개인화된 정보 필터링 에이전트를 위한 유전 알고리즘

손윤희, 박상호
안동대학교 컴퓨터공학과

Genetic algorithm for personalized imformation filtering agent

Yun-Hee Son, Sang-Ho Park
Dept of Computer Engineering Andong University
E-mail : chowon2@orgio.net, sahngpark@empal.com

요 약

유전 알고리즘을 이용한 정보 필터링 에이전트는 기존의 검색엔진에서 찾고자 하는 문서에 대해 검색된 문서의 유사도가 낮은 문제점을 해결한다. 본 논문에서는 HTML 태그의 중요도 가중치와 HTML 태그 안의 위치에 대한 가중치를 유전 알고리즘을 이용하여 학습한다. 여기서 학습된 가중치가 높은 태그와 태그 안의 위치 그리고 출현하는 빈도수에 대한 중요도 가중치를 다시 유전 알고리즘을 이용하여 학습하고 여기서 학습된 가중치로 검색된 문서를 필터링 하여 정보검색 성능을 향상시킬 수 있는 방법을 제안한다. 이 때 태그의 중요도 가중치 값을 학습하는 방법으로 하나의 태그를 유전자로 매핑하고 일련의 태그 집합을 염색체로 표현한 유전 알고리즘을 이용한다. 태그 안의 위치에 대한 중요도 가중치 값도 같은 방법을 이용한다. 여기서 나온 태그와 위치 그리고 빈도 수에 대한 중요도 가중치 값을 다시 유전자 알고리즘 이용하여 계산한다. 이 값으로 검색된 문서를 필터링 하여 기존의 정보검색보다 검색자가 원하는 검색문서에 상당한 정확율을 제공하는 방법을 제안한다.

1. 서론

인터넷의 급속한 보급으로 인해 웹 상에서 정보의 제공자와 사용자가 급증하고 있다. 웹은 편리한 사용자 인터페이스와 멀티미디어 환경을 제공함으로써 점차 다양하고 방대한 양의 정보와 사용자 중심의 서비스를 제공하고 있다. 이러한 방대한 웹 문서에서 사용자가 원하는 정보를 정확하게 찾는다는 것은 쉬운 일이 아니다. 여기에 대한 해결책으로 정보 검색엔진에 대한 연구와 정보 검색 필터링, 비교 검색 등 다양한 방향으로 접근이 시도되고 있지만 사용자의 요구와 100% 일치하는 검색기는 현재 존재하지 않는다.

이러한 사용자의 검색요구에 근접하기 위해 본 논문에서는 유전자 알고리즘을 이용하여 정보 검색에서 보다 더 효과적인 정보검색 필터링 방법을 소개한다. 기존의 전형적인 정보검색 방법에서는 단지 문서상에 나타난 단어들만을 기준으로 단서를 삼아 검색을 수행하였다. 그러나 본 논문에서는 전형적인 정보 검색에 덧붙여 HTML문서의 구조를 정보검색에 이용하고 이를 통해 검색 성능을 향상시킬 수 있는 방법을 제안한다. 이 방법에서 한 태그는 한 개의 유전자로 매핑 되며, 일련의 태그 집합은 염색체로 표현된다. 다음 유전 알고리즘을 이용하여 태그의 중요도에 대한 가중치를 학습하며, 이를 이용한 문서 검색에서 상위에 위치한 검색 결과 문서에 대해 검색 성능이 향상되는 방법을 제안한다.

마찬가지 방법으로 태그 안에서의 위치에 대한 가중치를 학습하며, 최종적으로 문서에서의 빈도 수와 태

그와 태그 안에서 존재하는 위치의 중요도에 대한 가중치를 학습하여 그 결과를 문서 검색 결과에서 정보 필터링하여 검색 효과를 향상시키는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 웹 문서에 대한 정보 검색 및 검색 모델, 그리고 정보필터링에 대해서 알아본다. 3장에서는 가중치 학습을 위한 유전 알고리즘을 소개한다. 4장에서는 실험과 결과에 대하여 소개하고 5장에서는 결론 및 향후 연구 방향을 밝힌다.

