질의 내부 단어 인접도를 이용한 검색 효율 향상 기법

(A Search Efficiency Improvement Method using Internal Contiguity in Query Terms)

윤 성 웅 † 채 진 기 † 이 상 훈 ††
(Soungwoong Yoon) (Jinki Chae) (Sanghoon Lee)

요 약 수많은 웹 정보 중에서 사용자가 원하는 정보를 찾아내는 것은 매우 어렵다. 검색 엔진은 웹 정보를 요약하였다가 사용자의 질의에 따라 상대적 중요도와 정보의 적합도를 반영한 검색순위를 제공한다. 그러나 이 순위는 개별 사용자가 원하는 정보를 상위 순위에 보여주는데 제한이 있다. 본 논문에서는 사용자의 검색 의도가 질의에 가장 잘 나타난다고 보고 질의의 의미를 잘 반영하는 웹 정보를 선택적으로 상위 순위화하기 위하여 질의 내부의 단어 인접도를 이용한 재순위화 방법을 제시하였다. 실험 결과 매우 간단한 방법으로 사용자가 요구하는 정보를 75.8%의 확률로 찾아낼 수 있으며, 선별된 정보들의 선택적인 순위 상승으로 13~20%의 검색 효율 향상을 기대할 수 있다.

키워드 : 정보 검색, 인접도, 선택적 순위 상승

Abstract It is difficult to get relevant information on vast Web data. Search engines summarize and store Web information and show the ranked lists based on user queries affected by relative importance and user-adaptation. But these have limitation with showing user-intended information at the top priority. User intention is presented in general within query itself. In this paper, we propose the selective rankup methodology of user-intended search results based on weighting internal contiguity in query terms. With experimental results, we can find user-intended results with 75.8% probability using this simple method only, and efficiency of rerank proposed outperforms ordinary case by 13~20%.

Key words : Web Retrieval, Contiguity, Selective Rankup

1. 서 론

웹은 무한한 속도로 발전하여 정보의 바다를 이루고 있다. 이에는 다양한 주제(*Topic*)의 수많은 정보들이 여러 가지 형태로 존재하는데, 이중에서 특정 정보를 직접 찾는 것은 대단히 어려운 문제이다. 따라서 많은 사용자들이 검색 엔진에서 질의(Query)를 이용하여 필요한 정보를 포함하는 웹 정보를 찾아내는 것이 일반적이다.

검색 엔진은 수집자(crawler)를 통해 웹 페이지를 수집, 요약하여 저장하고 저장된 역인덱스에서 고유의 순위 판단 기법을 통해 결과를 선별한다. 사용자가 질의를 입력하면 선별된 결과를 질의에 따른 중요도를 반영하여 재순위화하여 보여준다. 이때 대부분의 사용자는 제시된 결과의 상위에 있는 순위[1]에서 필요한 정보의 존재 여부를 판단하며[1,2], 따라서 검색 결과를 순위화하는 방법은 아주 중요하다.

검색 엔진의 순위 산정 방식으로 대표적인 것은 PageRank가 있다[3]. PageRank는 대중성에 기반한 순위 산정 방식으로 특정 페이지가 다른 많은 페이지로부터 연결되어 있다면 그 페이지를 우수한 대중성을 가진 것으

† 학생회원 : 국방대학교 전산정보학과
ysw1209@gmail.com
nice5556@nate.com
†† 종신회원 : 국방대학교 전산정보학과 교수
hoony@kndu.ac.kr
논문접수 : 2007년 4월 19일
심사완료 : 2008년 1월 4일

정보과학회논문지: 데이타베이스 제35권 제2호(2008.4)

1) 통상 10~20순위 내외의 상위순위

로 판단하는 방법이다. 그러나 사용자가 요구하는 정보는 그 범주를 명확히 규정할 수 없으며, 높은 대중성을 가지는 페이지와 사용자 요구의 연결(matching)은 매우 어려운 문제이다. 또한 정보의 공급자 입장에서 웹 상에 새로 포함된 정보들의 순위를 향상시키기 위해서는 많은 시간이 요구된다[2][4]. 이러한 요구에 따라 PageRank의 순위 산정을 위해 사용자의 검색 성향을 반영하는 방법(Personalized PageRank)[1], 검색의 범주를 한정하여 재순위화하는 방법(Topic-Sensitive PageRank)[5], 사용자의 질의에 대한 반응 기록을 저장하고 활용하는 방법(Query-biased Personalized PageRank)[1] 등 여러 가지 방법이 다양하게 연구되었다. 다른 한편으로는 웹 정보를 인간의 인식 체계에 따라 분류하고자 하는 의도에 따라 페이지를 분류하여 필요한 정보를 분류(filtering)하는 방법이 연구되고 있다[6,7].

