94-a-4

불논리검색, 퍼지검색, 확률검색의 효율 비교연구

A Comparative Study on Effectiveness of Boole logic retrieval, Fuzzy retrieval and Probabilistic retrieval

이 젬마, 사공 철 숙명여자대학교 문헌정보학과

GemMa Lee, Chul SaKong Dept.of Library and Information Science, Sookmyoung Women's University

본 연구에서는 불논리검색의 단점을 보완하기 위한 가장 강력한 검색모형인 퍼지검색과 확률검색의 효율을 불논리검색과 상호비교하였다. 실험데이터로 정보학 분야의 한국어 test collection인 KT Test Set을 이용하였고 색인어와 색인어의 문헌내 출현빈도를 바탕으로 퍼지시소러스를 생성하여 시소러스의 NT, BT로 탐색식을 확장한 다음 각각에 대해 3가지 검색을 행하고 검색효율을 평균재현율과 평균정확률로 측정하였다. 실험결과 검색효율은 재현율에서는 확률검색, 불논리검색, 퍼지검색 순으로, 정확률에서는 퍼지검색, 확률검색, 불논리검색 순으로 나타났다.

1. 서 론

정보검색의 이론적 발전은 불논리검색의 한계에 대해 많은 대안들을 제시했으며 그 중에서도 퍼지집합이론과 확률이론은 가장 강력한 대안으로 연구되어 왔다.

퍼지검색과 확률검색은 각각 독자적인 이론적 체계를 가지고 연구되어 왔으며 검색과정에서 부여되는 가중치에 대한 해석 및 접근방법에서 서로 상이한 특성을 나타내고 있다.

불논리검색을 대체할 수 있는 가장 강력한 검색모형으로서 이 두가지 검색기법에 대한 비교의 문제 또한 오래전 부터 논의되어 왔으나 실제적인 검색실험을 통해 비교·분석된 결과는 제시되지 않았다. 또한 시소러스의 자동생성이나 시소러스에 의한 질문확장 효율에 대해서도 오랫동안 많은 연구가 수행되어 왔으며 계속적인 실험과 연구를 통해 검색효율을 향상시킬 수 있는 방법을 모색하고 있다.

따라서 본 연구에서는 정보과학 분야의 한국어 test collection을 대상으로 퍼지시소러스를 생성하고 시소러스에 의한 질문확장을 통해 전통적인 불논리검색과 퍼지검색, 확률검색의 효율을 비교·분석 해 보고자 한다.

2. 실험방법

이러한 3가지의 검색효율을 비교하기 위하여 1994년 한국통신에서 개발한 KT Test Set을 이용하였다. KT Test Set는 「정보과학회 논문지」,

「한국정보과학회 Proceedings」, 「정보관리학회지」에 수록된 논문의 초록 1,053건으로 구성되어 있으며 이 중 문헌정보학 분야를 제외한 937건의 정보과학 분야에서만 200개의 레코드를 무작위 추출하여 실험데이터로 사용하였다. 색인어는 <keywords>필드에 있는 자연언어 색인어에 수정과 통제를 가하여 사용하였으며, 질문 작성시에는 200개 실험데이터를 대상으로 분류번호 <classification>에 따라 주제분야를 선정하여 주제전문가를 통해 10개의 질문식을 작성하였다. 표제와 초록에서 색인어의 출현빈도를 바탕으로 계층관계의 퍼지시소러스를 구축하였으며 시소러스의 NT(하위어), BT(상위어)로 질문을 확장하여 3가지 검색을 모두 행하였다. 각각의 용어확장에서 3가지 검색결과의 효율은 평균 재현율과 평균 정확률로 측정하였다.

본 실험에서 조설된 총 색인어 수는 1,683개였으며 중복되지 않은 고유한 색인어 수는 1,432개, 문헌 1건당 부여된 평균 색인어수는 8.4개였다. 퍼지시소러스는 Miyamoto의 퍼지시소러스 생성 공식 중 계층관계의 용어를 생성하는 1공식을 이용하여 구축하였다. 문헌에 부여되어 있는 모든 고유한 색인어 w_i와 w_j에 대해 문헌d_k에서 출현빈도를 각각 h_{ik}, h_{jk}라 했을 때 다음과 같은 1공식에 의해 용어간의 관련도가 계산되었다.(Miyamoto, 1990)

$$t(w_i, w_j) = \frac{\sum_k \min(h_{ik}, h_{jk})}{\sum_k (h_{ik})}$$

그러나 퍼지시소러스는 어디까지나 용어의 출현빈도 내지는 동시출현빈도라는 통계적 속성에 의해 작성된 것이므로 시소러스에 의한 질문확장시 전문가가 판단하기에 NT나 BT로 적당하다고 생각되는 용어만을 확장용어로 사용하였다.