2. 관련 연구

전 세계적으로 수많은 정보들이 인터넷 상에 놓여져 있다. 그러나 사용자가 원하는 정보가 어느 위치에 있는지 쉽게 알지 못하게 때문에 웹 상에서 많은 검색 도구들과 함께 다양한 검색 방법이 등장하고 있다. 대부분의 인터넷 검색은 대개 큰 주제에 대해 간단한 질의를 통해 검색하는 방법을 이용한다. 그런데 이 경우 각각의 질의에 대해 1000개를 클릭해서 겨우 10개를 만족시키는 정도이다. 많은 개체에서 개체집단을 유지하면서 그들 모두가 동시에 최적값을 찾는 방법으로 유전자 알고리즘이 탁월하다. 본 논문은 정보 검색의 정확성에 대한 문제점을 해결하기 위해서 유전자 알고리즘을 이용하여 질의에 대해 검색 결과 문서를 거르는 필터링 기법을 이용한다.

2.1 웹 문서에 대한 정보 검색

웹 문서를 대상으로 한 정보 검색의 연구는 내부구조와 외부구조로 나눌 수 있다.

먼저 웹 문서 외부 구조 즉 하이퍼링크에 대한 연구로 Spertus는 어떤 페이지 a가 링크를 통해 b와 페이지 c를 가리키고 있다면 페이지 b와 c는 서로 내용이 관련되어 있을 가능성이 있다는 것을 관찰했다. 한편 웹 문서를 분류하기 위해 링크와 링크의 순서를 이용하는 방법이 연구되었다.[1]. 하이퍼링크 정보를 이용한 대표적인 웹 검색 사이트의 예는 구글이다.[2]. 구글은 각 웹 페이지에 대한 랭킹의 품질을 페이지랭크라고 정의하고, 이를 계산하기 위해 링크 구조를 활용한다.

웹 문서의 내부구조는 몇 개의 검색 사이트를 통해 제한된 범위에서 활용되어 왔다. 라이코스[3](Lycos)는 웹 페이지를 타이틀, 헤더 및 바디 언더라인 부분으로 구별해 어느 위치에 질의어가 발생하는 지에 따라 질의와 문서의 연관 점수를 조정하는 것으로 알려져 있으며, 알타비스타(AltaVista)와 야후 같은 검색 사이트는 웹 페이지의 타이틀 부분에 출현하는 질의에 대해 더 높은 점수를 준다. 그러나 HTML 태그를 기준으로 한 내부구조가 검색 성능에 주는 영향에 대한 평가나 검색 시스템에 어떻게 적용되는 지에 대한 구체적인 방법은 제시되어 있지 않다. 한편 태그의 중요도를 단어의 빈도수에 적용하는 방법이 Cutler에 의해 연구되었다.[4]

검색 시스템에 있어서 그 성능에 영향을 주는 요소는 웹 문서의 구조 외에도 여러 사항들을 포함한다. 또한 이 요소들은 상호 독립적으로 존재할 수도 있고 연관되어 있을 수도 있다. 따라서 각 파라미터의 최적의 값을 찾는 것은 정보 검색에 있어서의 중요 문제라고 할 수 있다. 검색 시스템 LASER는 이와 같은 파라미터들을 제공할 수 있도록 구현되었다.[5]. 파라미터들은 구 값에 따라 검색 엔진이 HTML 필드 안에 있는 단어들에 영향을 받는 정도나 하이퍼링크들의 연관성에 영향을 받는 정도, 또는 부분 단어 매칭이나 질의-팀 인접성에 어느 정도의 영향을 받을 것인지를 결정짓는다. 이 연구에서 그들은 앞의 파라미터를 가지고 있는 검색 함수를 최적화하기 위해 모의 담금질(simulated annealing)기법을 적용하였다.