검색의 유형에는 검색할 내용을 미리 알고(생각하고) 검색하는 경우(물품 구매, 특정 정보 찾기)와 자주 사용하는 검색 도구에서 검색하고자 하는 의도부터 시작하여 연관되는 정보를 검색하는 경우(뉴스, 시사 등)가 있다[3][8]. 이소영 등[9]에 따르면 후자의 경우가 더욱 많은 것으로 나타나고 있으며, 이는 더욱 다양한 방식의 범주별 분류를 요구하거나[4] 다수 사용자가 요구하는 정보 외의 검색 개념을 요구한다. 따라서 검색의 유형으로 본 사용자의 검색 의도는 질의를 분석하는 데서 출발하여야 한다. 사용자의 검색 의도는 자연어 검색 시에는 질의에 가장 잘 반영되어 있으며[6], 간접적으로 검색 결과의 사용자별 이용자료(user click-through data)를 분석함으로써 확률적인 파악이 가능하다. 그러나 이 자료를 수집하기 위해 필수적인 사용자 식별 절차(로그인, IP 점검 등)는 사용자에게 불편함을 야기할 수 있다.

질의 방법에는 여러 가지가 있으며, 가장 간단하게는 단일 단어 질의와 문장(Sentence) 또는 단어구(Phrase)로 구성된 복합 질의로 구분된다. 단일 단어 질의 검색에 있어 효율성을 향상시키기 위한 많은 방법이 제시되었는데, 질의 자체를 분석하는 방법과 추가적인 질의를 요구하는 방법[2], WordNet[10]을 이용하여 질의를 분석하고 적절한 정보를 추가하여 검색하는 방법[11,12] 등이 연구되었으며 특히 WordNet을 이용하는 방법은 질의의 중의성이 강할 때 유용한 결과를 나타내었다.

그러나 검색은 단일 단어로 이루어진 경우보다는 여러 의미를 가진 단어의 집합인 복합 질의가 일반적이다. 복합 질의 검색은 보다 분명한 사용자 의도를 파악하기 위하여 단어 사이의 관계를 이용함으로써 단일 단어 질의에서 연구된 관점보다 검색 효율 향상을 꾀할 수 있다. 현재 사용되고 있는 복합 질의 검색 결과는 사용자에게 적절하지 않은 페이지가 상위 순위에 분포하는 경우가 많으며, 이를 해결하기 위하여 입력된 질의와 색인된 정보의 색인어 공유 정도를 판단하여 검색 순위에 영향을 미치고자 하는 연구도 진행되었다[6].

본 논문에서는 사용자의 의도를 가장 잘 반영하고 있는 질의에 집중하여, 질의를 구성하는 단어구 내부 인접도에 따라 검색 정보를 재순위화하는 방법을 제안하였다. 이는 기존 검색 방법에서 찾아내기 어려운 인지도가 낮은 정보나 새롭게 포함된 정보를 상대적으로 상위 순위로 위치시키는데 도움이 된다.

본 논문의 구성은 2장에서 기존 연구를 살펴보고 3장에서 질의 내부의 인접도를 표현하는 방법론을 제시하며, 4장에서는 실험을 통하여 인접도를 적용한 성과를 보였다. 5장에서는 결론 및 향후 연구 과제를 서술하였다.

2. 관련 연구

검색 엔진에서 순위를 결정하기 위해 사용하는 다양한 기법중에서 PageRank와 관련 순위화 기법의 발전 내용을 살펴보고, 순위를 결정하는 데 영향을 미치는 요인들을 도출한다. 앞으로 사용될 공통적인 범례는 다음과 같다.

n : 웹페이지 전체의 수

m : 고려하는 모든 주제 t의 수

T(i) : i번째 주제 (Topic)

2.1 PageRank

PageRank(PR)의 기본 개념은 어떠한 페이지 p_0가 다른 페이지 p로 연결(outlink)되어 있다면 p_0의 작성자는 p가 중요하다고 생각한다는 것이다. 즉 얼마나 많은 페이지가 연결(inlink)되어 있는가의 여부가 그 페이지의 중요도를 결정한다는 것이다[3]. 아래 식은 PageRank를 계산하기 위한 것이다.

$$PR(p) = d * \sum_{p_0 \in A_p} PR(p_0)/l_{p_0} + (1-d) * E(p)$$

d : 사용자가 임의의 연결을 선택할 확률

1-d : 연결을 선택하지 않고 임의의 다른 페이지로 바로 옮겨갈(Random Jump) 확률

Ap : p로 연결되어 있는 전체 페이지 수

lp_0 : p_0로부터의 외부 연결(Out-link) 개수

E(p) : Random jump 확률 값(Probability vector), 수학적으로는 1/n

PR(p)는 페이지의 순위 값을 나타내므로 n개의 페이

2) Search-dominant model[1]에서는 사용자에게 약 1,650회 노출되면 대중성이 급격히 상승하는 현상을 보였다.

