공학석사학위논문

유전 알고리즘을 이용한 웹 문서 검색

Web-Document Retrieval Using Genetic Algorithms

2001년 2월

서울대학교 대학원
컴퓨터공학과
김 선

유전 알고리즘을 이용한 웹 문서 검색

Web-Document Retrieval Using Genetic Algorithms

지도 교수 장 병 탁

이 논문을 공학석사 학위논문으로 제출함

2000년 10월

서울대학교 대학원

컴퓨터공학과

김 선

김선의 공학석사 학위논문을 인준함

2000년 12월

위원장 印

부위원장 印

위 원 印

초 록

본 논문은 웹 문서를 대상으로 한 정보 검색에서 문서를 구조화하는 데 사용하는 태그를 이용, 그 중요도를 학습하고 이를 웹 문서 검색에 이용해 검색 성능을 향상시키는 방법을 제시한다.

기존의 문서는 주로 텍스트만으로 구성되는 데 반해 웹 문서는 일반적으로 HTML을 기반으로 문서가 구조화된다는 특징을 가지고 있다. HTML은 문서의 형태를 이루게 하는 여러 종류의 태그들로 구성되어 있으며 문서 작성자는 이를 이용, 자신의 의도를 홈페이지에 반영한다. 본 논문에서는 HTML의 태그 정보를 이용하여 검색 성능을 향상하는 방법을 제시한다. 제시된 방법은 유전 알고리즘을 사용, 태그의 중요도를 학습한다. 한 태그는 한 개의 유전자로 매핑 되며, 일련의 태그 집합은 염색체로 표현된다. 학습을 통해 얻어지는 결과는 각 태그에 대한 가중치 정보들이며, 이는 검색엔진에서 문서 가중치 정보로 사용된다.

제안된 방법은 한 개의 인공 데이터 집합과 정보 검색 평가대회인 TREC에 사용된 데이터 집합에 적용되었다. 실험을 통하여 제안된 알고리즘은 태그의 중요도에 따라 가중치를 학습하며, 이를 이용한 문서 검색에서 상위에 위치한 결과 문서에 대해 검색 성능이 좋아짐을 보여 주었다.

주요어: 정보 검색, 웹 문서, 태그 가중치, 유전 알고리즘

제목 차례

1. 서론

1.1 연구 배경

인터넷은 탄생된 해인 1969년 이래로 성장을 거듭해 왔다. 특히 HTML(HyperText Markup Language) 문서를 기반으로 한 WWW(World Wide Web)이 등장하면서 그 속도는 기하급수적으로 증가해 인터넷은 웹 페이지만 3억 개가 넘는 규모로 발전하게 되었다[Lawrence and Giles, 1998]. 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 사용자가 원하는 문서를 검색하는 일은 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 웹 문서는 텍스트를 위주로 하나, 그 형태를 유지하기 위한 구조가 있다는 특징이 있다. 일반적으로 대부분의 웹 문서는 HTML을 바탕으로 작성된다. HTML은 웹 문서를 만드는 기본 틀을 제공하며 이 틀은 문서상의 태그 정보로 표시되어 웹 브라우저를 통해 번역된다. HTML은 다음과 같이 일반 텍스트 문서가 가지고 있지 않은 두 종류의 구조로 구성되어 있다[Boyan et al., 1996].

- 첫째는 내부구조이며, 이는 HTML 태그에 의해 표시된 텍스트 조각으로 구성된다. HTML은 문서의 텍스트에 대해 일정한 역할을 부여한다. 어떤

것은 볼드체나 이탤릭체 등과 같은 포맷과 관련되어 있으며, 헤드라인이나 앵커(anchor), 다른 문서와의 연결을 위한 하이퍼링크 부분에 해당하는 텍스트 조각과 같이 보다 풍부한 의미를 내포하게 하는 역할도 있다.

· 둘째는 외부구조이다. HTML 페이지는 무수히 많은 다른 페이지들과 잠재적으로 연결되어 있다. 이것은 하이퍼링크를 통해 연결되어 있으며, 자신이 링크하고 있는 다른 페이지와 자기 페이지를 링크하고 있는 다른 페이지 등으로 서로 얽혀 있다.

문서 작성자는 이러한 구조를 이용, 작성하는 웹 페이지에 자신의 의도를 담는다. 따라서 태그를 통해 일반 텍스트와 구별되어 있는 부분은 그 웹 문서의 특징을 담고 있다고 볼 수 있으며, 이를 이용하는 것이 문서 검색 결과를 향상시키는데 도움을 줄 수 있다고 생각할 수 있다.

전형적인 정보검색 방법에서는 단지 문서상에 나타난 단어들만을 기준으로 단서를 삼아 검색을 수행하였다. 그러나 본 논문에서는 전형적인 정보검색에 덧붙여 위에서 언급한 HTML 문서의 내부구조를 정보 검색에 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 학습한다. 제안된 알고리즘은 한 개의 인공 데이터 집합과 정보 검색 평가대회인 TREC(Text REtrieval Conference)에 사용된 데이터 집합에 적용되었다. 실험을 통하여 제안된 알고리즘은 태그의 중요도에 따라 가중치를 학습하며, 이를 이용한 문서 검색에서 상위에 위치한 결과 문서에 대해 검색 성능이 좋아짐을 확인할 수 있었다.

1.2 논문의 구성

본 논문의 구성은 다음과 같다. 2장에서는 웹 문서 검색에 관련된 연구와 정보 검색에 진화 연산을 이용한 연구에 대해 설명한다. 3장에서는 웹 문서 검색을 위해 사용한 검색 시스템에 대해 기술하며 4장에서는 태그의 중요도에 따른 가중치 학습을 위해 제안된 학습 알고리즘을 기술한다. 5장에서는 한 개의 인공 데이터 집합과 TREC 데이터 집합에 대한 실험 결과를 기술한다. 마지막으로 6장에서는 연구 내용을 요약하고 앞으로의 연구 과제와 함께 결론을 맺는다.

2. 관련 연구

제 1장에서 상기한 바와 같이 웹 문서의 태그 정보를 이용한 검색방법에 있어서의 특징은 문서 구조를 형성하는 태그를 이용한다는 점과 유전 알고리즘을 정보 검색에 이용한다는 점으로 구분될 수 있다. 특히 본 논문에서는 이 특징과 관련해 웹 문서를 대상으로 한 정보 검색과 진화 연산을 이용한 정보 검색, 두 가지로 나누어 관련 연구를 설명하고자 한다. 웹 문서에 대한 정보 검색은 웹 문서가 가지고 있는 내부구조 또는 외부구조를 검색에 이용한 연구에 대해 설명하며, 진화 연산을 이용한 정보 검색에서는 기계 학습 기법의 하나인 진화 연산이 정보 검색에 어떻게 적용되었는지를 설명한다.

2.1 웹 문서에 대한 정보 검색

웹 문서를 대상으로 한 정보검색의 연구는 주로 문서의 내부구조가 아닌 외부구조, 즉 하이퍼링크에 집중되어 있다[Bharat and Henzinger, 1998][Picard, 1998][Weiss et al., 1996]. 이 연구들은 공통적으로 다음 내용을 가정하고 있다.

- 만일 어떤 페이지 *a*로부터 페이지 *b*로 연결된 링크가 있다면, 페이지 *a*의 작성자는 페이지 *b*를 추천하고 있는 것이고 다수의 경우는 관련된 페이지들이 링크된다.

Spertus는 공동 인용이 두 페이지가 서로 연관되어 있음을 나타낼 수 있다는 것을 관찰했다[Spertus, 1997]. 다시 말해 어떤 페이지 *a*가 링크를 통해 페이지 *b*와 페이지 *c*를 가리키고 있다면 페이지 *b*와 *c*는 서로 내용이 관련되어

있을 가능성이 있다는 것이다. 한편 웹 문서를 분류하기 위해 링크와 링크의 순서를 이용하는 방법이 연구되었다[Chakrabarti et al., 1997]. 이 연구에서 그들은 문서 내에서 한 링크에 가까이 위치한 링크의 경우, 같은 주제를 가진 페이지를 가리키고 있을 가능성이 높다는 것을 보였다. 하이퍼링크 정보를 이용한 대표적인 웹 검색 사이트의 예는 구글(Google)이다[Brin and Page, 1998]. 구글은 각 웹 페이지에 대한 랭킹의 품질을 페이지랭크(PageRank)라고 정의하고, 이를 계산하기 위해 링크 구조를 활용한다. 만일 한 페이지가 있고 그 문서를 가리키는 다른 많은 문서들이 존재한다면 그 페이지는 높은 페이지랭크 값을 가질 수 있다. 또한, 높은 페이지랭크 값을 가진 문서를 링크하고 있는 어떤 문서가 있다면 역시 그 문서도 높은 페이지랭크 값을 가지게 된다.

