유전자 알고리즘 기반 유사도 변환을 이용한 순위 재조정 검색 모델

Re-Ranking Retrieval Model Using Similarity Transformation Based on Gene Algorithm

이재훈*, 이성주**
*조선대학교 일반대학원 전자계산학과
**조선대학교 컴퓨터공학부
e-mail : nuridepo@chosun.ac.kr

요 약

정보·통신과학의 발달로 다양한 영역에서 수많은 정보들이 발생하고 있다. 그 결과 사용자의 요구에 무분별한 응답을 제시하는 검색 모델도 발생하였다. 본 논문은 정보들 사이의 유사도를 변환하고 순위를 재조정하여 더욱 적합한 정보를 상위 순위에 제시함으로써 사용자 요구에 더욱 적합한 정보를 획득할 수 있는 모델에 대해 연구하였다.

1. 서론

최근 정보·통신과학의 발달은 다양한 영역에서 수많은 정보들을 발생시키고 있으며, 이는 현재 '정보홍수'라고 명명할 만큼 방대한 정보를 축적되어 왔다. 이러한 대량의 정보들 중에서 사용자가 원하는 정보를 찾아내기란 쉽지 않은 일이다.

또한 방대한 정보들 중에서 원하는 정보를 찾기 위해서 정보 검색 시스템을 사용하지만 정보 검색 시스템이 제시하근 검색 결과는 사용자가 하나씩 읽어보면서 확인하기에는 너무 많은 양이다. 이러한 정보 과적재(information overload) 문제는 정보 검색 시스템에서 해결해야 할 과제로 남아 있다.

이에 따라, 사용자가 요구하는 정보를 더욱 신속하고 적합하게 획득할 수 있는 모델에 대한 연구가 끊이지 않게 되었다. 본 문은 정보들 사이의 유사도를 변환하고 순위를 재조정하여 더욱 적합한 정보를 상위 순위에 제시함으로써 사용자 요구에 더욱 적합한 정보를 획득할 수 있는 모델에 대해 연구하였다.

본 논문은 다음과 같이 구성되었다. 2장에서는 관련 검색 모델과 유전자 알고리즘의 정보검색 적용의 동향을 고찰하고, 3장에서는 정보들 사이의 유사도를 변환하고 순위를 재조정하는 유전자 알고리즘을 제시하며, 4장에서는 제시한 알고리즘을 이용하여 시스템을 구현하였고 5장에서는 실험하고 평가하였고 끝으로 6장에서 결론에 대해 언급하였다.

2. 관련연구

2.1 정보 검색 모델

불리언 모델 : 불리언 모델은 집합 이론(Set Theory)과 부울 대수(Boolean Algebra)를 기반으로 한 모델로서 AND, OR, NOT 등의 연산을 제공한다. 간단한 공식을 제공하며 사용자가 원하는 의미를 정확하게 표현할 수 있다는 장점이 있는 반면, 부분 정합이 불가능하므로 지나치게 많은 문서가 검색되거나 극소수의 문서만이 검색될 수 있으며, 질의와의 유사도가 무조건 0 아니면 1로 정해지므로 랭킹을 할 수 없다는 단점이 있다.

벡터 모델 : 불리언 모델에서는 질의나 문서의 키워드에 모두 이진가중만을 할당하는데 이로 인해 그 성능에 한계가 있는 반면, 벡터공간 모델에서는 질의나 문서의 키워드에 이진이 아닌 적절한 가중치를 할당할 수 있다. 불리언 모델에서 할 수 없는 부분정합이 가능하다. 따라서 질의와 문서의 유사한 정도에 따라 랭킹을 매길 수 있다는 장점을 갖는다. 이 모델에서는 키워드들이 서로 모두 독립이라는 가정을 한다. 그러나 실제로는 단어들이 서로 독립이 아닌 어떤 관계를 갖게 되므로 가정이 옳다고는 할 수 없다. 이러한 가정을 완화시킨 것이 일반 벡터 공간 모델이다.

확률 모델 : 벡터 공간 모델의 경우 '모든 단

어들이 서로 독립이다'라는 가정으로 인해 단어들 사이의 관계를 표현할 수 없다는 단점을 갖는 것 이외에도, 질의와 문서간의 유사도 등 상당수의 파라미터가 임의로 정해질 수 있는, 즉 모델 자체에서 해결할 수 없다는 문제점이 있다. 반면, 확률 모델에서는 단어 사이의 관계, 의존, 질의어의 가중치나 질의와 문서 사이의 유사도 등이 모두 모델 안에서 정의된다. 확률 모델에서는 우선 문서집합 전체가 크게 두 부분, 즉 사용자의 질의에 상대적인 부분과 그렇지 않은 비상대적인 부분으로 나뉘어져 있다고 가정한다. 따라서 질의에 대한 가장 이상적인 결과 문서 집합은 질의어에 대한 결과로 나온 문서가 상대적인 문서일 확률을 극대화 시켜줄 수 있는 그러한 문서집합이 된다고 할 수 있다.