2.2 검색 모델

불린 검색 모델에서는 각 문서는 색인어의 집합으로 표현되고 질의어는 불린 수식, 즉 불린 연산자로 연결된 색인어로 구성된다. 시스템은 질의에 해당하는 불린 연산식을 만족시키는 문서들을 검색하지만 사용자 질의에 대한 문서의 유사도는 검색하지 않는다. 따라서 문서 정렬 기능은 없다. 이런 단점을 보완하기 위해 여러 확장 모델이 제안되었다.[6][7][8]

벡터 모델에서 모든 질의와 문서는 벡터로 표현되며

질의나 문서의 키워드에 적절한 가중치를 할당하여 질의와 문서의 유사도가 높은 문서들을 검색한다.
색인어 가중치는 여러 방법으로 계산될 수 있는데, Salton과 McGill[9]이 다양한 용어 가중치 기법에 대해 조사했다. 최적의 방법으로 가장 적당한 가중치를 준다면 문서의 유사도는 상당히 높아질 것이다.

확률 모델은 특정 질문과 문서와의 관련 확률을 산출하여 확률이 큰 문서를 검색한다. 단, 기본적인 가정이 적합한 문서인지 부적합한 문서인지는 검색 이전에 결정되어 있다는 것이다. 따라서, 적합 문서들이 알려져 있는 상태에서는 유용하겠지만 사용자가 처음으로 질의를 입력하는 최초 검색에서는 적합 문서들이 알려져 있지 않으므로 추정 값을 적용하여 검색하여야 한다.[10].

2.3 정보필터링

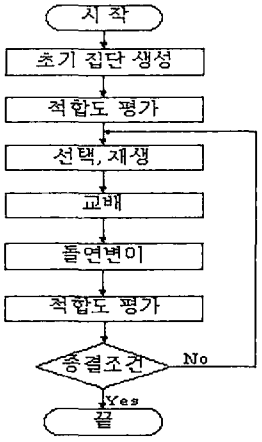
정보 필터링은 주식 정보나 신문기사 정보 전송같이 질의가 상대적으로 정적이고 새로운 문헌이 시스템에 들어왔다 분배되는 동작 모드이다. 특히 정보 필터링은 매일 발생하는 수천 개의 신문 기사 중에서 관심있는 기사를 선택하거나 재판 판결문을 얻는데 유용하다. 정보 필터링 작업에서는 사용자 기호에 근거한 프로파일 작성이 중요하다. 가장 단순한 방법은 사용자에게 필요한 키워드 집합을 쓰게 하는 것이다. 그러나 사용자가 전체 문헌에 익숙하지 않으므로 사용자에게 필요한 키워드를 직접 쓰게 하는 것은 현실적이지 못하다. 좀더 정교한 방법으로 사용자의 기호 정보를 수집하여 프로파일을 동적으로 구축하는 방법이다.
본 논문에서는 무수히 많은 개체에서 세대를 거듭하면서 최적의 개체를 찾아가는 유전자 알고리즘을 이용하여 정보필터링 하는 방법을 제안한다. 정보 필터링 함에 있어서 문제점은 처음에 사용자 프로파일에 대한 정보가 전혀 없다는 점이다. 무수히 많은 개체집단에서 아무런 정보 없이 최적의 값을 찾는 유전자 알고리즘은 이 문제를 해결하기에 적당한 방법이다.

3. 가중치 학습을 위한 유전자 알고리즘

유전자 알고리즘은 유전적 계승과 적자생존이라는 개념을 모델링 한 확률적 최적화(optimization)탐색방법이다.

3.1 유전 알고리즘

개체집단에서 개체(population)들은 스트링 또는 염색체라고 한다.