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[Boyan et al., 1996]. 파라미터들은 그 값에 따라 검색 엔진이 HTML 필드 안에 있는 단어들에 영향을 받는 정도나 하이퍼링크들의 연관성에 영향을 받는 정도, 또는 부분 단어 매칭이나 질의-텀 인접성에 어느 정도의 영향을 받을 것인지를 결정짓는다. 이 연구에서 그들은 앞의 파라미터를 가지고 있는 검색 함수를 최적화화기 위해 모의 담금질(simulated annealing) 기법을 적용하였다.

웹 문서 내부구조는 이미 몇 개의 검색 사이트를 통해 극히 제한된 범위에서 활용되어 왔다. 알타비스타(AltaVista)와 야후(Yahoo)같은 검색사이트는 웹 페이지의 타이틀 부분에 출현하는 질의에 대해 더 높은 점수를 주는 것으로

알려져 있으며, 라이코스(Lycos)[Mauldin, 1997]는 웹 페이지를 타이틀, 헤더 및 바디 부분으로 구별해 어느 위치에 질의어가 발생하는 지에 따라 질의와 문서의 연관 점수를 조정한다. 그러나 HTML 태그를 기준으로 한 내부구조가 검색 성능에 주는 영향에 대한 평가나 검색 시스템에 어떻게 적용되는지에 대한 구체적인 방법은 제시되어 있지 않다. 한편 태그의 중요도를 단어의 빈도수에 적용하는 방법이 Cutler에 의해 연구되었다[Cutler et al., 1999].

2.2 진화연산을 이용한 정보 검색

진화 연산은 결정적인 문서 검색 방법이 아직 존재하지 않는 상황에서 문서 색인, 질의 수정, 특징 선택 등 정보 검색의 각 분야 연구에 적용되어 왔다.

문서를 색인하는 과정에서 문서 기술(description)에 해당하는 키워드 집합은 여러 후보들이 존재할 수 있다. 이 후보들 중에서 적합한 키워드 집합을 선택하는 과정에 진화연산이 사용되기도 한다[Gordon, 1988]. 이때 한 개의 키워드는 유전자로 표현되며, 키워드 리스트는 한 개체로 표현되어 진다. 그리고 초기 사용자에 의해서 관련성이 판단된 문서 집합이 초기 해집합(population)을 구성하게 된다. 초기 해집합은 적합도 측정 함수에 기반해 세대를 거쳐가면서 진화하며 점차적으로 최적의 해집합으로 수렴하게 된다.

진화 연산을 이용해 질의를 자동으로 수정해 나가는 적응적 학습 기법 또한 소개되어 있다[Yang et al., 1993][Yang and Honavar, 1998]. Yang의 연구에서는 진화 연산의 연산자 및 병렬 검색 능력을 이용해 사용자 질의를 수정하는 방법을 제시하고 있으며, 또한 문서 분류에 있어서 특징 선택(feature selection) 문제에 진화 연산을 적용하는 기법을 개발하였다. 문서 분류 문제에 있어서 분류자(classifier)의 성능 및 비용은 분류자를 만들기 위해 사용되는 특징들의 선택에 영향을 받는다. 그리고 이 선택의 과정에서 최적의 특징 집합을 찾는데 진화 연산이 사용되었다.

3. 검색 시스템

정보 검색에서 있어서 주로 사용되는 검색 모델로는 불린(boolean) 모델, 벡터공간(vector space) 모델, 확률(probabilistic) 모델 등이 있다. 불린 모델은 AND, OR 등의 불린 연산자를 기반으로 하며, 벡터공간 모델은 문서를 벡터로 보고 사용자의 질의와 문서 사이의 관련성을 벡터 공간상에서의 거리에 의해 측정한다. 또한 확률 모델은 문서의 단어들에 대한 통계를 바탕으로 질의와 문서의 연관성을 확률로써 계산한다. 본 논문의 실험에 사용한 검색 엔진은 기본적으로 벡터공간 모델을 기반으로 하고 있으며, 기타 모델을 지원하도록 설계되었다. 이번 장에서는 앞에서 간략히 언급한 주요 정보 검색 모델에 대해 설명하고, 제안된 검색 방법에 사용한 엔진인 SCAIR를 기술한다.

3.1 검색 모델

3.1.1 불린 모델

불린 검색 모델은 시스템으로 구현하기 쉽고 질의를 처리하는 시간이 매우 빠르기 때문에 상용 검색 시스템에서 널리 사용되어 왔다. 불린 검색 모델에서 각 문서는 색인어의 집합으로 표현되고 질의어는 불린 수식으로 표현된다. 불린 수식은 불린 연산자 AND, OR, NOT으로 연결된 색인어들로 구성되며 시스템은 질의에 해당하는 불린 연산식을 만족시키는 문서들을 검색한다.

그러나 불린 검색 모델은 사용자의 질의에 대한 문서의 유사도(similarity)를 계산하지 않기 때문에 문서 정렬 기능을 제공하지 못한다. 따라서 이를 보강한

확장 불린 모델이 등장하였다. 불린 검색 시스템에 문서의 유사도에 따른 정렬 기능을 제공하기 위하여 MIN과 MAX와 같은 퍼지 집합 연산자를 사용하는 퍼지 집합 모델이 개발되었다. 그러나 MIN과 MAX 연산자가 질의와 문서의 유사도 계산에서 사용자의 생각과는 일치하지 않는 결과를 생성하는 문제가 발생되었고, 이를 극복하기 위하여 여러 확장 모델들이 제안되었다[Salton et al., 1983][Turtle and Croft, 1991][Callan et al., 1992]. 이 모델들은 AND와 OR에 대하여 서로 다른 연산자들을 사용하지만 확장된 불린 모델을 기반으로 하여 설명될 수 있다.

확장 불린 검색 방법의 하나인 P-norm 모델은 역시 불린 모델의 단점을 개선하기 위해 제안된 것으로, 색인어에 대한 가중치뿐만 아니라 질의어에 대한 가중치도 갖는다. P-norm 모델에서는 용어 A_1, A_2, …, A_n에 대한 색인 가중치 d_{A1}, d_{A2}, …, d_{An}을 갖고 있는 문서 D를 n차원 공간에 있는 (d_{A1}, d_{A2}, …, d_{An}) 좌표에 있는 것처럼 처리한다. d_{A1} OR d_{A2} OR … OR d_{An} 형태의 OR 질의가 있다고 가정하면 이 질의에 대해 모든 n좌표가 0인 문서는 검색에서 제외된다. d_{A1} AND d_{A2} AND … AND d_{An} 형태의 AND 질의에 대해서는 모든 n좌표가 1인 문서, 즉 모든 색인 용어 가중치가 1인 문서가 질의에 대한 결과로 가장 적합하다고 결정된다. OR 질의에 대해서는 (0, 0, …, 0) 좌표에서 떨어진 거리를 내림차순으로, AND 질의에 대해서는 (1, 1, …, 1) 좌표에서 떨어진 거리를 오름차순으로 정렬해서 문서 가중치를 부여한다.

P-norm 모델에서 사용하는 일반적인 질의 형태는 식 (1)과 같다. 상관 계수 p는 연산에 대한 제한 정도를 나타낸다.

$$Q_{OR_p} = (A_1, a_1)\ OR_p\ (A_2, a_2)\ OR_p\ \ldots\ OR_p\ (A_n, a_n) \qquad (1)$$

$$Q_{AND_p} = (A_1, a_1)\ AND_p\ (A_2, a_2)\ AND_p\ \ldots\ AND_p\ (A_n, a_n)$$

또한 질의 용어에도 상대적인 중요도에 따라 가중치 a_i를 부여한다. P-norm 모델에 대한 질의와 문서 사이의 유사도는 식 (2)를 사용해서 계산한다.

$$\begin{aligned}
SIM(Q_{or\ p}, D) &= \sqrt[p]{\frac{a_1^p d_{A1}^p + a_2^p d_{A2}^p + \cdots + a_n^p d_{An}^p}{a_1^p + a_2^p + \cdots + a_n^p}} \\
SIM(Q_{and\ p}, D) &= 1 - \sqrt[p]{\frac{a_1^p (1 - d_{A1})^p + a_2^p (1 - d_{A2})^p + \cdots + a_n^p (1 - d_{An})^p}{a_1^p + a_2^p + \cdots + a_n^p}} \qquad (2) \\
SIM(Q_{not}, D) &= 1 - SIM(Q, D)
\end{aligned}$$

3.1.2 벡터공간 모델

표준 불린 모델은 질의나 문서의 키워드에 모두 이진(binary) 가중치를 할당하는데 비해 벡터공간 모델은 질의나 문서의 키워드에 이진 값이 아닌 적절한 가중치를 할당할 수 있다. 또한 불린 모델에서 할 수 없는 부분 매칭이 가능하며, 따라서 질의와 문서의 유사도에 따라 랭킹을 줄 수 있다. 벡터 공간 모델에서 모든 색인어는 서로 독립이라고 가정을 하며 질의와 문서는 모두 식 (3)과 같은 벡터로 표현된다.

$$d_i = (w_{i1}, w_{i2}, \cdots, w_{in}) \qquad (3)$$

위 표현에서 d_i는 문서 또는 질의를 표현하며, w_{ik}는 문서 d_i에서의 색인어 t_k에 대한 가중치 값이다. 문서 i에 나타나지 않는 색인어에는 가중치 0이 할당된다.

