진화연산을 이용한 웹 문서의 특성 학습

Learning Web-Document Characteristics Using Evolutionary Computation

김 선, 장병탁
서울대학교 컴퓨터공학부
Sun Kim°, Byoung-Tak Zhang
Artificial Intelligence Lab (SCAI)
School of Computer Science and Engineering, Seoul National University
{skim, btzhang}@scai.snu.ac.kr

요 약

대용량의 문서를 대상으로 한 정보 검색은 인터넷과 WWW이 대중화되면서 웹 문서로 확장되었다. 기존의 문서는 주로 텍스트만으로 구성되는 데 반해 웹 문서는 HTML을 기반으로 문서가 작성된다. HTML은 문서의 형태를 이루게 하는 여러 종류의 태그들로 구성되어 있고 문서 작성자는 이를 이용, 자기 의도를 홈페이지에 반영한다. 따라서 태그 정보의 학습은 검색 효율을 향상시키는 데 도움을 줄 수 있다. 본 논문에서는 이러한 HTML의 태그 특성을 이용해 검색 효율을 향상하는 방법을 제시한다. 제시된 방법은 진화 알고리즘을 사용하여 질의와 검색결과를 담고 있는 데이터를 학습한다. 학습을 통해 얻어지는 결과는 각 태그에 대한 가중치 정보들이며, 이는 검색엔진의 문서 가중치 정보로 사용된다. TREC 데이터를 사용하여 실험하였으며 태그 정보를 이용함에 따른 검색 성능 변화를 비교 분석하였다.

1. 개 요

인터넷은 탄생된 해인 1969년 이래로 성장을 거듭해 왔다. 특히 HTML 문서를 기반으로 한 WWW(World Wide Web)이 등장하면서 그 속도는 기하급수적으로 증가해 인터넷은 웹 페이지만 3억 개가 넘는 규모로 발전하게 되었다 [1]. 이런 상황에서 사용자가 원하는 정보를 찾아 주는 웹 정보 검색이 등장하게 되었다. 그러나 비교적 규모가 작았던 인터넷 초기의 웹 검색 환경에 비해 그 크기가 커진 지금에 와서 검색 효율을 높이는 것은 그리 쉬운 문제가 아니다.

기존의 정보 검색은 대량의 문서를 보유하고 있는 도서관, 회사 등의 환경에서 이루어져 왔다. 이 때 문서는 주로 텍스트만으로 이루어져 있으며 임의의 문서를 대상으로 한 검색의 단서로 주로 텍스트만이 고려되었다. 반면 인터넷 정보 검색의 특징은 작성되는 문서의 형태가 다르다는 점에서 찾을 수 있다. 거의 모든 웹 문서는 HTML(HyperText Markup Language)을 바탕으로 작성된다. HTML은 웹 문서를 만드는 기본 틀을 제공하며 이 틀은 문서상의 태그 정보로 표시되어 웹 브라우저를 통해 번역된다. 문서 작성자는 이 점을 이용, 자신이 작성하는 웹 페이지에 자신의 의도를 담는다. 이

는 헤드문자, 볼드체, 이탤릭체 등으로 화면 또는 출력 문서상에 보여지게 된다. 따라서 태그를 통해 일반 텍스트와 구별되어 있는 부분은 문서의 특징을 담고 있다고 볼 수 있으며, 이에 대한 학습은 검색 성능을 향상시키는데 도움이 될 수 있다고 추측 할 수 있다.

본 논문에서는 이러한 HTML의 특성을 이용해 검색 결과를 향상시키는 방법을 제시한다. 사용된 검색엔진은 정보검색 컨퍼런스인 TREC(Text REtrieval Conference)을 위해 구축된 SCAIR(SCAI information retrieval engine)를 기반으로 하였다[2]. 그리고 태그의 특성을 학습하는 과정에 진화연산을 사용하였다.

2. 검색 시스템

검색엔진인 SCAIR는 벡터공간모델을 기반으로 한다[3]. 한 문서는 단어들의 집합으로 볼 수 있으며 이 때 한 단어를 텀(term)이라고 하면 각 문서는 텀들의 리스트 또는 텀 벡터로써 간주될 수 있다. 그리고 전체 문서의 집합은 텀과 문서의 행렬이 된다. 질의도 역시 단어로 구성된 문장이므로 문서와 마찬가지로 텀들의 리스트가 된다.

