명과 유래가 자세히 되어 있고 예시가 많아 이해가 쉬움", "좀 더 덜 복잡하게 정리되어 있다. 개념 이해에 도움이 된다." 등의 이유가 언급되어 있었다. 의사 결정 검색 과제에서는 "개인 영화 리뷰로 다양한 정보와 이미지들이 풍부하여 만족스러움.", "개인 영화 리뷰로 스틸 이미지가 부족하지만 영화와 관련된 설정과 이슈 정보들을 제공함으로써 새로운 정보 습득 측면에서 만족스러움", "스틸 이미지에 대한 기대가 컸으나, 텍스트 위주의 개인적인 나레이션 리뷰가 노출됨으로써 전체적인 만족도가 조금 떨어짐", "재밌게 구성된 내용. 사진 좋음", "텍스트가 너무 많아 읽지 않게 됨", "사진과 함께 자세한 설명 감상", "만화도 있음" 등이다.

그러나 의사 결정 검색 과제에서는 영화 내용을 미리 알려주어서 만족도가 떨어진다는 내용들이 상당수 확인되었다. "자세한 리뷰이지만 스포일", "자세하고 그림 많았는데 미리 영화 내용을 알려줘서 불만족", "영화에 대해 매우 전문적인 글이다(심지어 오마쥬 분석까지). 그러나 평가도 좋지만 내용을 다 알게 되어 버렸다", "내용 파악은 가능했다. 하지만 스포를 당하면 영화 보기가 싫다. 그래서 만족하지 않는다ㅠㅠ", "영화를 안 본 사람에게는 스포일러가 될 수 있는 내용", "스포일러까지 포함한 꽤 상세한 리뷰 문서이다.", "영화를 보고 나서 읽었으면 좋았을 글이다", "스포일러가 포함되어 있어 약간 실망", "스포일러라 흥미를 깰 수 있다", "결말이 다 나왔다. 스포" 등이었다. 이들은 모두 동적 검색 랭킹 모델이 적용된 시스템 2에서 확인된 사용자 피드백이었다. 이와 관련해서 의사 결정 검색 과제에서 시스템 주 효과가 유의미하지 않은 이유를 이미지가 많고 내용이 자세하긴 했지만 영화 내용을 미리 알려주어 만족도가 떨어진다는 것으로 추정할 수 있었다.

6. 결론

본 연구는 정보 검색 과제별 주요 적합성 판단 기준을 실제 정보 검색 시스템의 알고리즘에 반영하여 동적 검색 랭킹 모델로 구현하여 그 효과를 검증해 보는 것이었다. 이를 위해 모든 검색어에 동일하게 반응하는 정적 검색 랭킹 모델을 적용한 시스템과 정보 검색 과제별로 사용자 적합성 판단 기준의 변화에 따라 랭킹 요소 가중치를 달리한 동적 검색 랭킹 모델을 적용한 시스템을 준비하였다. 동적 검색 랭킹 모델은 정적 검색 랭킹 모델에 비해, 사실 검색 과제에서는 최신성 가중치(0.5 vs 1)와 구체성 가중치(0.05 vs 0.1)가 높이 반영되었고, 문제 해결 검색 과제에서는 반대로 최신성 가중치는 떨어진 반면(0.5 vs 0), 이해가능성과 관련한 이미지/동영상 가중치의 비중이 높이 반영되었다(0.1 vs 0.3). 의사 결정 검색 과제에서는 최신성 가중치(0.5 vs 1)와 이해가능성 가중치(0.1 vs 0.3), 그리고 구체성 가중치(0.05 vs0.1)가 모두 조금씩 높게 반영되었다. 사용자 평가 점수는 검색결과 상위

5개의 각 문서 별 인지 적합성(정보성), 상황 적합성(유용성), 정서 적합성(만족도)와 함께 직접적인 적합성 항목도 측정하였다. 이와 함께 전반적인 검색 결과 평가 항목도 측정하였다.