3) 사용자는 질의의 모든 의미들에 대해 정확히 알고 있는 경우가 드물며, 검색 질의로 사용한 단일 어휘가 올바른 의미로 판단하기 힘든 경우도 빈번하게 발생된다.

4) 검색 포탈에서는 웹페이지, 지식, 비디오, 블로그 등 서비스 가능한 범주별 순위를 제시하여 검색의 효율을 높이고 있다.

지에 대해 각각의 순위 값이 존재하며, 이 값은 웹 페이지를 종합, 외부 연결수로 계산하게 되므로 웹 정보를 종합하는 시기의 차이에 따라 다른 값을 가지게 되나 질의와는 독립적으로 단독적 순위 값(single global score)을 가지게 된다.

2.2 Topic-Sensitive PageRank

PageRank의 확장으로서 상이한 질의에 대한 올바른 순위를 제공하기 위하여 각각의 주제에 대한 페이지별 순위 값(TSPR score)을 산정한다[5]. 아래 식은 TSPR을 계산하기 위한 것이다.

$$TSPR_t(p) = d * \sum_{p_0 \in A_p} TSPR_t(p_0)/l_{p_0} + (1-d) * E_t(p)$$

$E_t(p)$: t번째 주제에 영향을 받은 Random jump 확률

각 페이지는 m가지 주제에 대한 m개의 TSPR 순위 값 집합을 가지며, 사용자의 질의에 대해 각 주제별 근접도에 따른 TSPR 순위 값을 이용하여 순위를 부여한다. 현재의 검색 엔진은 범주의 구분과 사용자들의 검색 결과 축적 및 분석을 통하여 가장 적절한 검색 결과를 산출하기 위한 다양한 기법들을 연구하여 적용하고 있으므로[5] 적어도 PageRank를 사용하는 구글에서는 이 방법에 근접하여 순위가 산정되고 있다고 가정할 수 있다.

2.3 Query-Biased Personalized Topic-Sensitive PageRank

사용자의 성향을 분석하여 PageRank에 반영한 Query-Biased Personalized Topic-Sensitive PageRank (PPR) 연구에 따른 순위값의 산정 방법은 아래 식과 같다[1].

$$PPR_T(p) = \sum_{t=1}^{m} \Pr(T(i) \mid q) \cdot TSPR_t(p)$$

$$= \sum_{t=1}^{m} T(i) \cdot \Pr(q \mid T(i)) \cdot TSPR_t(p)$$

T : 주제 선호도 벡터 (Topic preference vector)

Pr(T(i)|q) : 사용자의 질의 q 입력시 주제 T(i)와 관련될 확률

T(i) : i번째 주제에 대한 개인화된 선호도 벡터, 개인에게는 i번째 주제에 대한 선택확률 Pr(T(i))과 같다.

Pr(q|T(i)) : 사용자가 주제 T(i)에 대해 관심이 있을 때 질의 q를 입력할 확률

이전 기법에 비하여 PPR이 좋은 검색 효율을 보인 것으로 연구되었으며, 현재 이 방법을 적용한 검색 엔진의 존재에 대해서는 알려진 바 없다[6]. T(i)는 개인에게는 i번째 주제에 대한 선택 확률, 즉 Pr(T(i))와 같은 값이며, ODP[13]를 이용하여 이 값의 최초 결과를 받았다.

3. 단어 인접도의 산정

단어군으로 구성된 질의의 순위 산정에 인접도를 결정 요소로 부여하기 위하여 각 단어의 순위를 먼저 산정하고 단어 인접도에 대한 가중치를 부여하여 최종 순위를 산정하고자 한다. 인접도를 통하여 정보의 사용자 적합도를 평가할 수 있다면, 하위 순위에 위치한 사용자 적합도가 높은 페이지를 상위에 위치시키기 위한 방법으로서 질의 내부의 인접도를 이용하는 것은 매우 간단한 연산으로 선택적 순위 상승을 기대할 수 있다.

q = {q_1, q_2, ..., q_n}, 즉 n개의 단어로 이루어진 복합 질의 q를 이용한 검색 순위를 산정할 경우를 가정하여 보면, 이 단어들은 중요도와 관계없이 q라는 하나의 어구 내에 인접하여 있으므로 m가지 주제에 대한 인접도를 반영하여 검증하기 위해서는 m×n!회 만큼의 비교가 필요하다.

n개의 단어 중에서 사용자의 의도를 대표하는 것을 2~3개의 단어로 요약하고 이를 중요 단어 집합(Core word set)이라 한다면, 이는 문장에서 중요한 개념을 가진 단어를 찾아내는 것과 같은 방법으로서 사용자의 질의 입력이 보통 2단어 정도라는 선행 연구[14]와도 상통하는 것이다. 이때 다른 단어들은 이 단어 집합의 의미적 중요성(Semantic importance)을 보완하기 위한 역할을 수행한다고 할 수 있으며, 이때에도 m×(n-3)! ~ m×(n-2)! 회의 비교가 필요하게 된다.