문서 또는 질의 벡터들이 형성된 이후의 검색과정은 벡터의 연산에 의해 이루어진다. 문서 d가 $(w_{i1}, w_{i2}, \cdots, w_{in})$로 표현되고 질의 q가 $(w_{i1}, w_{i2}, \cdots, w_{in})$로 표현되었을 때, 문서 d와 질의 q사이의 벡터 유사도는 아래 식 (4)과 같이 계산되며, 이를 코사인 유사도(cosine coefficient similarity)라고 한다.

$$\mathrm{sim}(d_i, q) = \frac{\vec{d_i} \cdot \vec{q}}{|\vec{d_i}||\vec{q}|} \tag{4}$$

질의와 문서의 유사도 측정에 있어서 문서 값은 색인어들의 가중치에 의해 결정되기 때문에 가중치 부여 기법은 검색 효과에 영향을 미치는 중요한 요소이다. 정보 검색에 관한 많은 연구들은 색인어에 가중치를 부여하기 위하여 출현빈도(term frequency), 문서빈도(document frequency), 정규화(document length normalization)의 세 가지 요소를 고려한다. 출현빈도는 문서 내에서 자주 출현하는 색인어에 보다 높은 가중치를 부여한다. 문서빈도는 전체 문서들 중에서 적은 수의 문서에 출현하는 색인어에 보다 높은 가중치를 부여한다. 그리고 정규화는 문서 집합에 있는 모든 벡터들의 길이를 같게 만드는 것으로써 작은 크기의 문서가 문서 값 계산에 있어서 불공평하게 취급되는 것을 피하도록 한다.

3.1.3 확률 모델

확률 모델의 개념은 1960년대 초에 논의되기 시작했으며[Maron and Kuhns, 1960], 현재 주로 사용되고 있는 모델은 Robertson에 의해 정립되었다[Robertson and Sparck, 1976].

확률 모델을 개념적으로 설명하면 특정한 질문에 대해 각 문서가 관련이 있을 확률과 관련이 없을 확률을 산출하여 관련 확률이 관련이 없을 확률보다 큰 문서를 검색하는 것이다. 확률검색의 기본적인 가정은 적합성 정보가 미리 준비되어 있다는 것이다. 다시 말해 적합 문서와 부적합 문서에 대한 결정이 전체 문서집합에 대해서나 또는 일부 집합에 대해 검색 이전에 내려져 있다는 것이다.

대부분의 확률 모델은 적합 문서와 부적합 문서에서의 색인어 출현정보에 기초하고 있다. 각 문서 X가 (5)와 같은 색인어 벡터로 표현되어 있다면, x_i는 i번째 색인어의 유무를 나타내는 것으로 $x_i = 0$ 또는 1의 값을 갖게 된다.

$$X = (x_1, x_2, \cdots, x_n) \tag{5}$$

이때 문서가 질의를 만족하는 정도를 계산하는 유사도함수 $g(X)$는 다음 식 (6)과 같다.

$$g(X) = \log \frac{\Pr(X|rel)}{\Pr(X|nonrel)} \tag{6}$$

위 식에서 $\Pr(X|rel)$은 적합 문서가 문서 벡터 X로 표현될 확률이며, $\Pr(X|nonrel)$은 비적합 문서가 문서 벡터 X로 표현될 확률이다. 다시 말해, 문서 X가 질의를 만족시킬 가능성이 클수록 $g(X)$의 값도 증가한다. 여기에서 유사도함수 $g(X)$의 값을 계산하기 위해서는 $\Pr(X|rel)$과 $\Pr(X|nonrel)$의 값을 추정해야 한다.

문서 X의 변수 x_i에 대해 다음과 같은 가정을 하면, 유사도 함수 $g(X)$는 식 (7)과 같이 변환될 수 있다.

- 문서 X의 변수는 상호 독립이다.
- 색인어 가중치는 0 또는 1의 이진 값을 가진다.

$$g'(X) = \sum_{i=1}^{n} x_i \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)} + C$$

$$p_i = \Pr(x_i = 1|rel) \qquad (7)$$

$$q_i = \Pr(x_i = 1|nonrel)$$

위 식에서 상수 C는 질의에 대해 일정한 값이므로 문서의 랭킹에 영향을 미치지 않는다. 따라서 $g'(X)$의 값을 산출하기 위해서는 x_i에 대한 p_i와 q_i값만이 필요하다. $g'(X)$는 질의에 대한 적합 문서들이 알려져 있는 상태에서는 유용하게 사용될 수 있다. 그러나 질의에 대한 적합 문서들이 알려지지 않은 상태, 예를 들어 사용자가 처음으로 질의를 입력하는 최초 검색에 대해서는 어려움이 있다. 이러한 경우에는 p_i와 q_i에 대한 추정 값을 적용해 검색하는데 이에 관한 Croft의 초기 연구에서는 p_i를 0.5로 q_i를 n_i/N으로 추정하였다[Croft and Harper, 1979].

3.2 검색 엔진(SCAIR)

본 논문에서 사용된 검색 엔진은 정보 검색 학회인 TREC에 참여하기 위해 만들어진 SCAIR(SCAI Information Retrieval Engine)를 기반으로 하고 있다 [Shin and Zhang, 1998]. SCAIR는 기본적으로 벡터공간 모델을 바탕으로 한다[Salton et al., 1975]. 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. 참고로 여기에서의 텀은 앞에서 언급한 색인어와 비슷한 의미로 볼 수 있다. 따라서 전체 문서의 집합은 텀과 문서의 행렬이 된다. 문서와 마찬가지로 질의 역시 단어로 구성된 문장이므로 텀들의 리스트로 표현된다.

SCAIR에서는 주어진 문서에 대해 기본적으로 식 (8)에 나타나 있는 $tf \cdot idf$를 이용, 색인을 한다[Salton, 1989]. $tf \cdot idf$에서 tf는 텀 빈도수(term frequency)를 나타내며 idf는 역 문서 빈도수(inverse document frequency)를 나타낸다. 주어진 질의에 대한 가중치 역시 $tf \cdot idf$를 이용하여 계산한다.

$$w_{di} = tf_{di} \cdot \log\left(\frac{N}{df_i}\right) \tag{8}$$

w_{di} : 문서 d에서 i번째 텀의 가중치

tf_{di} : 문서 d에서 i번째 텀의 빈도수

N : 전체 문서의 수

df_i : i번째 텀을 포함한 문서의 수

본 연구에서는 SCAIR를 확장하여 문서 가중치 산정에서 $tf \cdot idf$ 이외에 확

률 모델 방법인 추론 네트워크 모델[Broglio et al., 1995]과 2-포아송 모델[Robertson et al., 1995]을 지원하도록 구현하였다. 추론 네트워크 모델과 2-포아송 모델에서의 가중치 계산을 위한 식은 각각 (9)와 (10)와 같으며, 여기에서의 질의 가중치는 tf를 이용한다.

$$w_{di} = \left[0.4 \times H + 0.6 \times \frac{\log(tf_{di} + 0.5)}{\log(\max tf_{di} + 1.0)}\right] \times \frac{\log \frac{N}{n}}{\log N} \quad (9)$$

w_{di} : 문서 d에서 i번째 텀의 가중치

tf_{di} : 문서 d에서 i번째 텀의 빈도수

N : 전체 문서의 수

n : 텀 i를 가지는 문서의 수

H : 1

$$w_{di} = \frac{tf_{di}}{k_1\left((1-b) + b\frac{\text{document length}}{\text{average document length}}\right) + tf_{di}} \times \log \frac{N - n + 0.5}{n + 0.5} \quad (10)$$

w_{di} : 문서 d에서 i번째 텀의 가중치

tf_{di} : 문서 d에서 i번째 텀의 빈도수

N : 전체 문서의 수

n : 텀 i를 가지는 문서의 수

k_1 : 2.0

b : 0.75

질의와 문서의 유사도는 질의 벡터 및 문서 벡터를 내적 하여 결정하며 이는 식 (11)에 나타나 있다.

$$sim(d,q) = \sum_{k=1}^{n} (\alpha_{dk} \times w_{dk}) \times w_{qk} \qquad (11)$$

w_{dk} : 문서 d에서 k번째 텀의 가중치

w_{qk} : 질의 q에서 k번째 텀의 가중치

α_{dk} : 문서 d의 k번째 텀에 대한 태그 상수

위 식에서 상수 a_{dk}는 텀 k가 속해 있는 모든 HTML 태그에 대응하는 태그 가중치의 곱으로 결정된다. 단, 텀 k가 어느 태그에도 속해 있지 않다면 a_{dk}는 1.0이 된다. 주어진 질의와 모든 문서에 대한 유사도가 결정되면 이를 내림차순으로 정렬해 리스트의 형태로 출력한다.