확장된 불논리 탐색식은 3가지 검색에 모두 이용되지만 퍼지검색에서는 각 탐색어에 가중치를 따로 입력할 수 있게 하였으며 이러한 가중치를 평균연산기호로 연산한 결과가 문헌 적합도값의 내림차순으로 출력되었다. 확률검색에서는 불논리 연산기호가 모두 무시되며 탐색어와 관련된 통계적 빈도에 의해 문헌의 적합성 확률값이 계산되어 내림차순으로 출력되었다.

3. 검색실험

불논리검색은 완전일치기법으로 AND, OR의 논리연산기호로 결합된 질문의 조건을 모두 만족하는 문헌만이 검색된다.

퍼지검색시에는 확장된 탐색식에서 이용자가 부여한 원래의 탐색어에는 1의 가중치가 부여되며 시소러스에 의해 확장된 용어에는 공식에 의해 계산 된 용어관계값으로 0에서 1사이의 가중치가 부여된다. 각 탐색어의 가중치를 결합하여 최종적으로 검색된 문헌의 적합도 값을 산출해 주는 함수는 다음과 같이 R_1, R_2로 구성된다.(김현희, 배금표, 1993, pp.44-45)

< R_1 >

탐색식을q,탐색어를t_1,t_2라할때,

- $q=t_1$: t_1을 포함하는 모든 문헌 검색
- $q=t_1\ AND\ t_2$: $R_1(q)=R_1(t_1)\cap R_1(t_2)$
- $q=t_1\ OR\ t_2$: $R_1(q)=R_1(t_1)\cup R_1(t_2)$
- $q=t_1\ AND\ NOT(t_2)$: $R_1(q)=R_1(t_1)$

< R_2 >

검색함수 R_2에서는 평균연산기호λ_γ를 사용한다. 또한 매개변수 γ=0.7이 가장 효율적인 것으로 판명되었으므로 본 실험에서도 γ=0.7을 사용한다. 탐색식을 q, 문헌을 d, 탐색어는 t_1, t_2, 탐색어가 문헌에 부여된 색인어로서 갖는 가중치를 $f_d(t_1)$, 확장된 탐색어 가중치를w_1,w_2라 정의할 때[1]

- $q=(t_1, w_1)$

$R_2(q, d)=w_1 \cdot f_d(t_1)$

- $q=(t_1, w_1)\ AND\ (t_2, w_2)\ AND\ldots(t_n, w_n)$

$$R_2(q, d)=\gamma \cdot \min(w_1 \cdot f_d(t_1),\ w_2 \cdot f_d(t_2)\ldots w_n \cdot f_d(t_n)) + \frac{(1-\gamma)\cdot(w_1 \cdot f_d(t_1)+w_2 \cdot f_d(t_2)+\cdots+w_n \cdot f_d(t_n))}{n}$$

- $q=(t_1, w_1)\ OR\ (t_2, w_2)\ OR\ldots(t_n, w_n)$

$$R_2(q, d)=\gamma \cdot \max(w_1 \cdot f_d(t_1),\ w_2 \cdot f_d(t_2)\ldots w_n \cdot f_d(t_n)) + \frac{(1-\gamma)\cdot(w_1 \cdot f_d(t_1)+w_2 \cdot f_d(t_2)+\cdots+w_n \cdot f_d(t_n))}{n}$$

- $q=NOT(t_1, w_1)$

$R_2(q, d)=1-w_1 \cdot f_d(t_1)$

---

1) 여기서 $f_d(t_i)$는 문헌 d에서 색인어 t_i의 중요도를 나타내는 함수로 본 실험에서는 색어 부여시 가중치가 부여되지 않았기 때문에 색인어 집합은 퍼지집합이 아닌 보통집합이 되므로 1의 가중치가 부여된 것과 같다.

본 실험에서는 여러번의 검색을 반복해 본 결과 문헌의 적합도 값이 α=0.5 이상일 때 적합문헌이 가장 많이 검색되는 것으로 나타남으로써 검색 기준치 $\alpha \geq 0.5$인 문헌만이 최종적으로 검색되어 순위가 부여되게 하였으며 이에 대해 검색효율을 계산하였다.