먼저 질의 q가 {q_1, q_2}, 즉 2개의 단어로 이루어진 단어구인 경우를 살펴보자. 예외적으로 단어구 전체의 의미가 함축적인 경우도 존재하지만, 일반적으로는 의미상 주된 기능을 가지는 (또는 요구하는) 단어인 q_1과 q_1에 부가하여 의미를 분명하게 하는 단어인 q_2로 구분할 수 있는데, 이때 복합 질의 q의 주제 선호도는 다음과 같이 정리될 수 있다.

관찰 : 복합 질의 q는 q_1에 q_2의 의미가 덧붙여져 단일한 주제로 수렴되었다고 할 수 있다. 이때 q_1은 단일 주제에 대한 PPR로, q_2는 다중 주제에 대한 PPR로 산정할 수 있다.

이때 각각의 PPR은 다음과 같이 정리된다.

$PPR_{q_1}(p) = TSPR_{T_k}(p)$ --- 단일 주제

$PPR_{T_k,q_2}(p) = \Pr(T_k \mid q_2) \cdot TSPR_{T_k}(p)$ -- 다중 주제

q_2의 주제 벡터를 산정할 때는 q_1에 종속되지는 않지만 q_1을 기반으로 한 주제를 주로 사용하게 되며, 다른 주제에 대한 선호도는 보편적인 경우와 다르게 선택되지 않을 확률이 매우 높게 된다. (즉 0과 매우 가까운

5) 야후 SoftBot 5.0i, 다음 Talkro IR, 엠파스 KONAN Docruzer 등

6) 구글은 개인화된 검색 서비스의 베타테스트를 시작하였으나 구체적인 검색 알고리즘은 알려지지 않았다.

값을 갖는다.) 이에 따라 PPRq2는 다음과 같이 정리할 수 있다.

$$PPR_{q_2}(p) = \Pr((T_k \mid q_1) \mid q_2) \cdot TSPR_{T_k}(p)$$

베이즈 정리(Bayes' Theorem)에 의하면 $\Pr((T_k|q_1)|q_2)$는 다음과 같으며, 주제 T_k에 대한 q_1의 선택 확률은 단일 주제라는 조건에 의하여 1이 된다.

$$\Pr((T_k \mid q_1) \mid q_2) = \frac{\Pr(T_k \mid q_1) * \Pr(q_2 \mid (T_k \mid q_1))}{\Pr(q_2)}$$

$$\propto \Pr(q_2 \mid (T_k \mid q_1))$$

PPRq는 q_1의 순위 값에 가중치 α를 부여한 q_2의 순위값을 더한 값에 비례한다. q_1이 지배적(dominant)인 순위 값이므로 가중치 α가 작을수록 q_2가 가지는 중요도는 낮아지고 단일 주제에 근접하는 순위 값을 가지게 되며, α가 클수록 q_1의 중의성이 높고 q_2가 가지는 결정 요소로서의 중요도가 높아진다고 할 수 있다.

$$PPR_q(p) \propto PPR_{q_1}(p) + \alpha \cdot PPR_{q_2}(p)$$

$$= (1 + \alpha * \Pr(q_2 \mid (T_k \mid q_1))) * TSPR_{T_k}(p)$$

여기서 $\Pr(q_2|(T_k|q_1))$은 인접도 가중치로서 사용자가 q_1에 의해 특성화된(biased) 주제 T_k에 대하여 q_2를 입력할 확률로 정의된다. 이 값은 기본적으로 q_1의 검색 결과 전체에서 q_2의 출현 빈도를 이용하여 산정할 수 있으나, 이는 q_1의 중의성과 주제 선호 방향을 염두에 두어야 하므로 매우 많은 자료를 요구한다. 따라서 대안으로 질의 q를 이용한 검색시 q_1이 단일 주제가 될 수 있도록 도운 q_2의 역할을 이용하였다.

검색 결과에서 q_1이 단일 의미를 가지는 경우는 q_2와의 복합 검색 결과(즉 q의 검색 결과)에서 사용자가 원하는 정보인 경우(적합도가 높은 정보의 집합)가 해당되며, q_2의 역할은 이 정보 집합에서 q_2가 출현하는 빈도를 측정하여 평가할 수 있다. 다시 말해서 단일 주제 q_1에 기반한 q_2의 선택 확률은 질의 q의 검색 결과중 q_1 적합 그룹 내 q_2의 출현 빈도와 비례한다. 이 빈도가 높다는 것은 인접도가 질의의 의미를 잘 반영한다는 것을 뜻하며, 4장에서 실험을 통하여 인접도가 평균적으로 정보의 중요도에 미치는 영향을 보이고자 한다.