한편 본 논문에서의 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 추가되었다. 한 문서를 구성하는 텀 중에 특정 태그에 속해 있는 텀들은 인덱싱 과정 중 별도로 표시를 한다. 그리고 태그에 대한 가중치는 유사도를 결정하는 데 적용되어 상수 α의 형태로 식 (11)에 적용된다.

4. 태그 중요도 학습을 위한 유전 알고리즘

본 논문에서 제안하는 웹 문서 검색에서의 태그의 중요도에 따른 가중치 적용은 유전 알고리즘을 바탕으로 한 가중치 학습에 기반하고 있다. 유전 알고리즘은 NP문제 등과 같이 고난도의 최적화 문제를 해결하는데 좋은 대안으로 사용되고 있다. 유전 알고리즘은 자연 진화의 원리를 문제 풀이 또는 모의 실험에 이용하는 연구 방법인 진화 연산(evolutionary computation)의 대표적인 한 분야이다[Zhang, 1995]. 본 장에서는 먼저 유전 알고리즘에 대한 용어 및 기본 알고리즘을 소개하며, 태그 가중치를 이용한 학습 알고리즘을 설명한다.

4.1 유전 알고리즘

진화 연산은 커다랗게 유전 알고리즘(genetic algorithm), 유전 프로그래밍(genetic programming), 진화 프로그래밍(evolutionary programming), 진화 전략(evolutionary strategy)으로 나뉘어 지며, 이는 모두 선택, 교차, 돌연변이의 자연 진화 원리를 바탕으로 하고 있다.

유전 알고리즘에서의 기본 요소는 생물학에서 말하는 유전 물질인 DNA, 즉 염색체(chromosome)이다. 개체(individual)들은 교차(crossover)에 의해 염색체를 부분 결합하고 돌연변이(mutation)에 의해 미소하게 변화된 새로운 염색체를 가진 새로운 개체들을 만들어내고, 개체들은 환경에 적응하는 정도에 따라 선택적으로 번성한다. 유전 알고리즘의 기본 구조는 이러한 생물의 진화 과정을 문제 해결 과정으로 옮겨 놓은 것이다. 문제 상의 임의의 해는 유전 알고리즘이 이해하는 형태로 표현되어 염색체로 불린다. 자연계에서처럼 가변적이고

제한이 없는 개체들의 집단 대신, 유전 알고리즘에서는 대부분 정해진 수의 염색체 집단을 운영하는데 이를 해집단(population)이라 하며 유전 알고리즘은 이 복수 개의 해들을 유지하면서 작동된다. 염색체상의 각 인자는 유전자(gene)라 하며 유전 알고리즘에서는 인자 하나가 유전자가 된다. 생물학에서 유전자형(genotype)은 유전자의 조합이고, 표현형(phenotype)은 관찰되는 형질을 말한다. 유전 알고리즘에서도 비슷하게 염색체 그 자체를 유전자형이라 하고, 이와 대응되는 해의 성격 또는 품질 등을 표현형이라 한다.

그림 1은 일반적인 유전 알고리즘의 구조를 보인다[Ballard, 1997]. 유전 알고리즘은 대부분 정해진 수의 해로 구성되는 해집단을 갖는다. 그림에서 해집단의 해의 수는 n이며, 먼저 n개의 해를 임의로 생성한다. 이 n개의 해집단으로부터 k개의 새로운 해를 만들어 내는데, 각각의 해는 선택, 교차, 돌연변이의 단계를 거쳐 만들어진다. 이렇게 만들어진 k개의 해는 해집단 내의 k개의 해와 대치된다. 이 과정은 정지 조건이 만족될 때까지 수행되며, 정지 조건이 만족된 후 해집단에서 가장 좋은 해가 리턴된다. 여기에서 상수 k는 해집단이 한 번에 얼마나 많이 대치되느냐를 결정하는데 k/n를 세대차(generation gap)라 하며, 세대차가 1에 가까운 경우, 다시 말해 절대 다수의 해가 대치되는 경우를 세대형 유전 알고리즘(generational GA)이라고 한다. 한편, 세대차가 $1/n$에 가까운 경우, 즉, 새로운 해가 생기는 대로 해집단에 넣어주는 방식을 안정 상태 유전 알고리즘(steady-state GA)이라고 한다. 안정 상태 유전 알고리즘은 대체로 해집단을 빨리 수렴시키는 경향이 있으며 대신 전역해가 아닌 지역해로 수렴할 가능성이 크다.

Choose a population size.
Choose the number of generations N_g.
Initialize the population.
Repeat the following for N_g generations:

1. Select a given number of pairs of individuals from the population probabilistically after assigning each structure a probability proportional to observed performance.
2. Copy the selected individuals(s), then apply *operators* to them to produce new individual(s).
3. Select other individuals at random and replace them with the new individuals.
4. Observe and record the fitness of the new individuals.

Output the fittest individual as the answer.

그림 1: **유전 알고리즘**

연산자들 중 선택 연산자는 교차를 위해 해집단에서 임의의 해를 선택하는 연산자이며 일반적으로 우수한 해에게 선택될 확률을 높게 준다. 이렇게 선택된 해는 부모해(parent)**가 된다. 교차는 두 개의 부모해로부터 자식해**(offspring)**를 만들어 내는 연산자이다. 교차는 유전 알고리즘의 대표적인 연산자로서 유전 알고리즘의 성능에 큰 영향을 미친다. 교차 연산의 방법은 다양하게 존재하지만 일반적인 특징은 복수 개의 해를 결합하여 하나를 만든다는 것이라고 말할 수 있다. 돌연변이는 자연계의 돌연변이처럼 해를 임의로 변형시키는 연산자이다. 교차가 부모해에 있는 속성들을 부분적으로 이용하는 역할을 하는 반면, 돌연변이는 부모해에 없는 속성을 다음 세대에 도입하는 역할을 한다.**

유전 알고리즘의 정지 조건도 다양하게 줄 수 있는데 가장 대표적인 두 가

지를 든다면 루프(loop)을 일정 횟수만큼 수행한 다음 정지시키는 방법과 해집단의 다양성이 일정 수준 이하로 떨어지는 시점에 정지시키는 방법을 들 수 있다. 다양성이 떨어지는 것을 판단하기 위해서는 해집단 내의 염색체들 중 일정 비율에 해당하는 부분이 똑같은 지를 확인하는 경우가 일반적이다.

4.2 태그 가중치 학습 알고리즘

이미 기술한 것처럼 웹 문서의 특징은 화면 표시 및 하이퍼링크(hyperlink) 등을 위한 태그를 담고 있다는 것이다. 태그의 중요도를 판단하고, 그에 대한 가중치를 학습하는 방법에는 여러 가지 것들이 있을 수 있다. 그러나 이 문제에 있어서의 중요한 특징은 중요도에 따른 가중치를 계산하는 결정적인(deterministic) 방법이 없다는 것이다. 따라서 본 논문에서는 진화연산을 이용해 주어진 문제를 해결하는 방법을 제시하였다.

태그 가중치 학습 문제를 유전 알고리즘에 적용하기 위해서는 염색체의 표현 및 세대별 진화를 위한 염색체의 우열 판단 과정에 대한 척도가 필요하다. 제안된 알고리즘에서 하나의 HTML 태그는 한 유전자와 대칭 되며 염색체는 유전자의 집합, 즉 HTML 태그들로 구성된다. 염색체는 태그의 가중치 열로 구현되며 각 태그의 가중치는 실수로 염색체에 포함된다. 가중치 학습에 있어서 해 집단에 대한 학습 알고리즘은 그림 2와 같으며 염색체의 표현과 학습에 대한 개괄은 그림 3에 나와 있다.