문서를 구성하고 있는 텀들은 $tf \cdot idf$ 방식을 사용해 인덱스 된다[4].

$$w_{ik} = tf_{ik} \cdot \log\left(\frac{N}{df_k}\right) \quad (1)$$

w_{ik} : 문서 i에서 k번째 텀의 가중치
tf_{ik} : 문서 i에서 k번째 텀의 빈도수
N : 전체 문서의 수
df_k : k번째 텀을 포함한 문서의 수

질의의 각 텀에 대해서는 식 (1)의 idf (tf를 제외한 부분)에 의해서 가중치가 부여된다.

문서와 질의의 유사도를 결정하기 위해 변경된 코사인 측정 방법을 사용하며 이는 식 (2)에 나타나 있다.

$$sim(d_i, q_j) = \frac{\sum_{k=1}^{n} \alpha_m \cdot w_{ik} \cdot p_{jk}}{\sqrt{\sum_{k=1}^{n} (\alpha_m \cdot w_{ik})^2 \cdot \sum_{k=1}^{n} p_{jk}^2}} \quad (2)$$

w_{ik} : i번째 문서의 텀 가중치
p_{jk} : j번째 질의의 텀 가중치
α_m : 상수값($m \in$ M, M : the set of tag)

질의와 문서에 대한 유사도를 계산한 후에는 이를 정렬해 리스트의 형태로 출력한다.

한편 검색대상인 문서는 HTML로 작성된 웹 문서이므로 HTML 태그를 처리하는 과정이 필요하다. 따라서 각 문서에 속해 있는 태그 정보를 별도로 저장하는 과정과 태그의 중요도에 따른 가중치를 적용하는 과정이 추가된다. 한 문서를 구성하는 텀 중에 특정 태그에 속해 있는 텀들은 인덱스 과정 중 별도로 표시를 한다. 그리고 태그에 대한 가중치는 유사도를 결정하는 데 적용되어 상수 α의 형태로 식 (2)에 적용된다.

3. 진화연산을 이용한 문서 특성 학습

이미 기술한 것처럼 웹 문서의 특징은 화면 표시 및 하이퍼링크(hyperlink)등을 위한 태그를 담고 있다는 것이다. 태그의 가중치를 학습하기 위한 기법에는 여러 가지 방법이 있을 수 있으나 여기에서는 진화연산을 이용해 학습을 하여 검색에 적용하였다.

각 염색체(chromosome)는 태그들로 구성되며 실수 값을 가진 태그의 가중치 열로 표현된다. 초기 해 집단은 임의로 구성된 염색체들로 구성되어 평가된다.

다음 세대로 진화를 하기 위한 부모 염색체의 선택은 다음과 같이 이루어진다. 부모 염색체는 해 집단 중 품질이 좋은 상위 반절에 속하는 개체 중에서 임의로 선택되며 품질은 적합도 함수(fitness function)에 의해 결정된다.

적합도 함수는 태그 가중치에 대한 검색 결과의 성능을 측정하는 함수이다. 성능을 측정하는 방법으로는 TREC의 검색 결과 측정 방법인 평균 정확율(average precision) 값을 사용하였다[5]. 정확율은 검색 결과 문서 중 관련된 문서(relevant document)의 수의 비율을 말한다.

선택된 부모 염색체는 교차(crossover)를 통해 자식 염색체를 생산한다. 교차는 염색체의 각 위치에 대해 두 부모 염색체의 평균을 내어 자식 해의 해당 위치로 값을 배정하는 산술적 교차(arithmetical crossover)로 이루어진다[6]. 교차를 통해 생성된 자식 해들은 해 집단 중 적합도가 낮은 반절의 염색체와 대체된다. 변이(mutation) 연산은 해 집단 중 임의의 해와 염색체의 임의의 위치의 값을 바꿈으로 이루어지며 돌연변이를 생성시킨다.

```
initialize chromosomes
for g = 1 to g_max
    evaluate all chromosomes by fitness function
    for i = 1 to M
        choose two chromosomes p1, p2
        offspring_i = crossover(p1,p2)
        offspring_i = mutation(offspring_i)
    end for
    replace M chromosomes by offsprings
end for
return optimal chromosome
```

<그림 1> 진화연산을 이용한 학습 알고리즘

해 집단에 대한 진화 알고리즘은 <그림 1>에 나타나 있다.