4. 실 험

3장에서 제시된 인접도 가중치를 산정하고 그 결과를 측정하기 위하여 실험을 진행하였다.

4.1 질의 선정

웹 페이지 순위의 가중치 영향 정도를 실험하기 위하여 2단어로 구성된 40개의 2단어구를 선정하였으며, 이 2단어구는 가급적 약자를 배제하고 단일 단어로는 중의성을 가지지만 2단어구의 연관성을 통해 중의성이 해결되는 21개의 2단어구[7]와 이를 고려하지 않은 19개의 2단어구로 분류하였고, 중의성 고려 2단어구의 의미 구분은 WordNet을 이용하였다[7]. 중의성 고려 단어구는 검색 결과 비교를 위하여 쌍으로 이루어져 있다.

검색 자료는 2007년 8월 9일 각 질의별 100순위까지 자료를 수집하였으며, 검색 엔진은 구글을 이용하였다.

검색 자료 분류는 가정된 범위로 그룹화하였는데, 두 단어가 순서대로 인접한 경우(A그룹), 순서대로이나 인접하지는 않은 경우(B그룹), 역순으로 인접한 경우(C그룹), 역순이나 인접하지는 않은 경우(D그룹), 기타의 경우(E그룹)로 나누었고 그룹별 순위 분포 및 평균 순위를 계산하였다. 이때 검색결과에서 정보 제목을 그룹 구분의 기준으로 하였으며 (내용은 중요 페이지 식별에만 사용), 문법 오류를 감안하지 않고 오직 두 단어의 근접 여부만을 고려하였다. 정보 제목에 두 가지 이상의 그룹이 공존할 경우에는 A그룹에 가까운 그룹으로 명시했다.

적합 정보 선별에는 정보의 적합률(Precision)와 권위도(Authority)를 이용하였는데[15], 이를 위하여 전산학 전공 대학원생 7명이 검색 결과를 보고 질의에 적합한 정보를 선별하고 선별된 적합 정보 중에서 1위부터 10위까지는 순위를 부여하도록 하였으며, 최종적인 적합 정보 식별에는 각 순위 정보별 선별수를 누적(majority vote)하여 산정하였다.

여기서 인접도가 최초 검색 엔진 순위상에 미치는 영향을 관찰하기 위해서 표 3과 같이 각 순위 정보의 사용자 적합 정도를 별도로 정리하였는데, 이를 중요도로

표 1 중의성 고려 단어구[8]

질의	의미	질의	의미	질의	의미
Java island	섬	Mass data	대량	Engagement rule	전투
Java language	언어	foreign Capital	자본	social Engagement	약속
Custom tax	통관	Capital seoul	수도	power Plant	공장
Custom tailor	관습	tennis Court	코트	water Plant	식물
head Coach	코치	civil Court	법정	christma Seal	봉인
Coach train	객차	Turkey trip	국가	Seal hunter	물개
Mass hall	대중	thanksgiving Turkey	칠면조	SEAL team	특수부대 (중의)

7) 중의성이 있는 단어의 선정은 결과값의 비교를 위하여 선행연구에서 인용하였으며, 검색도구 구현과 WordNet의 이용을 위해 영문 단어로 선택하였다.

8) 의미는 중의성이 있는 단어의 단어군으로서의 의미로 주된 의미에 해당하는 단어는 대문자로 표시하였다.

표 2 중의성 미고려 단어구[9)]

질의	의미	질의	의미	질의	의미
Computer science	학문	Cheap Ticket	상호 종속	toilet Paper	종류/ 가격
HP Computer	제조 회사	birthday Party	종류	Mobile phone	종류/ 가격
XML query	XML	Pizza Delivery	시간/ 위치	fruit Knife	종류/ 가격
query language	query	test Drive	종류	salad Dressing	정확한 의미
Meta data	메타	Stock investment	역순시 순위	Flash Game	프로 그램
Data mining	마이닝	Windows Explorer	상세 기능		
King Crab	갑각류	contact Lense	종류/ 가격		

표 3 중요도 산정 기준

식별수	7~6회	5~2회	1회 이하
중요도(값)	중요 (2)	보통 (1)	관련적음(0)

정의하였다. 중요도의 산정은 권위도 정보를 검색 엔진과 같이 주어진 순위 전체를 대상으로 해석해야 하는 어려움을 극복하기 위한 것이다.