<그림1. 기본적인 단순 유전 알고리즘의 흐름도>

각 개체는 문제에 대한 해가 될 수 있는 가능성을 나타내며 개체 집단 위에서의 진행과정은 해가 될 가능성이 있는 것들에 대한 탐색에 해당한다. 유전자 알고리즘은 탐색과 해의 가능 영역들을 균형 있게 이용하는 일반성 있는 부류의 탐색방법이다. 기존의 다른 탐색 방법들은 탐색 공간에서 최적값을 찾기도 전에 지역극소(local minimum)에 빠질 위험이 있지만 유전자 알고리즘은 해가 될 가능성이 있는 개체집단을 유지하면서 그들 모두가 동시에 최적값을 찾아나가기 때문에 지역 극소에 빠질 위험을 어느 정도 해결할 수 있다는 점이 중요한 특징이다. 유전자 알고리즘은 순환 동안에 해가 될 가능성이 있는 개체집단을 유지한다. 각 해는 평가되어 적합도(fitness)의 척도를 제공한다. 이로부터 더 적합한 개체들을 선택함으로써 새로운 개체집단이 구성된다. 이 새로운 개체 집단 중 일부 개체들은 교배와 돌연변이에 의해 새로운 해를 구성하고 이 해들의 적합도를 평가함으로써 새로운 해를 얻게 된다.

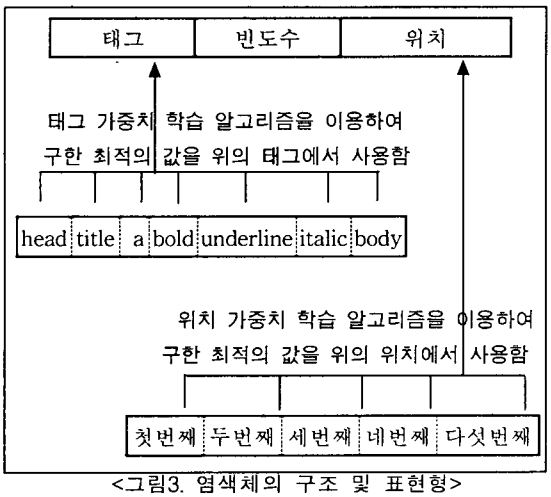
3.2 가중치 학습을 위한 유전 알고리즘

웹 문서에서 태그 중요도를 계산하고 그 가중치를 계산하는 결정적인 방법은 아직 없다. 본 논문에서는 <가중치 학습을 위한 유전자 알고리즘>을 이용하여 주어진 문제를 해결하는 방법을 제안한다.

```
모집단의 초기화;
for I = 1 to MaxGen
  적합도 함수에 의해 모든 염색체 평가
  for I = 1 to N
     Select chromosome for next poppulation;
     Crossover ;
     Mutation;
     적합도 함수 의해 모든 염색체 평가
  end for
  N개의 염색체를 자식염색체로 교체
end for
```

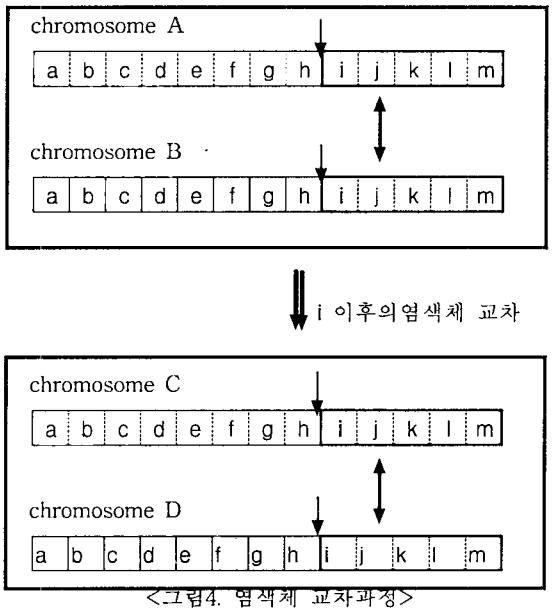
<그림2. 가중치 학습 위한 유전자 알고리즘 >

제안된 알고리즘에서는 하나의 태그는 한 유전자에 대칭 되며 염색체는 유전자의 집합 즉, HTML 태그들로 구성된다. 초기 해 집단은 임의로 구성되며, 적합성 함수는 태그의 가중치를 이용한 검색 결과의 성능을 측정하는 함수이다.[11] 다음은 염색체의 구조를 표현한 것이다.