초기 염색체 생성

for g = 1 to g_{max}

적합도 함수에 의해 모든 염색체를 평가

for i = 1 to M

두 염색체 p_1, p_2 선택

$offspring_i$ = crossover(p_1, p_2)

$offspring_i$ = mutation($offspring_i$)

end for

M 개의 염색체를 자식해로 교체

end for

N 세대 이후의 염색체 중 최상의 염색체를 return

그림 2: 태그 가중치 학습 알고리즘

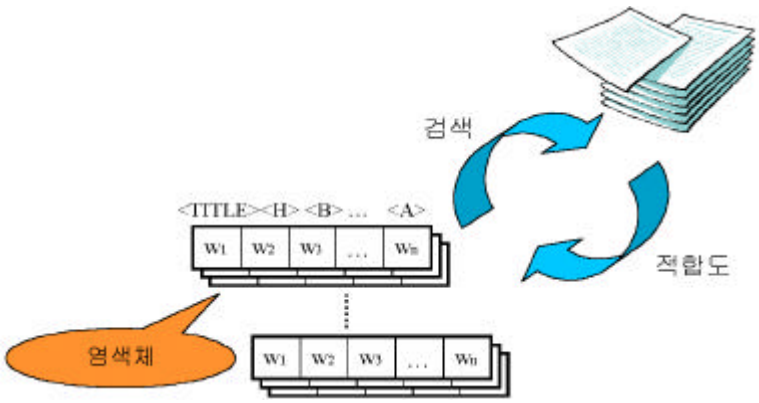

그림 3: 염색체 구조 및 표현형

초기 해 집단은 임의로 구성된 염색체들로 구성된다. 다음 세대로 진화하기 위한 부모해의 선택은 다음과 같이 이루어진다. 부모 염색체는 해집단 중 품질이 좋은 상위 반절에 속하는 개체 중에서 임의로 선택되며 품질은 적합도 함

수(fitness function)에 의해 결정된다. 적합도 함수는 태그 가중를 이용한 검색 결과의 성능을 측정하는 함수이다. 성능을 측정하는 방법으로는 TREC의 검색 결과 측정 방법인 11포인트 평균 정확율(11-point average precision) 값을 사용하였다[Voorhees and Harman, 1999]. 11포인트는 재현율(recall)을 0.0에서 0.1사이 값의 11개 지점으로 나누어 각 포인트에서의 정확율(precision)을 평균하는 방법이다. 정확율 P는 검색 결과 문서 중 질의와 관련된 문서(relevant document)의 수의 비율을 말하며, 재현율 R은 전체 관련 문서 중 검색 결과에 포함된 문서의 수를 나타내는 비율이다. 정확율과 재현율에 대한 수식은 (12), (13)과 같으며, 적합도 함수는 (14)와 같이 정의된다. 염색체의 적합도 함수로 11포인트 평균 정확율을 정한 이유는 태그의 중요도 가중치 학습의 목적이 문서검색 결과를 향상시키는데 있기 때문이다.

$$P = \frac{Number\ of\ retrieved\ relevant\ documents}{Total\ number\ of\ retrieved\ documents} \tag{12}$$

$$R = \frac{Number\ of\ retrieved\ relevant\ documents}{Total\ number\ of\ relevant\ documents\ in\ collection} \tag{13}$$

$$fitness = \frac{1}{\sum_{i=1}^{N} r(d_i)} \sum_{i=1}^{N} \frac{1}{i} \sum_{j=1}^{i} r(d_j) \tag{14}$$

$r(d)$: 검색된 문서 d에 대한 관련성 값을 리턴하는 함수

(만약 d가 관련문서라면 1을 리턴하고, 그렇지 않다면 0을 리턴한다)

N : 검색된 전체 문서의 수

선택된 부모 염색체는 교차를 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[Michalewicz, 1992]. 교차를 통해 생성된 자식 해들은 해 집단 중 적합도가 낮은 반절의 염색체와 대체된다. 돌연변이(mutation) 연산은 해 집단 중 임의의 해와 염색체의 임의의 위치의 값을 바꿈으로 이루어지며 돌연변이를 생성시킨다. 그림 4와 그림 5는 본 알고리즘에서 교차와 돌연변이가 일어나는 과정을 시각화한 것이다.

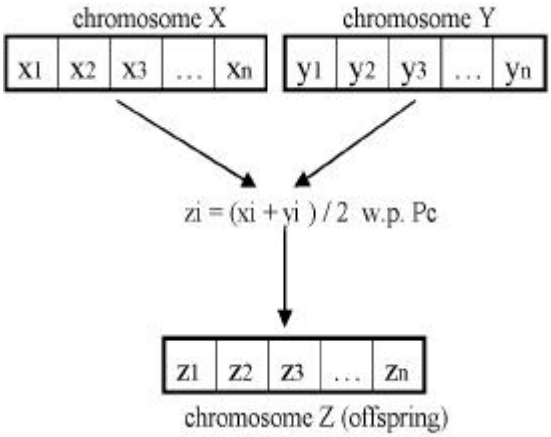

그림 4: 염색체 교차과정

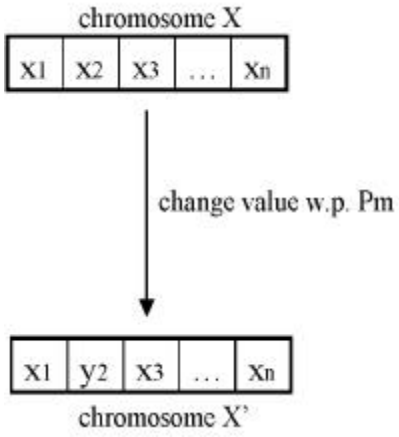

그림 5: 염색체 돌연변이과정

5. 실험 및 결과

웹 문서의 태그는 유전 알고리즘에 의해 그 중요도에 따라 가중치가 학습되며 학습된 가중치는 문서에 대한 태그 가중치 상수로 적용되어 정보 검색을 수행한다.

실험에서는 웹 문서의 모든 태그로 염색체를 구성하지 않고, 일부를 유전자로 선택해 염색체를 구성하였다. 본 논문에서는 HTML에서 문서의 내용에 대해 의미가 높을 것이라고 판단된 타이틀(title), 헤더(header), 볼드(bold), 이탤릭(italic), 앵커(anchor) 등을 선택하였다. 위 다섯 가지 구조에 대한 태그명은 각각 순서대로 <TITLE>, <Hx>, <B>, <I>, 그리고 <A>이다. 타이틀과 헤더 태그는 일반적으로 문서의 내용을 요약 또는 대표하는 용어에 대해 사용되기 때문에 선정했으며, 볼드와 이탤릭 태그는 어떤 단어 또는 구문을 강조하는데 사용되고, 따라서 문서의 일부 또는 전체 단락의 핵심 내용과 연관된 내용일 가능성이 크다고 가정했다. 앵커는 다른 문서와의 링크를 위해 사용되며, 제 2장의 하이퍼링크에 대한 관련연구에서 언급한 두 문서의 링크는 상호 연관성이 높다는 것을 의미한다는 가정을 바탕으로 앵커 태그에 의해 지정된 단어 또는 구문은 해당 문서에 의해 링크된 관련 문서를 대표하는 문구라는 가정 하에 선정되었다.

염색체는 위의 가정으로 선택된 다섯 개의 태그로 정의되고 해집합이 구성되었으며, 제안된 방법에 대해 한 개의 인공 데이터 집합과 정보 검색 학회인 TREC에 사용된 데이터 집합을 대상으로 실험하였다. 한편, 검색엔진 SCAIR의 문서 가중치 산정은 제공하는 기능 중 $tf \cdot idf$ 함수를 바탕으로 하였다.

5.1 실험 I : 인공 데이터

5.1.1 데이터 설명

실험 I은 인공 데이터를 작성한 후 이를 검색 대상으로 하여 제안된 알고리즘이 태그 정보의 중요도에 따라 가중치를 제대로 학습하는지 여부를 판단하기 위해 실시되었다.

웹 문서는 학회 논문 제출을 요청하는 홈페이지(call-for-papers homepage)와 학회 안내 홈페이지(conference homepage)를 대상으로 총 100개의 문서를 수집하였으며, 내용은 인공지능 관련 학회 및 생명과학, 정보검색 학회 등을 포함한다. 수집된 문서를 바탕으로 서로 다른 태그 구조 정보를 가지는 두 가지 문서 집합을 생성하였다. 문서 집합 A는 수집된 문서를 담은, 즉 어떠한 편집도 하지 않은 문서의 집합이며, 문서 집합 B는 원 문서에 대해 핵심단어를 헤더 태그를 이용해 강조시킨 변형된 문서들로 구성된다. 여기에서 핵심단어의 강조는 문서집합의 크기가 작고, 실험의 목적이 가중치의 학습 여부에 있기 때문에 해당 웹 문서의 주제와 관련 있는 모든 단어 또는 구문에 대해 수정하지 않고 질의와 관련된 부분만 수정을 하였다.