4.2 분석 결과

각 검색 페이지의 분석 결과는 다음과 같다.

- 2단어 검색의 경우 전반적으로 사용자 의도에 적절한 정보를 제공하였으며,[10)] 특히 중의성 고려 그룹에서는 사용자 의도와 다른 검색 결과가 나타나는 경우가 매우 적었다. 시사성이 강한 문제(Java island의 지진 관련 언론 기사, 교황의 turkey trip 등)는 사용자의 의도와 PageRank 순위를 연관시키기 어려우므로 통계적 접근이 곤란하나 일반적인 경우 매우 정확한 검색 결과를 보였으며, 검색 엔진은 관련어를 통한 검색 결과를 함께 보여주었다.[11)]
- 그림과 같이 그룹 분류는 A그룹(순서대로 근접)과 E그룹(기타)으로 대별할 수 있었으며, B/C/D그룹은 상대적인 연결수가 매우 적어(7.7%) 어느 그룹에 포함되어도 결과에 미치는 영향이 근소하게 보이나 A그룹에 포함되는 것이 검색 경향상 바람직할 것으로 판단된다.[12)]

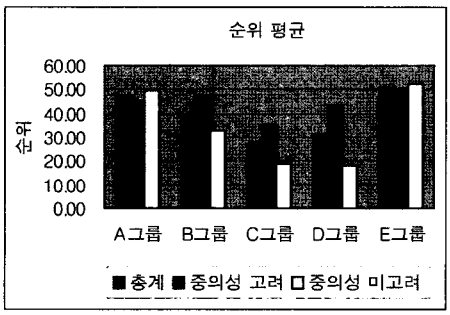


그림 1 순위 평균

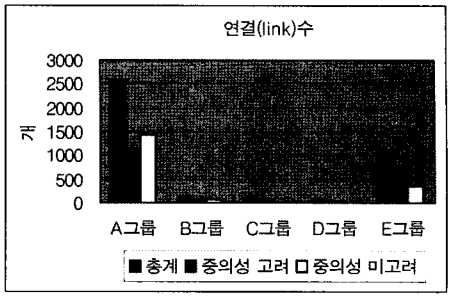


그림 2 연결수

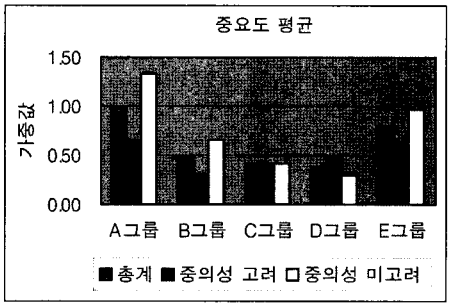


그림 3 중요도 평균

A그룹과 E그룹의 비교 결과 순위 평균(그림 1)으로는 차이가 근소하였다. 이는 검색 엔진의 2단어 검색 결과는 검색 순위 분포상으로는 별다른 차이를 보이지 않는다는 의미로서 100순위내의 분포로는 2단어 검색의 효율성을 찾을 수 없다는 의미이다.

그러나 관련 연결 수(그림 2)와 중요도 평균(그림 3), 10순위까지의 권위도(그림 4)로 비교할 때는 A그룹이 E그룹보다 각각 37%, 23%, 49% 높은 결과를 보였다. 연결수의 차이는 검색 엔진이 검색 결과를 찾을 때 2단어의 복합된 주제를 사용자에게 제공하기 위하여 단어 간의 근접 정도에 비중을 두었다고 할 수 있

9) 검색의 주된 의도를 나타내는 단어를 대문자로 표시하였으며 동일 비중인 단어군은 모두 대문자로 표시하였다. 의미란에는 검색 목표 정보를 나타내었다.

10) 사용자 검색 의도에 따라 결과에 대한 해석이 달라질 수 있으나, 이 경우에도 각 그룹 사이의 차이는 여전할 것으로 보인다.

11) Mass data (mass storage), turkey trip (tour, travel), water plant (aquatic plant), XML query (XQuery, XQL), pizza delivery (papa jones, domino 등 상표), fruit knife (hankel 등 상표) 등

12) A그룹과 동일한 비중의 B/C/D그룹(예 : Engagement rule, java island, capital seoul)의 경우가 다수 출현하였으며, 영문법 특성을 고려할 때도 A그룹에 포함하는 것이 타당하다.