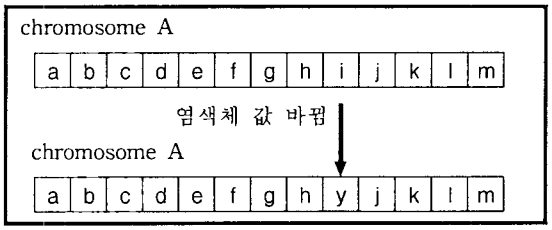


<그림3. 염색체의 구조 및 표현형>

선택된 부모 염색체는 선택과, 교차, 돌연변이에 의해서 자식 염색체를 생산한다. 선택은 적합도 함수에 의해 계산된 값이 높은 것으로 선택한다. 교차는 해당위치의 값 이후의 유전자를 서로 교환하고, 돌연변이는 해 집단 중 임의의 해와 염색체의 위치의 값을 바꿈으로 돌연변이를 생성한다. 다음은 염색체의 교차과정과 선택과정이다.



<그림4. 염색체 교차과정>



<그림5. 염색체 돌연변이과정>

4. 실험 및 결과

실험은 문서 집합에 대해 20회 반복 시행하였으며 태그의 중요도에 따른 최적의 가중치는 20세대 이후의 모든 해집합을 대상으로 가장 적합도가 높은 염색체를 학습된 최적 가중치 후보로 간주하였다. 다음은 실험을 위해서 설정된 파라미터이다.

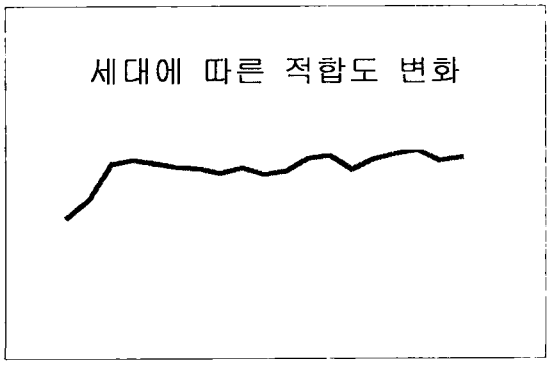
<실험에 사용된 파라미터>

- Total Population size = 50
- Chromosome length = 13
- Maximum # of generations = 20
- Crossover probability = 0.6
- Mutation probability = 0.0333

태그 가중치를 적용한 검색 방법은 다음과 같다.

1. 학습을 위한 질의 데이터에 대해 제안된 학습 알고리즘을 이용하여 태그 가중치를 얻는다.
2. 일정 횟수만큼 1의 과정을 반복한 후 얻어진 태그 가중치 중에서 가장 높은 값을 가지는 염색체, 태그 가중치 집합을 최적의 태그 가중치 값으로 간주하여 최종 태그 값으로 한다.
3. 일정 횟수만큼 1의 과정을 반복한 후 얻어진 태그 안의 위치 가중치 중에서 가장 높은 값을 가지는 염색체, 태그 가중치 집합을 최적의 태그 가중치 값으로 간주하여 최종 태그 안의 위치 값으로 결정한다.
4. 학습된 가중치가 높은 태그와 태그 안의 위치 그리고 출현하는 빈도수에 대한 가중치를 2,3과 같은 방법으로 결정한다.
5. 질의에 대해 위의 가중치를 이용하지 않고 일반 검색을 시행, 200개의 문서를 검색한다.
6. 검색된 문서에 대해 선택된 가중치 및 가중치 값을 이용하여 유사도 값을 조정한다.
7. 변경된 유사도 값을 기준으로 검색 문서를 재정렬한다.