2.2 정보검색에서 유전자알고리즘의 연구 동향

스키마 이론의 발전은 Holland에 의해 연구되었고, GA에 대해서도 이론적으로 연구 되었다.

Gordon는 보다 좋은 문서를 기술(descriptions)하기 위해 적응점 유전자 알고리즘에 대하여 연구를 하였다. 각 문서에는 N개의 개체가 할당되고 각 개체는 인텍싱 단어의 집합으로 구성된다. GA는 확률적인 모델에 의한 것 보다 좋은 문서 개체를 만들어 낸다. 재생산는 co-relevant 문서 방식으로 생산된 것 보다 20 세대에서 39.74%가 향상 되었고, 40 세대에서는 56.61%가 향상되었다.

Yang and Korfhage는 GA에서 문서 단어 인데싱를 하는데 가중치를 변경하여 질의 최적화하는 방법을 제안하였다. 선택은 확률적인 연산 방법을 사용하였고, 교배는 이중 교배점을 사용하는 blind 교배를 사용하였다. 돌연변이는 질의 개체군을 재생산하는 전통적인 돌연변이 방법을 사용하였다. 실험에서는 6 세대 이후의 적합한 문서 질의에 집중되어 있다.

Chen and Kim는 GANNET(하이브리드 유전자 and 뉴럴 네트워크)를 제안했다. 이 시스템은 GA를 이용하여 사용자가 선택한 문서를 키워드로 최적화하는 형태이다.

Kraft et alsms는 가중치 불리언 질의 정형화를 향상시키기 위해서 GA 프로그래밍을 제안하였다. 문서는 인덱스 단어의 벡터의 관점이다. 가중치 블리언 질의는 Koza의 유전자 모델에서 염색체로 표현되었다. GA의 목표는 단어의 재현율과 정확율 검색 성능을 향상시키기 위해 질의를 수정하는 것이다.

3. 유전자 알고리즘 기반 유사도 변환을 이용한 문서 순위 재조정 알고리즘

본 논문에서는 축소용어 집합을 기반으로 작성된 시소러스를 통한 각 문서의 검색상태 값(P_r)과 원시 문서베이스를 기반으로 작성된 각 문서의 검색 상태 값(P)은 1차 질의확장과 2차 내용 질의에 대한 검색 상태 값을 의미한다. 따라서 1단계 질의 확장으로 재현율을 유지하고, 정확률을 높이기 위한 2단계 유사관계 행렬 기반의 클러스터 검색을 수행하였다. 본 논문에서는 검색 순위를 재조정(Re-ranking)을 통해 적합한 문서가 상위에 검색될 수 있도록 검색 상태 값을 재조정한다.

$$Si_{combined} = \alpha P_r + \beta P = \{RSV(d_1), RSV(d_2), ..., RSV(d_n)\} \quad \cdots(식1)$$

P_r : 축소용어 행렬 기반 퍼지검색(문서상태 값)
P : 유사관계 행렬 기반 클러스터 검색
α, β : 1로 설정
$Si_{combined}$: 문서 상태 간의 순위 재조정 결과

(식1)에서 도메인의 특성에 따라 α, β을 적용하여 재조정한다.

최종적인 문서 상태 값은
$P_r = \{d_1/0.71, d_2/0.36, d_3/0.63, d_4/0.39, d_5/0.39\}$이고,
$P = \{d_1/0.84, d_2/0.40, d_3/0.73, d_4/0.24, d_5/0.32\}$,
여기서 α와 β 값을 1:1로 설정하면
$Si_{combined} = \{d_1/0.78, d_2/0.38, d_3/0.68, d_4/0.32, d_5/0.36\}$
이다.
따라서 검색상태 값이 0.5 이상을 선택하면
$Si_{0.5} = \{d_1/0.78, d_3/0.68\}$이다.

위와 같이 각 단계별로 계산하는 문서 순위 재조정 알고리즘은 다음 [알고리즘1]과 같다.