연구 결과, 전체적으로 동적 검색 랭킹 모델을 적용한 정보 검색 시스템에서의 사용자 평가 점수가 전반적인 항목에서 높은 것으로 나타났고, 이 차이는 통계적으로 유의미하였다. 정보 검색 과제별로는, 사실 검색 과제에서와 문제 해결 검색 과제에서는 전반적으로 동적 검색 랭킹 모델을 적용한 정보 검색 시스템에 대한 사용자 평가 점수가 높았고, 차이 또한 유의미한 것으로 나타났다. 그러나 의사 결정 검색 과제에서는 두 시스템간의 차이가 유의미하지 않은 것으로 나타났다. 그 이유로는 내용이 너무 자세하면서 영화 내용을 미리 알 수 있는 스포일러 내용을 포함하고 있어서 불만족스럽다는 참가자 피드백에서 이유를 미루어 짐작해 볼 수 있었다. 결과적으로 본 연구의 결과는 사용자가 정보 검색 과제 별로 적합성을 판단하는 기준이 달라진다는 앞선 연구 결과를 지지하는 것으로 나타났고, 이는 정보 검색 과제 별 동적 검색 랭킹 모델의 필요성을 제안하는 것으로 볼 수 있다.

본 연구는 다음과 같은 실용적 의의를 가진다.

첫째, 사용자 연구 접목을 통한 시스템 구현 및 개선 가능성을 보여 주었다는 점에서 실용적 의의를 가진다. 본 연구는 사용자 적합성에 관한 연구 결과를 실제로 정보 검색 시스템의 검색 랭킹 모델로 구현하여 효과를 검증하였다. 이를 통해 사용자 연구 결과의 타당성을 검증함과 동시에 사용자 연구를 접목한 시스템 성능 개선의 중요성을 시사한다는 점에서 의의를 가진다. 일반적으로 검색 랭킹 모델은 휴리스틱이나 기계 학습에 기반한 랭킹 요소를 추출하여 사용하고 있는 실정이다. 본 연구는 Park과 Sohn(2009)의 사용자 적합성 판단 모형에 관한 연구에서 밝혀진 사용자 적합성 판단 기준들을 랭킹 요소로 활용하였다. 사용자 연구와 시스템 연구의 시너지 발생은 매우 중요한 것으로, 향후 사용자 연구를 통해 궁극적으로 사용자를 보다 만족시킬 수 있는 정보 검색 시스템 개선으로 이어질 수 있도록 이와 같은 접근은 더욱 적극적으로 장려되어야 할 것이다.

둘째, 정보 검색 과제 별 사용자 판단 기준 변화를 고려한 정보 검색 시스템 디자인이 필요하다는 실용적 시사점을 제공한다. Schamber(1990)는 사용자는 상황에 따라 동적으로 적합성을 인지한다고 하였다. Limberg(1999), Kelly 등(2002), 그리고 Tombros, Ruthven과 Jose(2005)의 연구 결과들은 정보 검색 과제에 따라 적합성 평가에 사용되는 특징이나 문서 요소가 다르다는 것을 알려주었다. Park과 Sohn(2009) 또한 사용자 적합성 판단 모형을 통해 정보 검색 과제별로 사용자 적합성 판단 기준이 달라진다는 것을 밝혀내었다. 실제로 본 연구에서는 정보 검색 과제별 동적 검색 랭킹 모델을 구현하여 적용해 봄으로써 사용자 평가가 향상되는 것을 보여주기도 하였다.

셋째, 검색어 별 동적 랭킹 모델을 위한 검색어 분류의 세분화 기준을 제공하였다는 점에서 실용적 의의가 있을 수 있다. 최근 연구들에서 검색어 별 동적 검색 랭킹 모델의 필요성은 언급되고 있으나 구체적으로 검색어가 어떤 식으로 분류되어야 하는지는 명확하게 밝혀져 있지 않다. 또한 실제로 검색 랭킹 모델을 제안한 연구는 더욱 찾기 힘들다. 본 연구는 검색 목적에 따른 정보 검색 과제별로 적합성 판단 기준이 달라진다는 선행 연구(Park et al., 2009) 결과를 실제 정보 검색 과제별 동적 검색 랭킹 모델로 구현하여 효과를 실증적으로 입증함으로써, 정보 검색 과제 별 분류가 동적 검색 랭킹 모델 적용의 기준이 될 수 있다는 가능성을 보여주었다.