제안된 학습 알고리즘에 의한 학습결과는 다음과 같다.



<그림6. 실험 결과>

세대가 진행됨에 따라 해집합 전체의 평균 적합도 값이 변화하는 과정을 그래프로 나타낸 것이다. 4세대까지는 급격히 증가하다가 그 이후로는 약간의 증가가 있다. 세대가 어느 정도 진행되면은 해집단의 다수가 우수한 염색체에 가까워지므로 적합도 함수의 값이 완만하게 증가한다.

5. 결론 및 고찰

제안된 검색 방법은 실험을 통해 태그의 중요도 가중치를 학습하고 학습된 태그의 가중치 정보를 다시 유전인자로 한 것에 학습된 태그 안의 위치에 대한 가중치를 다시 유전인자로 한 태그 속의 위치 가중치 정보와 빈도수를 유전자로 하여 최종적으로 계산된 가중치 정보를 정보 검색에 사용하였으며 그 변화를 비교하였다. 확인 결과 가중치를 이용한 정보검색에서 검색 결과 향상에 대한 가능이 보여진다. 현재 일반 검색기와 비교했을 때 검색 결과에 상당한 차이가 나는 것은 아니지만 향후 정보 검색의 효율성을 고려한 문서와 구조화가 잘 되어 있는 웹 문서에서는 검색결과에서 상당한 결과가 기대된다. 또한 특정 사용자에 대한 개인화에 유전적 알고리즘을 적용하여 수정된 최적의 사용자 프로파일의 값으로 정보 검색에 성능을 향상시킬 수 있으며 이에 대한 연구가 요구된다.

[참고문헌]

[1] Chakrabarti s., Dom, B., Gibson, D., Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A., Expcrimcnts in Topic Distillation, ACM-SIGIR '98 Post-Conference Workshop on Hypertext Information Retrieval for the Web, 1998.

[2] Brin, S. and Page, L., The Anatomy o a Large-scale Hypertextual Web Search Engine, The Seventh International world Wide Web Conference(WWW7), pp.107-117, 1998

[3] Mauldin, M. L., Lycos: Design Choices in an Internet Serch Service, IEEE Expert, 12(1), pp. 8-11, 1997.

[4]G.Salton and M.J. McGill. Introduction to Medern Information Retriewal. McGraw Hill Book Go., New York, 1983.

[5] Boyan, J., Freitag, D., and Joachims, T., A Machime Learning Arcitecture for Optimizing Web Search Engines, Proceedings of the AAAI Workshop on Inernet-Based Information Systems, pp. 1-8, 1996.

[6] Salton , G., Fox, E. A., and Wu, H., Extended Boolean Information Retrieval, Communications of the ACM, Vol. 26, No. 11, pp. 1022 1036, 1983

[7] Turtle, H. and Croft, W. B., Evaluation of an Inferencee Nework-based Retrieval Medel, ACM Transactions on Information Systems, Vol. 9, No. 3, pp. 187-222, 1991.

[8] Callan J. P., Croft, W. B. and Harding, S. M., The INQUERY Retrieval System, Proceedings of the Third International Conference on Database and Expert Systems Applications, Springer, pp. 78-83, 1992.

[9] Salton, G., Wong, A and Yang, C. S., A Vector Space Model for Automatic Indexing, Communications of the ACM 18, pp. 613-620, 1975.

[10] Croft, W. B. and Harper, D. J., Using Probabilistic Models of Document Retrieval without Relevance Information, journal of Documentation, 35(4), pp. 285-295, 1979.

[11] Baeza-Yates and Ribeiro-Neto, Modern Information Retrieval(korean), pp. 86-91, 2001.