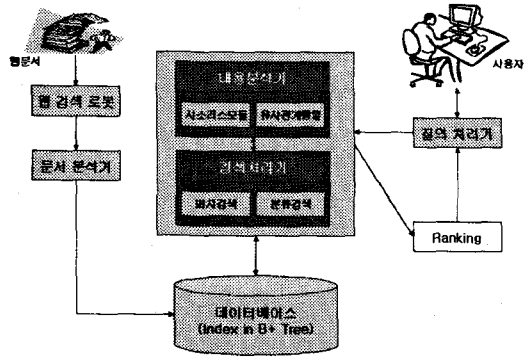
[알고리즘1] 문서 순위 재조정 알고리즘

입력 : 각 문서
출력 : 문서 순위가 재조정 된 문서 집합

```
function Re_ranking-document(){
  get document_set() ; // 수집된 문서 입력
  occurrence-frequency();
  // 형태소 분석가를 이용한 색인어의 위치 정보와
  // 빈도수 계산
  source_base();
  // 문서분석기(원시 문서베이스 생성)
  reduction_term_set(); // 축소용어 행렬 생성
  thesaurus_creation(); // 시소러스 생성
  similarity_relation_matrix(); //유사관계 행렬 생성
  get_use._query();
  query_expansion(); //시소러스기반 질의어 확장
  fuzzy_search(); //문서탐색(퍼지 검색)
  cluster_search();
  //유사관계 행렬기반 질의어 확장기 및
  // 문서 클러스터
  similarity_combine_rerank();
  //유사도 결합을 통한 순위 재조정
  put-document_set();
}
```

4. 시스템 구현

본 논문에서 제안한 검색 모델의 시스템은 [그림1]과 같이 입력된 문서들을 대상으로 문서를 분석하는 부분, 퍼지 함수와 관계성을 이용하여 문서내용을 분석하는 부분 그리고 확장된 질의를 이용하여 문서를 검색하고 유사도 변환을 통한 문서 순위를 재조정하는 부분으로 구성된다.



[그림1] 검색 시스템 구조

문서 분석기는 내용 분석 및 질의 처리 부분과는 별도로 정보검색 시스템의 중요한 부분이 문서에 대한 정보를 추출하여 색인을 생성하고 관리하는 부분이다. 따라서 저장구조는 문서에 대한 정보를 추출하여 색인을 생성하고 저장하는 부분이고 역화일을 만들기 위해 인덱스 파일은 B+ tree 파일과 포스팅 파일로 구성된다.

문서 검색 시스템에서 문서는 문서 구조를 고려한 색인어의 집합에 의해 표현되며, 색인어 연관관계 정도에 의해 문서의 내용을 구조화하거나 문서집단으로 대표되는 특정 주제 영역의 지식구조를 파악할 수 있다. 따라서 본 논문에서는 원시 문서베이스를 기반으로 키워드간의상호 의존관계를 통해 문서 내용을 분석하기 위한 유사관계행렬을 구성한다. 또한 높은 시간 복잡도를 해소하고 도메인 지식을 표현하는 축소행렬을 정의하고 이를 기반으로 색인어들의 의미관계를 정의하기 위한관점에서 시소러스를 생성하여 도메인의 영역 지식을 분석한다.

시스템의 문서검색 및 순위 재조정기는 위의 문서의 내용 분석과정에서 계산된 유사관계 행렬과 시소러스를 통해 질의를 확장하여 문서를 검색 하는 과정이다. 검색 분석 과정은 퍼지검색, 분류검색 및 재 순위화 모듈로 구성된다.

5. 실험 및 평가

실험에 사용된 문서의 집합은 컴퓨터 과학 분야의 10개 주제의 논문들로 인터넷을 이용하여 국회도서관에서 검색 가능한 석·박사 학위 논문과 한국정보과학회와 한국정보처리학회의 논문지에 실린 논문을 분야별로 약 50편씩 498편에 해당하는 문서집합이다.

정보검색 분야에서 가장 널리 사용되는 성능 평가의 척도에는 정확률과 재현율이 있다. 본 논문에서는 적합성 정도에 따라 검색결과로 얻은 문서들의 순위를 결정하는 것이 목적이므로 정확률에 기반을 둔 적합률을 제안하여 평가에 사용하며 일반적인 정확률 공식은 (식2)와 같고 이를 변형한 적합률은 (식3)과 같다.

$$\text{정확률} = \frac{\text{검색된 적합 문헌 수}}{\text{검색된 문헌 총수}} \quad \cdots(\text{식2})$$

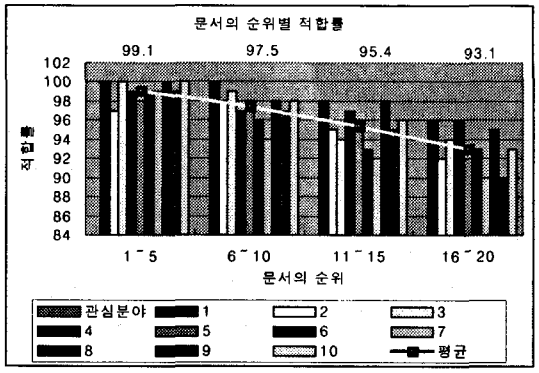
$$\text{적합률} = \frac{\sum_{i=1}^{n} R_{score}}{\sum_{i=1}^{n} R_{\max}} \quad \cdots(\text{식2})$$

유사도에 따른 문서순위결정은 유사한 관심분야라도 문서순위결정이 다르게 이루어져야 하므로, 세부 전공이 다른 사용자들을 대상으로 순위별 적합률을 측정하였다. 의사문서를 작성하고 이를 이용하여 문서의 순위를 결정하였다. <표1>의 값은 세부분야별 각 5명의 사용자들이 평가한 평균 적합률이다.