검색을 위해 생성한 질의는 그림 6과 같다. 질의에 있어서 제목(title) 및 설명(description)은 실험 II에 사용된 TREC 데이터 집합의 형식을 사용한 것이다. 검색엔진은 생성한 질의에 대해 빈도수가 높은 단어 또는 불용어 등을 제거함으로써 질의어에서 의미가 없다고 생각되는 단어들을 삭제한 후 남은 단어들을 질의 벡터로 사용한다. 따라서 그림 6 질의에 대한 최종적인 질의어는 ‘genetic’, ‘algorithms’, ‘conference’, ‘especially’, ‘information’, ‘retrieval’ 등 여

섯 개 단어가 된다. 검색엔진에서는 <title>, <desc>, <narr> 등 각 필드에 따른 추가적인 가중치 산정은 하지 않고 모두 같은 비중의 질의어로 간주한다. 한편 질의에 대한 관련 문서 집합은 수작업에 의해 결정했으며, 100개 문서 중 10개의 문서가 질의와 관련이 있다고 판단되었다.

```
<title> genetic algorithms
<desc> Description:
Is there a conference on genetic algorithms,
especially containing call for papers on information retrieval?
<narr> Narrative:
none.
```

그림 6: 실험 I에 사용한 질의문

5.1.2 실험 및 결과

실험 I에서 사용한 학습 알고리즘의 파라미터는 다음과 같이 설정했으며, 문서 집합 A와 B에 대해 동일하게 설정되었다.

- 개체군 크기(population size) : 100
- 세대 수(number of generation) : 30
- 돌연변이 확률(probability of mutation) : 0.04
- 태그 가중치 범위 : 0.0 4.0

실험은 문서 집합 A, B에 대해 각각 10회에 걸쳐 반복 시행되었으며 태그의 중요도에 따른 최적의 가중치는 지정한 세대수가 완전히 지난 후, 남은 해집합에서 가장 적합도가 높게 측정된 염색체를 학습을 통해 얻은 최적의 가중치로

간주하였다.

그림 7은 문서 집합 A와 문서 집합 B에 대한 가중치 학습과정에 대해 세대가 진행함에 따라 변화하는 적합도 값을 나타낸 것이다. 각 세대에 대한 적합도 값은 그 세대에 있어서 해집합에 존재하는 전체 염색체의 적합도 평균을 말한다. 그림에서 보듯이 문서 집합 A와 B에 대한 실험에서 모두 세대가 진행함에 따라 일정한 적합도 값에 해집합이 수렴해 감을 알 수 있다. 한편 결과상에서 문서 집합 A와 문서 집합 B의 적합도 간에 차이를 보이는 것은 문서 집합 B의 경우, 문서의 주요 내용을 헤더 태그로 구조화함으로써 문서 집합 A에 비해 보다 더 구조화가 잘 되어 있기 때문이다. 다시 말하자면 구조화가 더 잘 되어 있음으로 인해 평균적으로 태그 중요도를 이용한 정보 검색에서 보다 더 좋은 검색 성능을 보이게 된 것이다.

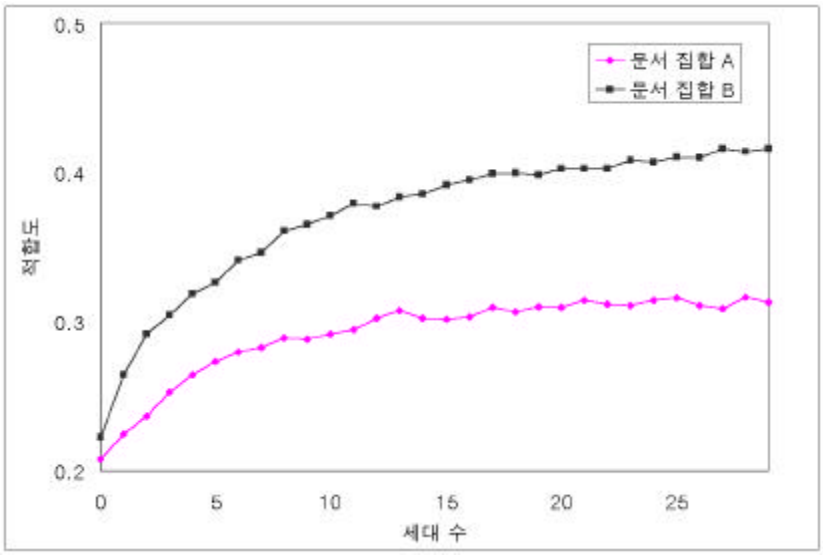

그림 7: 세대에 따른 적합도 변화

30세대에 걸친 가중치 중요도 학습이 지난 후 얻어진 태그에 대한 중요도 평균값은 타이틀, 헤더, 볼드, 이탤릭, 앵커 태그에 대해서 문서 집합 A의 경우 각각 0.77, 0.13, 2.15, 1.16, 0.08이며, 문서 집합 B의 경우 각각 1.12, 3.98, 2.23, 1.80, 0.10이 얻어졌다. 이를 막대그래프로 도식화 한 것이 그림 8, 그림 9에 나타나 있다. 그림에서 보는 것처럼 문서 집합 A에 대해서는 볼드가 가장 중요한 태그로 학습되었지만, 문서 집합 B에 대해서는 헤더가 가장 중요한 태그로 학습되었다. 이는 헤더 태그로 내용을 강조한 문서 집합 B에 대한 태그 중요도 학습이 제안된 알고리즘에 의해 올바르게 이루어 졌음을 의미한다.

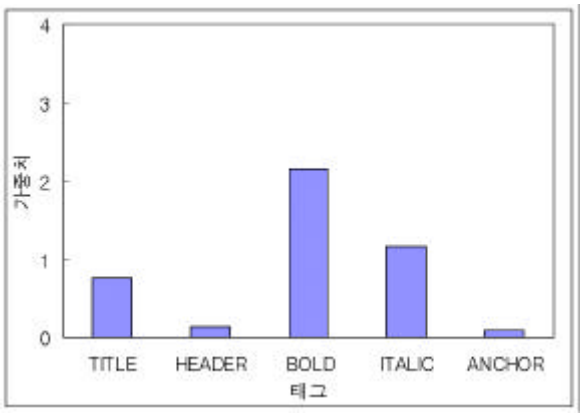

그림 8: 문서 집합 A에 대한 태그 중요도

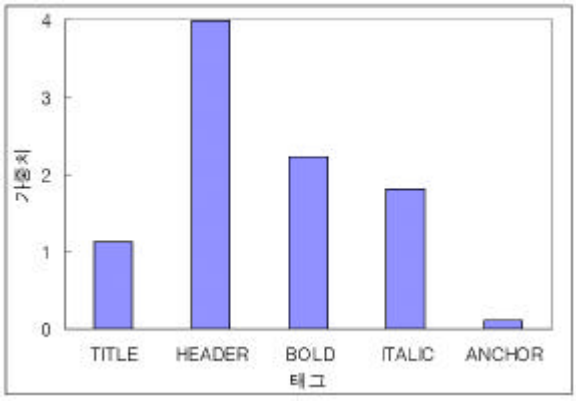

그림 9: 문서 집합 B에 대한 태그 중요도

5.2 실험 II : TREC 데이터

5.2.1 TREC

실험 II는 태그의 중요도를 제안된 알고리즘에 의해 학습한 후, 얻어진 태그 중요도 가중치 값을 바탕으로 웹 문서 검색을 수행하였을 때의 변화를 관찰하기 위해 실시되었으며, 정보 검색 학회인 TREC에 사용되는 데이터 집합을 대상으로 하였다.

TREC(Text REtrieval Conference)은 미국 NIST에서 주관하는 국제적인 정보검색 평가대회로 매년 여러 분과의 문제에 대한 경기(competition)를 바탕으로 진행되고 있다[NIST]. TREC의 분과로는 본 논문의 실험에서 사용한 웹 트랙(web track) 외에 필터링 트랙(filtering track), 질의답변 트랙(question answering track), 교차언어 트랙(cross-language track) 등 정보 검색에 있어서 관심의 대상이 되고 있는 여러 분야를 망라하고 있다.

실험 II에서 사용한 TREC 데이터는 TREC의 분과 중 웹 트랙에 사용되는 문서집합을 이용하여 실시되었다. 문서 집합은 1999년에 열렸던 TREC-8에 사용된 웹 문서(WT2g)이며 이 문서는 Internet Archive [Archive]라는 인터넷 사이트에 의해 수집된 모든 웹 페이지에서 일부 문서를 선택하여 구성한 것이다. 전체 문서 집합의 크기는 2기가 바이트에 이르며, 총 247,491개의 웹 페이지를 포함하고 있다.

<title> foreign minorities, Germany
<desc> Description:
What language and cultural differences impede the integration of foreign minorities in Germany?
<narr> Narrative:
A relevant document will focus on the causes of the lack of integration in a significant way; that is, the mere mention of immigration difficulties is not relevant. Documents that discuss immigration problems unrelated to Germany are also not relevant.