확률검색시 적합성 정보를 알 수 없는 초기 탐색에서는 Croft & Harper의 초기 가중치 공식을 이용해 문헌의 초기 가중치를 구하게 된다. 이 때 공식의 한 항목이라도 0이되는 경우 무한대값이 산출될 가능성을 배제하기 위하여 각 항목에 0.5를 더해 준 Croft & Harper의 초기가중치 공식은 다음과 같다.(Croft & Harper, 1979. p.287)

$$g(x) = \sum x_i \log \frac{N-n_i+0.5}{n_i+0.5}$$

여기서 N은 전체 문헌집합의 수, n_i은 주어진 탐색어가 부여된 문헌의 수이다.

초기 검색결과에 대해 적합문헌을 판정해 주기 위한 기준치로는 α=3.0을 사용하였다. 이는 검색을 여러번 반복해 본 결과 α=3.0 이상일 때 검색효율이 가장 좋았기 때문이다. 여기서 얻어진 적합성 피드백에 의해 다음과 같은 Robertson & Sparck Jones의 공식에 따라 문헌의 최종 적합성 가중치를 계산하게 된다.(Robertson & Sparck Jones, 1976, p.143; van Rijsbergen, 1979, p.119)

$$g(x) = \sum_{i=1}^{n} x_i \log \frac{\frac{r+0.5}{(R-r+0.5)}}{\frac{(n-r+0.5)}{(N-n-R+r+0.5)}}$$

여기서 N은 전체 문헌 수, R은 주어진 탐색식에 적합한 문헌의 수, n은 탐색어가 부여된 문헌의 수, r은 적합문헌 중 탐색어가 부여된 문헌의 수이다.

이렇게 순위화 된 검색결과에 대해 실제로 적합문헌들은 검색 기준치 α=3.0 이상에 분포하는 것으로 나타났다. 따라서 문헌의 적합성 가중치가 3.0이상이 되는 문헌만이 내림차순으로 순위화 되어 최종 검색결과로 제공되며 이에 대해 검색효율이 계산되었다.

4. 검색효율 측정 및 비교분석

이상에서와 같이 3가지 검색기법의 검색효율을 비교 해 보면 재현율은 NT, BT확장시 공통으로 확률검색이 각각 49.83%, 39.37%로 가장 높고, 그 다음이 불논리검색 41.29%, 36.62%, 마지막으로 퍼지검색 32.62%, 35.62%순이었다.

확률검색의 재현율이 가장 높은 이유는 시스템이 문헌의 적합성 가중치를 추정하기 위해서, 주어진 탐색어에 대한 전체 문헌집합내에서 이용가능한 최대한의 정보를 이용하기 때문이라고 판단된다. 또한 van Rijsbergen(1979, pp.35-38)이 말했듯이, 확률검색에서는 계산된 적합성 확률의 순으로 문헌을 순위화함으로써 일정 cut-off에서 기대된 재현율과 정확률을 최대화하는 것이 가능하다. 불논리검색이나 퍼지검색은 불논리 구조를 탈피하지 못한다는 점에서 검색되는 문헌의 수에 한계가 있을 수 밖에 없다. 퍼지검색은 일단 NOT연산기호를 제외하면[2] 불논리를 만족하는 대상문헌집합을 구하고 여기에 탐색어 가중치를 이용하여 문헌의 적합도 값을 구하게 된다. 역시 확률검색과 마찬가지로 검색결과 순위화 된 문헌에 대해 cut-off를 적용함으로써 검색되는 문헌 수 및 검색된 적합문헌의 수를 조정할 수 있다. 퍼지검색이 불논리검색보다 재현율이 낮은 이유는 cut-off를 설정함으로써 상대적으로 검색된 적합문헌의 수가 줄어들게 되어 전체 적합문헌에 대한 재현율이 낮아지는 대신 소수의 검색된 문헌에 대한 적합문헌 수는 많아지므로 정확률이 높아지기 때문이다.

한편 정확률에서는 NT, BT공통으로 퍼지검색이 각각 82.5%, 81.66%로 가장 높았으며, 다음으로 확률검색 77.14%, 77.16%, 불논리검색76.71%, 70.55% 순이었다.

퍼지검색의 정확률이 가장 높은 이유는 불논리에 의해 정확하게 매치된 문헌집합에서 평균

2) 불연산과는 달리 NOT으로 연결된 탐색어도 검색될 수 있다.