그러나 본 연구에서는 사용자 적합성 판단 기준을 정량화하여 문서의 특징을 이용하여 랭킹 요소로 반영하였지만, 이에 대한 검증은 연구 범위에서 해당되지 않기 때문에 본 연구에서 이루어지지 않았다. 사용자 적합성 판단 기준과 문서 요소와의 연결은 관심을 가지고 연구된 바 있으나 아직 명확하게 밝혀지지 않았다(Tombros, Ruthven & Jose, 2005). 향후 사용자 적합성 판단 기준과 문서 요소의 연결에 관한 지속적인 연구를 통해, 사용자 적합성 연구를 시스템 개선에 적극 활용할 수 있을 것이다.

정보 검색에서 적합성은 사용자와 시스템의 중요한 상호작용이다(Saracevic, 2007). 본 연구는 이러한 적합성을 사용자 중심으로 접근하여 사용자 적합성 판단 모형 기반으로 실제 시스템으로 연결해 봄으로써 사용자 적합성에 대한 이해를 넓히고 이러한 접근을 통한 시스템 개선 가능성을 제안하였다. 사용자와 시스템의 상호작용에는 적합성 외에도 인터페이스도 중요한 영향을 미치는 요소이다. 향후 연구는 연구 범위를 보다 확장하여 인터페이스와 같은 외적인 요소에 대한 고려도 필요할 것이다. 그리고 적합성과 인터페이스와의 영향 관계나 상호작용과 같은 보다 확장된 사용자 중심의 적합성 연구가 필요하겠다. 이러한 노력

을 통해 궁극적으로 사용자의 경험을 향상시키는 정보 검색 시스템 개선이 이루어질 수 있을 것이다.

REFERENCES

Agichtein, E., Brill, E., & Dumais, S.T. (2006). Improving Web search ranking by incorporating user behavior. *In Proceedings of the 29th Annual International ACM SIGIR '06,* 19-26.

Baeza-Yates. R. & B. Ribeiro-Neto. (1999). *Modern information retrieval*. Addison-Wesley.

Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science, 45,* 149-159.

Barry, C. L. & Schamber, L. (1998). User-defined relevance criteria: A comparison of two studies. *Proceedings of the American Society for Information Science*, Chicago, IL, 103-111. Medford, NJ : InformationToday, Inc.

Bateman, J. (1998). Changes in relevance criteria: A longitudinal study. In: ASIS Proceedings. 1998, 23-32.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology, 54(10),* 913-925.

Choi, Y. & Rasmussen, E. M. (2002). Users' relevance criteria in image retrieval in American history. *Information Processing & Management, 38,* 695-726.

Clarke, C.L.A., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher S., MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. *In Proceedings of the 31th Annual International ACM SIGIR '08*, 659-666.

Cosijn, E. & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management, 36,* 533-550.

Cuadra, C. A. & Katter, R.V. (1967). Opening the black box of relevance. *Journal of Documentation , 23,* 291-303.

Cutler, M., Shih, Y., & Meng, W. (1997). Using the structures of html documents to improveretrieval. In Proceedings of the USENIX Symposiumon Internet Technologies and Systems, 241–251.

Drori, O. (2002). Algorithm for documents ranking: Idea and simulation results. In Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering. *ACM International Conference Proceeding Series, 27,* 99–102.

Fitzgerald, M. A. & Galloway, C. (2001). Relevance judging, evaluation, and decision making in virtual library: A descriptive study. *Journal of the American Society for Information Science and Technology, 52,* 989-1010.