그림 10: TREC 토픽의 예

웹 트랙에서의 질의는 토픽(topic)이라고 불리며, 한 토픽은 제목(title), 설명(description), 상세 설명(narrative)의 세 필드로 구성된다. 그림 10은 웹 트랙 질의의 한 예를 보인 것이다. 제목 필드는 매우 짧은 질의를 대상으로 한 실험을 가능하기 위해 디자인이 되었으며 토픽을 가장 잘 나타내는 단어들로 구성되어 있다. 제목에 들어갈 수 있는 단어의 최대 수는 세 단어로 한정되어 있다. 설명 필드는 토픽에 대해 토픽을 설명할 수 있는 한 문장으로 구성된다. 즉, 타이틀 필드를 좀 더 자세히 설명할 수 있는 한 개의 문장이 설명 필드에 삽입된다. 설명 필드의 특징은 타이틀 필드에 나타나 있는 모든 단어를 포함하고 있다는 것이다. 마지막으로 상세 설명 필드는 토픽에 대해 가장 자세히 설명되어 있는 필드이며, 질의에 대한 문서를 찾는데 있어서 관련 문서가 포함해야 하는 필수적인 설명이 모두 포함되어 있는 부분이다.

웹 트랙에서의 질의에 대한 답, 즉 관련 문서의 집합은 질의를 만드는 과정에서 미리 정해지지 않는다. 웹 트랙에서 관련 문서의 집합은 참가 그룹이 제

출한 질의에 대한 검색 결과를 바탕으로 만들어진다는 특징을 갖는다. 정보 검색에서 이렇게 관련 문서 집합을 찾는 것을 풀링 방법(pooling method)이라고 한다[Voorhees and Harman, 1999][Zobel, 1998]. 풀링 방법에 있어서 관련 문서를 찾기 위해 관련 문서의 후보가 될 수 있는 문서들을 모아 풀(pool)을 구성하는데, 이 후보 문서들은 참가 그룹이 제출한 검색결과를 바탕으로 한다. 다시 말해, 참가 그룹이 제출한 검색 결과 중 상위에 해당하는 일정 수의 문서를 모든 그룹들에 대해 모은 후 이를 풀에 집어넣는다. 웹 트랙에서는 각 토픽에 대해 참여 그룹이 제출한 검색 결과 중 상위 100개의 문서를 관련 문서의 후보 집합으로 보고 이를 풀에 추가한다. 이 때 관련 문서를 찾는 과정은 구성된 풀의 범위 내로 한정된다. 따라서 판단자로 정해진 사람들은 풀 안에 존재하는 문서들에 대해서만 질의에 대한 관련 문서에 해당되는지 판단하게 되며, 이를 거친 문서들만이 최종적으로 관련 문서로 정해지게 된다. 웹 트랙의 풀링 방법에 있어서 후보 관련 문서를 상위 일정수의 문서로 하는 이유는 모든 참여 시스템이 가장 관련문서일 가능성이 높다고 판단된 문서들을 상위에 랭크하는 형식으로 결과를 출력하기 때문이다.

한편 풀링 방법은 위에 설명한 내용에서 드러나는 것처럼 문서 집합 전체에 대해서 모든 관련 문서를 찾지 못하는 한계가 있다. 따라서 풀링 방법에 의해 각 검색 엔진에 대한 절대적인 성능 치수를 측정하는 것은 어렵지만, 검색 엔진을 상호 비교하는데 있어서 그 상대적인 차이를 보여줄 수는 있다.

5.2.2 실험 및 결과

실험은 전 절에서 언급한 웹 트랙 데이터 집합을 사용하였으며, 웹 문서는 WT2g 데이터 집합, 질의는 TREC-8에 사용되었던 번호 401 420번 토픽을 대상으로 하였다. 토픽은 제목과 서술, 상세 서술 부분으로 나뉘어 지지만 실험에 사용한 부분은 토픽 중 제목과 서술 부분이며, 각 필드에 대한 별도의 의미부여 없이 같은 비중으로 고려해 질의로 사용하였다. 선택된 전체 20개의 토픽 중, 태그 중요도 학습을 위하여 401 410번 토픽을 사용했으며, 학습된 태그 가중치를 이용한 검색을 위해서는 411 420번까지의 토픽이 사용되었다.

태그 가중치 학습 과정과 학습된 가중치에 대한 문서검색에 있어서 모든 검색 결과는 문서의 질의에 대한 관련성에 따라 토픽당 200개의 문서를 랭킹 출력하며, 검색 성능 판단은 11포인트 평균 정확율을 기준으로 측정하였다. 검색되는 문서를 200개로 제한한 이유는 일반적으로 사용자는 질의에 대한 결과로 작은 개수의 문서를 받기 원하기 때문이다.

실험 II에서 사용한 학습 알고리즘의 파라미터는 다음과 같이 설정하였다.

- 개체군 크기(population size) : 100
- 세대 수(number of generation) : 25
- 돌연변이 확률(probability of mutation) : 0.04
- 태그 가중치 범위 : 0.0 4.0

실험은 문서 집합에 대해 20회 반복 시행하였으며 태그의 중요도에 따른 최적의 가중치는 실험 I과는 다르게 20세대 이후의 모든 해집합을 대상으로 가장 적합도가 높은 염색체를 학습된 최적 가중치 후보로 간주하였다. 태그 가중치를 적용한 검색 방법은 다음과 같은 과정으로 이루어진다.

1. 학습을 위한 질의 데이터에 대해 제안된 학습 알고리즘을 이용하여 태그 가중치를 얻는다.
2. 일정 횟수만큼 1의 과정을 반복한 후 얻은 태그 가중치 중에서 가장 높은 적합도 값을 가지는 염색체, 즉 태그 가중치 집합을 최적의 가중치 값으로 간주한다.
3. 최적의 가중치 집합에서 그 값이 1.0보다 큰 가중치를 검색을 위한 태그 가중치로 선택한다.
4. 질의에 대해 태그 가중치를 이용하지 않고 일반 검색을 시행, 200개의 문서를 검색한다.
5. 검색된 문서에 대해 선택된 가중치 및 가중치 값을 이용하여 그 유사도 값을 조정한다.
6. 변경된 유사도 값을 기준으로 검색 문서를 재정렬을 한다.

위 과정을 그 성격을 기준으로 나누면 다음과 같이 태그 가중치 학습(1, 2) 과정, 특징 선택(feature selection) 과정 및 검색 과정(4, 5, 6)으로 나누어 생각할 수 있다.

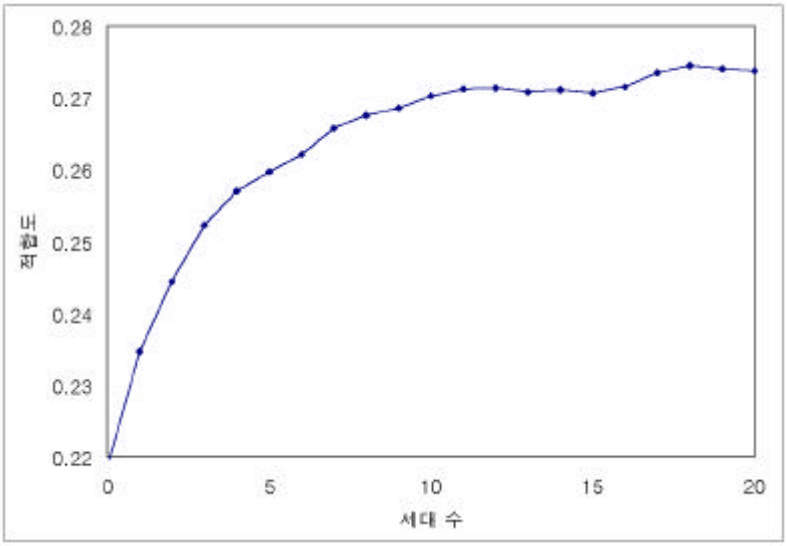

그림 11: 세대에 따른 적합도 변화

그림 11은 제안된 알고리즘에 의한 학습 결과이다. 세대가 진행됨에 따라 해 집합 전체의 평균 적합도 값이 변화하는 과정을 그래프로 나타낸 것이다. 세대가 증가함에 따라 8세대까지는 적합도가 급격히 증가하다가 그 이후에는 완만하게 증가함을 보이고 있다. 이는 사용한 진화연산의 산술교차를 원인으로 생각할 수 있다. 산술교차는 적합도가 높은 두 염색체의 평균값으로 자식 해를 생산하며, 자식해로 교체되는 염색체 수는 해 집단의 절반이다. 따라서 세대 초기의 해 집단은 한 세대의 변화에도 적합도가 우수한 해 집단의 염색체에 빠르게 수렴하는 경향을 보인다. 반면 세대가 어느 정도 진행한 후에는 해 집단의 다수가 이미 기존의 우수한 염색체에 가깝게 수렴을 했으므로 적합도 값의 증가가 둔화되는 것이다.