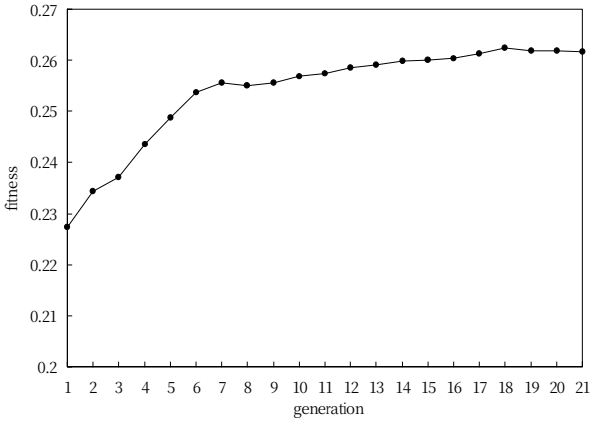
4. 실험 및 결과

실험을 위해 사용한 데이타는 TREC의 웹 트랙에서 사용하는 질의와 문서집합이다[7]. TREC에서 질의는 토픽(topic)이라고 하며 한 토픽은 제목(title), 서술(description), 상세 설명(narrative)로 구성된다. 서술은 제목에 대해 간단한 기술을 한 부분이며, 상세 설명은 서술보다는 더욱 자세하게 찾고자 하는 정보를 기술한 부분이다. 실험을 위해서 사용된 것은 토픽 중 제목과 서술 부분이다.

실험은 토픽 중 401번에서 420번 토픽을 이용해 이루어 졌다. 학습을 위해서 401번에서 410번까지의 토픽을 사용했으며, 학습된 태그 가중치에 대한 검색을 위해서는 411번에서 420번까지의 토픽을 사용했다.

검색에 따른 결과는 문서의 관련도에 따라 한 토픽당 1000개의 문서를 출력하며 평균 정확율에 의해 성능이 측정된다.

가중치를 학습하기 위한 태그 정보는 HTML의 태그 중 유의할 가능성이 있다고 판단된 다섯 개의 태그(<TITLE>, <Hx>, <B>, <I>, <A>)를 사용했다. 태그는 각각 제목(Title), 헤더문자(Header), 볼드체(Bold), 이탤릭체(Italic), 링크지시어(Anchor)를 뜻한다. 문서의 작성자는 웹 페이지를 작성할 때 강조하고 싶은 부분 또는 페이지의 특성에 맞는 부분에 대해 일반 문자와는 다른 형태로 표시하고자 하는 경향이 있다. 따라서 제목, 헤더문자, 볼드체, 이탤릭체를 주요한 태그 정보로 보았다. 그리고 다른



<그림 2> 세대에 따른 적합도 변화

문서와의 링크를 위해 사용되는 링크지시어를 태그 정보로 추가했다. 이는 문서와 문서를 링크할 때 관련이 있는 문서와 연결을 시킨다고 보았기 때문이다.

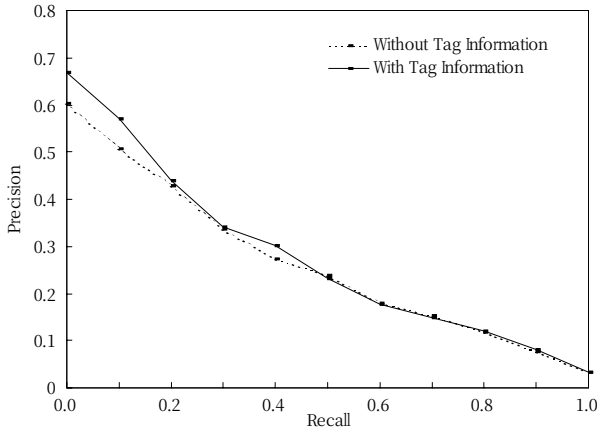
다음 실험 결과는 10회에 걸쳐 얻은 데이터를 분석한 것이다. <그림 2>는 진화연산에 의한 학습 결과이다. 세대가 진행됨에 따라 해집합 전체의 평균 적합도 값이 변화하는 것을 그래프로 나타낸 것이다. 세대가 증가함에 따라 8세대까지는 적합도가 급격히 증가하다가 그 이후에는 완만하게 증가함을 보이고 있다. 이는 사용한 진화연산의 산술교차가 원인이다. 산술교차는 적합도가 높은 두 염색체의 평균값으로 자식 해를 생산한다. 또한 교체되는 염색체 수는 한 세대의 해 집단의 2분의 1이다. 따라서 세대 초기의 해 집단은 한 세대의 변화에도 적합도가 우수한 해 집단의 염색체에 급격하게 수렴하는 경향을 보인다. 반면 세대가 어느 정도 진행한 후에는 해 집단의 다수가 이미 기존의 우수한 염색체에 가깝게 수렴을 했으므로 적합도 값의 증가가 둔화되는 것이다.

진화연산을 통해 학습한 태그의 가중치 정보는 검색을 위해 적용되었다. 20세대 이후의 해 집단 중 가장 높은 적합도를 가지는 염색체를 태그의 적합한 가중치로 간주했다.