Geng, X., Liu, T.-Y., Qin, T., Arnold, A., Li, H., & Shum, H.-Y. (2008). Query dependent ranking using K-nearest neighbor. *In Proceedings of the 31th Annual International ACM SIGIR '08,* 115-122

Hirsh, S. G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science, 50,* 1265-1283.

Ingwersen, P. & Järvelin, K. (2005). The turn: Integration of information seeking and retrieval in context. Dordrecht, The Netherlands: Springer.

Järvelin, K. & Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems, 20,* 422–446.

Joachims, T. (2002). Optimizing search engines using clickthrough data. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 133-142.

Kang, I. H. & Kim, G.. C. (2003). Query type classification for web document retrieval. *In Proceedings of the 26th annual international ACM SIGIR '03,* 64–71.

Kekäläinen, J. & Järvelin, K. (2002). Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In Bruce, H., Fidel, R., Ingwersen, P. & Vakkari, P., eds. Emerging Frameworks and Methods, Seattle, 2002. Colerado: Libraries Unlimited, 253-270.

Kelly, D., Murdock, V., Yuan, X., Croft, W. B., & Belkin, N. J. (2002). Features of documents relevant to task and fact-oriented questions. *In Proceedings of the Eleventh International Conference*

on Information and Knowledge Management, 645-647. New York, NY: ACM.

Limberg, L. (1999). Experiencing information seeking and learning: a study of the interaction between two phenomena. *Information Research, 5(1)*. Available at: http://informationr.net/ir/5-1/paper68.html

Maglaughlin, K. L. & Sonnewald, H. (2002). User perspective on relevance criteria: A comparison among relevant, partially relevant, and not-relevant. *Journal of the American Society for Information Science and Technology, 53,* 327-342.

Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science, 48,* 810-832.

Park, J. A. & Sohn, Y. W., (2009). User-oriented judgment model in information retrieval. *Korean Journal of the Science of Emotion and Sensibility, 12(4)*, 489-500.

Park, T. K. (1993). The nature of relevance in information retrieval: *An empirical study. Library Quarterly, 63,* 318-351.

Robertson, S. (1997). Overview of the okapi projects. *Journal of Documentation, 53(1),* 3-7.

Rees, A. M. & Schultz, D. G., (1967). A field experiment approach to the study of relevance assessments in relation to document searching, 2 vols. Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, Cleveland, OH.

Salton, G. (1971). The SMART Retrieval System-Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Salton, G. & Lesk, M.E. (1968). Comparative evaluation of indexing and text processing. *Journal of the Association for Computing Machinery, 15*, 8-36.

Saracevic, T. (1970). The concept of relevance in information science: A historical review. In Saracevic, T. (Ed.), *Introduction to information science,* 111-151. New York: R.R. Bowker.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science, 26,* 321-343.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology, 58,* 1915-1933.

Savolainen, R. & Kari, J. (2006). User-defined relevance criteria in web searching. *Journal of Documentation 62(6),* 685-707.

Schamber, L. (1991). Users' criteria for evaluation in a multimedia environment. *In Proceedings of the 54th ASIS Annual Meeting, 28,* 126-133.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing & Management, 34,* 599-621.

Spink, A. & Jansen, B.J. (2004). *Web search: Public searching of the Web.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Su, L. T. (2003). A comprehensive and systematic model of user evaluation of Web search engines II: An evaluation by undergraduates, *Journal of the American Society for Information and Technology, 54,* 1193–1223.

Tang, R. & Solomon, P. (1998). Towards an understanding of the dynamics of relevance judgments: An analysis of one person's search behavior. *Information Processing & Management, 34,* 237-256.

Vaughan, L. (2004). New measurements for search engine evaluation proposed and tested, *Information Processing & Management, 40,* 677–691.

Xu, Y. & Chen, Z. (2006). Relevance judgment - What do information consumers consider beyond topicality? *Journal of the American Society for Information Science and Technology, 57(7),* 961-973.

원고접수: 2012.06.21
수정접수: 2012.09.03
게재확정: 2012.09.03