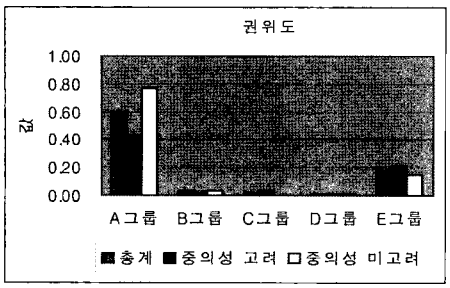


그림 4 권위도

으며, 중요도의 차이는 질의의 2단어가 인접한 경우가 그렇지 않은 경우에 비해 사용자에게 필요한 정보가 상대적으로 더 많이 제공되었다고 할 수 있다.[13] 중의성 고려 그룹과 미고려 그룹으로 구분 분석한 경우에도 각각 21% / 3% / 30%, 53% / 37% / 64% 높은 결과를 보였다.

- 검색 결과 표현 시 질의의 2단어구가 제목에 형태 그대로 존재하는 경우가 제목 내 문장 속에 존재하는 경우보다 더욱 높은 검색 정확도를 보였다. 또한 순위의 상위에서 부여하였던 중요도는 유사한 연결 페이지의 하위 순위로 갈수록 떨어지는 주관적 경향과 함께 더욱 중요한 정보가 하위 순위에 위치할 경우 상위 순위의 중요도를 다시 판단하게 하는 경향이 있다. 따라서 정보 검색 시 상위 순위는 더욱 중요한 결정 요소가 됨을 알 수 있었다.

  사용자의 성향도 검색순위 결정에 상당한 영향을 미치며, 이는 사용자의 선택 수만으로는 결정되기 어려운 확률로서 정보의 필요시기, 사용자의 일정 등 개인 정보와 매우 밀접한 관계가 있음을 알 수 있다.

4.3 인접도 가중치의 영향 평가

전장에서 q_1의 단일 주제 기반 q_2의 선택 확률은 검색 결과 전체 중 q_1 적합 그룹 내 q_2의 출현 빈도와 비례한다고 하였으므로, 가중치의 초기값은 중요도 그룹에서 중요(2)를 받은 순위 페이지 중 A그룹에 해당되는 항목의 점유도를 이용하였다. 산출결과 $Pr(q_2|(T_k|q_1))$의 초기값은 아래 표와 같다.

이 0.75845라는 값은 단어 인접도만을 이용하여 기존 검색 결과에서 사용자가 원하는 정보를 얻어낼 확률이 75.8%에 이른다는 것을 뜻하며, 이는 인접도가 높은 A

표 4 인접도 가중치 초기값의 질의별 산출 결과

구 분	총계	중의성 고려	중의성 미고려
$Pr(q_2\|(T_k\|q_1))$	0.75845	0.61700	0.89989

13) A그룹에 속하면서도 사용자의 의도를 반영하지 못한 결과를 중요도에서 배제(예 : head coach, mass hall, capital seoul)하였기 때문이다.

표 5 검색효율 비교 (중요도합의 평균) - 상위 10순위

구 분	최초	인접도부여			
		α = 1	%	α = 2	%
계	12.56	14.17	13	14.68	16
중의성 고려	9.95	11.19	12	11.62	17
중의성 미고려	15.15	17.16	13	17.74	17

표 6 검색효율 비교 (중요도합의 평균) - 상위 20순위

구 분	최초	인접도부여			
		α = 1	%	α = 2	%
계	23.17	26.01	12	27.68	19
중의성 고려	18.33	20.24	10	21.67	18
중의성 미고려	28.00	31.79	14	33.68	20

그룹이 확률적으로 중요도도 높다는 것을 나타낸다. 이의 반영을 통한 검색 효율 향상을 살펴보기 위하여 최초 검색 순위와 가중치 부여 순위의 중요 페이지 관련 상위 10순위와 상위 20순위를 비교하였는데, 비교 방법은 평가된 중요도 값의 10순위 및 20순위까지의 합산으로 평가하였으며, 가중치 부여 순위는 최초 검색 순위에 인접도 그룹에서 A그룹에 해당되는 항목에 인접도 가중치의 초기값을 부여하여 재산정하였다. 다음 표는 중요도합 평균의 비교이다.

비교 결과 인접도 가중치 부여시 검색 효율이 13~20% 향상되는 것을 알 수 있다. α가 증가하면 중요도가 높은 페이지가 선택적으로 더 많은 가중치를 얻으므로 상위 순위로 접근하는데, 100순위를 중요도 순으로 재배열한 이상적인 순위에 대한 Kendall's τ distance[16] 측정에 대해서는 최초 검색 순위가 인접도 부여 순위보다 이상적인 순위와 유사한 것으로 보아 인접도 가중치를 적용한 순위는 중요도가 특히 높은 페이지를 선택적으로 상위 순위화하는 방법임을 알 수 있다.

5. 결론 및 향후 연구

단어 인접도를 이용하면 검색 결과를 판단할 때 사용자에게 유용한 정보를 75.8%의 확률로 식별할 수 있으며, 인접도 가중치를 이용한 재순위화는 사용자에게 유용한 페이지의 순위를 선택적으로 상위 순위로 보내게 되어 전체적인 검색 효율이 약 13~20% 향상되었다.