평가에 적합률을 사용하였는데, 이 실험의 경우에 n은 순위별 제시되는 문서의 개수이다. 예를 들어 문서의 순위 1~5에서 n은 5이고, 순위 11~20에서의 n은 10이다. [그림2]는 <표1>을 그래프로 나타낸 것이다.

<표1> 문서의 순위별 평균 적합률

순위 / 관심분야	1~5	6~10	11~15	16~20
1	100	100	98	96
2	97	98	95	92
3	100	99	94	94
4	99	97	97	96
5	98	98	96	92
6	99	96	93	93
7	99	94	92	90
8	100	98	98	95
9	99	97	95	90
10	100	98	96	93
평균	99.10	97.50	95.40	93.10



[그림2] 문서순위결정 기법의 성능

실험 결과, 1~5순위의 결과에 대해서는 99.1% 이상의 적합률을, 20 순위 내에서는 95% 이상의 적합률을 보임으로써 본 논문에서 제안한 문서순위결정 기법이 사용자의 요구를 만족시키기에 충분히 우수함을 알 수 있었다.

6. 결론 및 향후 연구

여러 전문 분야에서 산출하는 지식 정보의 양이 기하급수적으로 증가하고 정보의 가치 또한 중요시되는 정보화 사회에서 적합한 정보를 손쉽고 빠르게 얻는 것은 현대인의 당연한 욕구가 되었다. 이러한 요구에 따라 수많은 지능형 정보검색 시스템들이 등장하여 여러 분야에서 활발히 이용되고 있다.

그러나 검색결과의 양이 방대하여 최적의 결과를 찾는데 소요되는 시간과 노력이 증대되고 있다. 따라서 사용자의 관심분야와 선호하는 것을 고려하여 최적의 정보만을 제공해주는 시스템이 필요하게 되었다. 이러한 요구에 따라 사용자의 질의어와 검색된 문서들이 얼마나 유사한가에 따라 문서의 순위를 결정하는 문서순위결정 기법과 사용자의 기호를 저장해 놓고 이를 참조하여 필요 없는 정보를 여과시켜주는 정보 필터링 기법이 연구되고 있다.

일반적인 문서순위결정 기법은 사용자가 입력한 질의로 질의 벡터를 구성하고 빈도수와 같은 문서 자체의 정보를 이용하여 문서 벡터를 만든 후, 코사인 유사 공식으로 질의 벡터와 문서 벡터를 비교하여 문서의 순위를 결정한다. 본 논문에서는 문서들 간의 유사도를 사용자 선호에 맞게 변환하여 문서순위결정 기법에서는 문서를 문서들의 유사도를 사용자의 선호에 맞게 변수를 조정함으로써 문서순위를 재조정하였고 이는 방대한 양의 정보들에서 사용자가 원하는 문서를 상위 순위에 제시함으로써 정보검색 시간을 더욱 단축시킬 수 있음을 보였다.

향후 유사도 변환을 유전자 알고리즘을 적용하여 군집화된 정보들로 적재하고 기계학습을 통해 규칙을 생성하고 또한 생성된 군집에 해당하는 문서군에 대해 유사도를 2차 변환하면 사용자 질의에 더욱 정확한 정보를 획득할 수 있을 것이라 사료된다.

참고문헌

[1] 은희주, 김용성, "유사관계행렬을 기반으로 한 순위 재조정 검색 모델", 정보과학회논문지, 제31권 제11호, 2004.11

[2] 유영준, "문헌정보학에서 지식 구조에 관한 연구", 연세대학교 대학원 박사학위논문, 2003.8

[3] Bracha Shapira, Uri Hanani, Adiraveh, Peretzshoval, "Information Filtering: A New Two-Phase Model Using Stereotypic User Profiling," Journal of Intelligent Information System, Vol. 8, 1997.

[4] Maria R. Lee, "Context-Dependent Information Filtering," Proceedings of ISDL '97, November Tsokuba, Ibaraki, Japan, pp.19-21, 1997.

[5] Scott Deerwester, Susan T.Dumais, George W.Furnas, Thomas K. Landauer, Richard Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society for Information Science, Vol. 41, No. 6, pp. 391-407, 1990.