연산기호로 계산된 문헌의 적합성 가중치를 일정 수준에서 절단함으로써 검색된 문헌이 줄어드는 대신 그 안에 포함된 적합문헌 수는 늘어나게 되기 때문이다. 퍼지검색 다음으로는 확률검색이 높은 정확률을 얻었다. 이는 일단 탐색어가 하나라도 출현하는 문헌이면 다 검색이 되는 망라성이 높은 결과에서 적합문헌이 고루 분포되게 되는데 여기에 일정수준의 cut-off를 적용하여 정확률 또한 향상시킬 수 있었기 때문이다. 그러나 확률검색에서는 탐색어간의 논리적 관계가 무시되고 단지 탐색어의 통계적 속성에 의해서만 검색이 이루어지므로 이용자가 요구하는 적합한 의미의 문헌만을 검색하는 검색의 완벽을 기하기에는 한계가 있었다고 볼 수 있다. 다음으로 완전매치기법인 불논리검색의 정확률도 높은 편이었으나 퍼지검색이나 확률검색에 비하면 상대적으로 낮은 편이었다.

5. 결론 및 제언

이상에서와 같이 불논리검색, 퍼지검색, 확률검색의 효율을 측정·비교한 결과 NT, BT 질문확장시 공통으로 재현율은 확률검색, 불논리검색, 퍼지검색 순으로, 정확률에서는 퍼지검색, 확률검색, 불논리검색 순으로 나타났다. 이를 불논리검색과 비교해 보았을 때, 확률검색은 보다 향상된 효율을 보여 주었으며 퍼지검색은 정확률에서는 불논리검색 보다 월등히 향상된 효율을, 재현율에서는 다소 낮은 효율을 보여주었으나 그 차이는 매우 미소하였으므로 두가지 검색기법 모두 불논리검색보다 향상된 결과를 보여주었다고 할 수 있다.

퍼지검색은 탐색어간의 의미적 관계 내지는 탐색어 자체의 중요도를 가중치로 표현할 수는 있으나 근본적으로 불대수에 의한 연산이라는 범위를 벗어나지 못함으로써 탐색어가 완전히 일치하는 문헌만이 검색된다는 불논리검색의 한계를 그대로 갖게 된다. 또한 적합성 피드백을 이용함으로써 검색효율을 향상시킬 수 있는 방법이 없다는 점에서 개선의 여지가 있다고 생각된다. 또한 색인어와 탐색어에 부여된 가중치를 수학적으로 연결하여 문헌의 가중치를 계산할 수 있는 검색모형에 대한 통일이 이루어져야 할 것이다.

확률검색은 불논리가 무시되고 용어의 통계적 출현빈도에 의해 문헌이 검색됨으로써 탐색어와 관련된 문헌이 보다 많이 검색될 수는 있지만 용어간의 의미적 관계를 전혀 나타낼 수 없기 때문에 역시 이용자에게 적합한 문헌만을 검색하는데 한계를 갖고 있다. 이는 확률검색이 기본적으로 용어간의 독립성 가정을 전제로 하고 있기 때문에 발생하는 결과라 생각된다. 따라서 용어간의 의존적 관계나 관계의 중요도를 검색모형에 포함시킬 수 있는 방법의 개발이 필요하다 하겠다. 또한 1차 검색시 한번의 적합성 피드백을 거치기 때문에 적합문헌을 판정, 검색의 정확성을 기할 수 있지만 여기서 얻어진 적합성 정보가 과소평가되어 간접적으로만 반영되는 경향이 있으므로 적합성 피드백 정보를 최대한 이용할 수 있는 모형의 개발도 필요하다 하겠다.

참 고 문 헌

김현희, 배금표. 1993.
"퍼지정보검색시스템의 검색효율에 관한 연구." 정보관리학회지 7(1) : 79-95.

Bookstein, N.J. & Croft, W.B. 1987.
"Probability and Fuzzy-Set Applications to Information Retrieval." *ARIST* 20 : 117-149.

Croft, W.B. & Harper, D.J. 1979.
"Using Probabilistic Models of Document Retrieval Without Relevance Information." *JD* 35(4) : 285-295.

Miyamoto, Saddaki, 1990.
Fuzzy Sets in Information Retrieval and Cluster Analysis. Dordrecht : Kluwer Academic Publishers.

Robertson, S.E. & Sparck Jones, K. 1979.
"Experiments in Relevance Weighting of Search Terms." *IPM* 15 : 133-144.

van Rijsbergen, C.J. 1979.
Information Retrieval. London : Butterworths.