이때 계수 α는 인접도 가중치가 검색 결과에 미치는 영향을 나타내는 변수로서 차후 연구에서 2단어로 된 전형적인 검색 방법으로 나타내질 질의를 선정하고 인접도 가중치를 일반화하기 위한 광범위한 분석을 해야 하겠다. 또한 2단어로 제한한 연구를 확장, 문장 등으로 구성된 자연어 질의에 대한 인접도를 산정하고 기존 순위 산정 방식과 조합하여 사용자가 원하는 정보에 더욱

접근할 수 있는 방안을 연구해야 하겠으며, 뉴스, 블로그, 멀티미디어와 같이 일반적인 검색 정보가 아닌 경우의 인접도를 이용한 접근에 대해서도 연구하고자 한다.

참 고 문 헌

[1] F. Qiu and J. Cho, "Automatic Identification of User Interest For Personalized Search," In Proceedings of the 15th international conference on World Wide Web, pp. 727-736, 2006.

[2] B.J. Jansen, A. Spink, and T Saracevic, "Real life, real users, and real needs: A study and analysis of user queries on the Web," Information Processing and Management, 36(2):207-227, 2000.

[3] S. Brin and L page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," In Proceedings of 7th international conference on World Wide Web, pp. 107-117, 1998.

[4] J. Cho and S. Roy, "Impact Of Search Engines On Page Popularity," In Proceedings of the 13th international conference on World Wide Web, pp. 20-29, 2004.

[5] T. H. Haveliwala, "Topic-Sensitive PageRank," In Proceedings of the 11th international conference on World Wide Web, pp. 517-526, 2002.

[6] 박의규, 나동열, 장명길, "문장-질의 유사성을 이용한 웹 정보 검색의 성능 향상", 한국정보과학회 논문지 소프트웨어 및 응용 제32권 제5호, pp. 406-415, 2005.

[7] 김창환, 임지희, 최호섭, 윤화묵, 옥철영, "사용자 어휘 지능망을 이용한 의미적 정보검색", 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호, pp. 157-160, 2006.

[8] 김형일, 김준태, "질의어 의미별 사용자 선호도를 이용한 웹 검색의 성능 향상", 한국정보과학회 논문지 소프트웨어 및 응용 제31권 제8호, pp. 1101-1112, 2004.

[9] 이소영, 조영환, "검색포탈에서 사용자 질의분석을 통한 검색형태 연구", 한국정보과학회지 제22권 제4호, pp. 47-51, 2004.

[10] WordNet. a lexical database for the English language, Princeton Univ, http://wordnet.princeton. edu/

[11] 김형일, 김준태, "워드넷 기반 협동적 평가와 하이퍼링크를 이용한 검색엔진의 성능 향상", 한국정보처리학회 논문지 B 제11-B권 제3호, pp. 369-380, 2004.

[12] 조미영, 김판구, "정보량과 개념적 밀도를 이용한 단어 의미 중의성 해결", 제24회 한국정보처리학회 추계학술발표대회 논문집 제12권 제2호, pp. 445-448, 2005.

[13] The Open Directory Project, http://www.dmoz.org

[14] J. R. Wen, J. Y. Nie and H. J. Zhang, "Clustering user queries of a Search Engine," In Proceedings of the Internation World Wide Web conference, pp. 162-168, 2001.

[15] G-R. Xue, H-J. Zeng, Z. Chen, W-Y. Ma, H-J. Zhang and C-J. Lu, "Implicit Link Analysis for Small Web Search," In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 56-63, 2003.

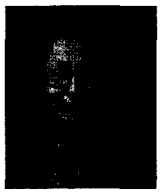
[16] M. Kendall and J. Gibbons, "Rank Correlation Methods," Edward Arnold, London, 1990.

윤 성 웅

1996년 한양대학교 도시공학과(공학사) 2004년 국방대학교 전산정보학과(공학석사). 2004년~현재 국방대학교 전산정보학과 연구원. 관심분야는 정보검색, 시멘틱웹, 자연어 처리

채 진 기

1999년 육군사관학교 전산학과(학사). 2005년~현재 국방대학교 전산정보학과 석사과정. 관심분야는 정보검색, 텍스트마이닝, 데이타마이닝



이 상 훈

1978년 성균관대학교 전자공학과(공학사). 1989년 연세대학교 전자계산학(공학석사). 1997년 일본 교토대학 정보공학(공학박사). 1998년 충남산업대학교 멀티미디어과 교수. 2000년~현재 국방대학교 전산정보학과 교수. 관심분야는 협조작업처리(CSCW), 데이타베이스, 정보 검색, 멀티미디어시스템, 정보 보호