<표 1>은 관련문서 검색 수를 비교한 것으로 태그 가중치의 적용에 따른 변화가 거의 없음을 보이고 있다.

	일반 검색 결과	태그 적용 검색 결과
검색된 관련문서 수	262	261±5

<표 1> 관련 문서 검색 수 비교



<그림 3> Interpolated Recall-Precision Averages

<그림 3>은 학습된 태그의 가중치를 적용했을 때와 그렇지 않았을 때를 평균 정확율-재현율(Average Precision-Recall)로 비교한 그림이다. 재현율(Recall)이란 전체 관련 문서 중 검색 결과에 포함된 문서의 수를 나타내는 비율이다. 재현율이 0.5이하일 때는 태그 정보를 적용한 경우의 정확율이 더 높다가 그 이후에는 차이가 없음을 보인다. 비슷한 관련 문서 검색 수에 대해 재현율이 작을 때 정확율이 더 높다는 것은 관련된 문서가 결과 값의 상위에 존재한다는 것을 의미한다. <표 2>는 재현율에 대한 평균 정확율의 변화를 비교한 것이다. 재현율이 작을 때 태그를 적용한 검색 결과가 더 좋은 정확율을 보이고 있다.

한편 일부 검색 결과는 평균 정확율이 일반 검색 결과에 비해 떨어지는 결과를 보였다. 이는 학습이 지역해로 수렴했거나 일부 문서에만 잘 맞도록 학습된(overfitting) 결과이다.

재현율	일반 검색 결과	태그 적용 검색 결과
0.0	0.6042	0.6551±0.0385
0.1	0.5073	0.5555±0.0276
0.2	0.4286	0.4426±0.0220
0.3	0.3377	0.3293±0.0129
0.4	0.2731	0.2871±0.0155

<표 1> 재현율에 따른 평균 정확율의 변화 비교

5. 결 론

본 논문에서는 웹 문서 검색 결과를 향상시키기 위한 방법으로 유전자 알고리즘을 이용해 HTML 태그의 가중치를 학습하는 방법을 제시하였다. 학습된 태그의 가중치 정보를 검색에 사용해 그 변화를 비교하였다. 태그 가중치를 적용한 경우, 가중치를 적용하지 않았을 때에 비해 그 성능이 향상됨을 보였다.

유의할 점은 검색된 관련 문서 수만을 고려할 때 태그 가중치에 따른 변화는 거의 없다는 것이다. 그러나 검색된 문서에 대한 관련 문서의 분포는 태그 정보를 적용한 경우, 그렇지 않았을 때에 비해 관련 문서들이 결과의 상위에 더 많이 나타나 있다. 이는 태그 정보의 학습이 검색된 문서 중 관련 문서 수를 높이는 것보다는 관련 문서 수를 상위에 위치시키는 데 도움을 준다는 것을 뜻한다. 실제 사용자가 검색 엔진을 사용할 때의 체감 검색 효율은 관련된 문서가 얼마나 먼저 보여지느냐에 달려 있다는 것을 볼 때 태그 정보의 활용은 중요하다고 보여진다.

이 실험은 토픽 20개를 대상으로 이루어졌다. 따라서 보다 많은 토픽을 적용했을 경우에 어떤 변화를 보이는 지에 대한 분석이 추가로 필요하다. 또한 학습이 지역해로 수렴하거나 학습 데이터에 대한 가중치로만 진화하는 것에 대한 문제 해결이 필요하다.

감사의 글

본 연구는 정보통신부 대학기초 연구(과제번호 98-199)에 의해 일부 지원되었음.

참고 문헌

[1] Lawrence, S. and Giles, C. L., *Searching the World Wide Web*, Science Vol. 280 pp. 98-100, 1998

[2] Shin, D. H. and Zhang, B. T., *A Two-Stage Retrieval Model for the TREC-7 Ad Hoc Task*, The Seventh Text Retrieval Conference(TREC-7), 1998

[3] Salton, G., Wong, A., Yang, C. S., *A vector Space Model for Automatic Indexing*, Communications of the ACM 18, 1975.

[4] Salton, G., *Automatic Text Processing*, Addison-Wesley. pp.279-281, 1989

[5] Voorhees, E. M. and Harman, D., *Overview of the Eighth Text Retrieval Conference*, TREC-8, 1999

[6] Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolutionary Programs*, Springer. pp.104-105, 1992

[7] ACSys, *TREC Web Tracks homepage*, http://pastime.anu.edu.au/TAR/WT/