HTML 태그	가중치
<TITLE>	0.6
<Hx>	1.6
<B>	0.7
<I>	0.6
<A>	1.6

표 1: 학습된 태그 중요도 가중치

앞에 설명한 유전 알고리즘에 의한 학습과정을 통해 얻은 태그 중요도에 따른 가중치는 표 1과 같다. 타이틀, 헤더, 볼드, 이탤릭, 앵커 태그에 대해서 각각 0.6, 1.6, 0.7, 0.6, 1.6의 값이 태그 가중치로 학습되었다. 따라서 본 실험에 의한 태그의 중요도는 헤더,앵커 > 이탤릭 > 타이틀,볼드의 순서가 된다. 이것은 헤더, 앵커 태그에 의해 구조화된 문서상의 단어 또는 문구는 문서 내용에 대한 키워드를 포함할 가능성이 큰 반면, 상대적으로 타이틀 태그의 경우는 그 가능성이 작다는 의미이다. 여기에서 유의할 점은 웹 문서의 타이틀 태그 가중치 값이 작다는 것인데, 이 것은 타이틀 태그에 의해 구조화된 문구가 해당 문서를 대표하는 제목이 될 수는 있지만 헤더나 앵커에 의해 구조화된 문구 이상의 추상화된 개념으로 인해 질의와는 거리가 멀어졌을 가능성 또는 사용자 질의어의 경향이 문서 전체보다는 문서 부분에 국한되어 있을 가능성 등으로 분석할 수 있다.

	11포인트 평균 정확율
일반 검색	0.2383
태그 가중치 적용 검색	0.2503

표 2: 평균 정확율 비교

표 2는 일반 검색의 경우와 태그 중요도에 따른 가중치를 적용한 검색에 대

한 결과를 나타낸 것으로 11포인트 평균 정확율을 비교하고 있다. 검색된 문서에 대해 일반 검색의 경우 0.2383, 태그 가중치 적용 검색 결과의 경우 평균 0.2503의 평균 정확율을 나타내었다.

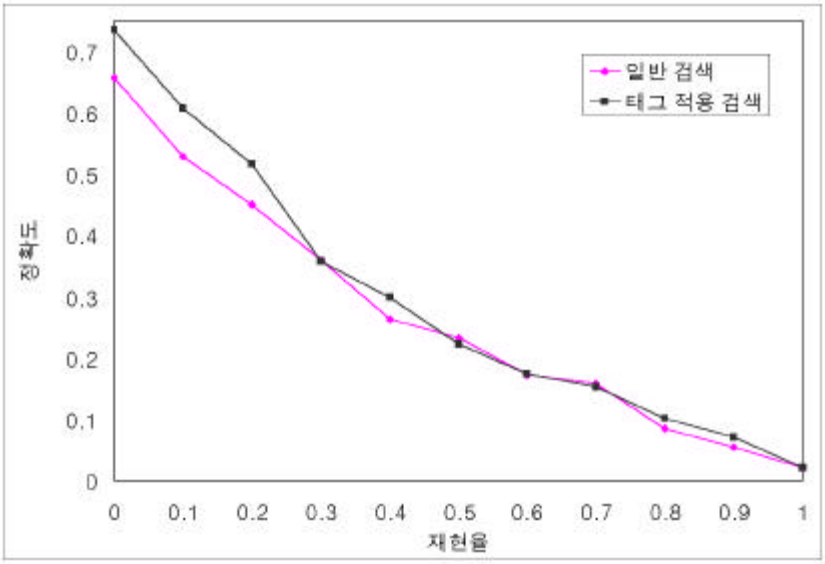

그림 12: 재현율에 따른 정확율

그림 12는 태그 중요도 가중치를 적용했을 경우와 그렇지 않았을 경우를 평균 정확율-재현율(Average Precision-Recall)로 비교한 그림이다. 재현율이 0.3 이하일 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 3은 재현율에 대한 평균 정확율의 변화를 표로 비교한 것이다. 재현율이 작을 때 태그를 적용한 검색 결과가 더 좋은 정확율을 보이고 있다.

재현율	일반 검색	태그 가중치 적용 검색
0.0	0.6574	0.7350
0.1	0.5284	0.6081
0.2	0.4514	0.5158
0.3	0.3601	0.3580
0.4	0.2645	0.2990
0.5	0.2339	0.2227
0.6	0.1735	0.1745
0.7	0.1583	0.1525
0.8	0.0862	0.1007
0.9	0.0561	0.0705
1.0	0.0219	0.0219

표 3: **재현율에 따른 평균 정확율의 변화 비교**

6. 결론 및 고찰

본 논문에서는 웹 문서에서 그 내부 구조를 형성하는 태그를 정보 검색에 이용하는 방법을 제시하였다. 이를 위해 태그에 대한 중요도 가중치를 검색 엔진에 적용하였으며 태그의 중요도에 적합한 가중치를 학습하기 위해 유전자 알고리즘을 이용한 가중치 학습 알고리즘을 제안하였다.

제안된 검색 방법은 실험을 통해 태그의 중요도 가중치를 학습하고 학습된 태그의 가중치 정보를 정보 검색에 사용하였으며 그 변화를 비교하였다. 제안된 학습 알고리즘은 문서 구조의 주요 태그에 대해 그 중요도에 따라 가중치를 학습했으며, 태그 가중치를 이용한 정보 검색에서 검색 결과 향상에 대한 가능성이 확인되었다. 특히 검색된 문서에 대한 관련 문서의 분포는 태그 가중치를 적용한 경우 일반 검색 결과에 비해 관련 문서들이 상위에 더 많이 존재하였다. 이는 학습된 태그 정보가 관련 문서를 검색 결과의 상위에 랭킹 시키는데 도움을 준다는 것을 의미한다. 실세계 사용자가 검색 엔진을 사용할 경우 체감 검색 효율은 관련된 문서가 얼마나 먼저 출력되느냐에 달려 있다는 것을 볼 때 태그의 중요도에 따른 가중치 정보의 활용은 중요하다고 할 수 있다.

향후 과제로는 다음과 같은 내용의 연구가 필요하다. HTML로 만들어진 실세계 웹 페이지는 태그를 이용한 구조화에 있어서 부족한 점이 있다. 즉, 일반적으로 문서가 준 구조화(semi-structured)되어 있기 때문에 나타나는 태그 가중치 검색의 효과 감소를 생각할 수 있다. 따라서 XML 문서 집합 등 구조화가 잘되어 있는 집합에 대한 추가적인 평가가 요구된다. 또한 제안된 방법은 질의의 경향과 문서의 속성을 학습하기 때문에 특정 사용자에 대해 개인화(personalized)된 검색 에이전트의 개발 또는 일정한 특성을 가지는 구조화된 문서 집합을 대상으로 한 정보 검색에 이용할 수 있으며 이에 대한 연구가 필요하다.

참고 문헌

[Archive] Internet Archive, *Building an Internet Library*, http://www.archive.org.

[Ballard, 1997] Ballard, D. H., *An Introduction to Natural Computation*, MIT Press, pp. 263-275, 1997.

[Bharat and Henzinger, 1998] Bharat, K. and Henzinger, M. R., Improved Algorithms for Topic Distillation in a Hyperlinked Environment, *Proceedings of the ACM SIGIR'98 Conference*, pp. 104-111, 1998.

[Boyan et al., 1996] Boyan, J., Freitag, D., and Joachims, T., A Machine Learning Architecture for Optimizing Web Search Engines, *Proceedings of the AAAI Workshop on Internet-Based Information Systems*, pp. 1-8, 1996.

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

[Broglio et al., 1995] Broglio, J., Callan, J. P., Croft, W. B., and Nachbar, D. W., Document Retrieval and Routing Using The INQUERY System, *The Third Text REtrieval Conference (TREC-3)*, pp. 29-38, 1995.

[Callan et al., 1992] 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.

[Chakrabarti et al., 1997] Chakrabarti, S., Dom, B., Gibson, D., Kumar, S. R., Raghavan, P., Rajagopalan, S., and Tomkins, A., Experiments in Topic Distillation, *ACM-SIGIR '98 Post-Conference Workshop on Hypertext Information Retrieval for the Web*, 1998.

[Croft and Harper, 1979] 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.

[Cutler et al., 1999] Cutler, M., Deng, H., Maniccam, S and Meng, W., A New Study on Using HTML Structures to Improve Retrieval, *The Eleventh IEEE Conference on Tools with AI*, pp. 406-409, 1999.

[Goldberg, 1989] Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, 1989.

[Gordon, 1988] Gordon, M., Probabilistic and Genetic Algorithms for Document Retrieval, *Communications of the ACM* 31, pp. 1208-1218, 1988.

[Holland, 1975] Holland, J. H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, 1975.

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

[Maron and Kuhns, 1960] Maron, M. E. and Kuhns, J. L., On Relevance, Probabilistic Indexing and Information Retrieval, *Association for Computing Machinery*, 7(3), pp. 216-244, 1960.

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

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

[NIST] NIST, *Text REtrieval Conference homepage*, http://trec.nist.gov.

[Picard, 1998] Picard, J., Modeling and Combining Evidence Provided by Document Relationships Using Probabilistic Argumentation Systems, *Proceedings of the ACM SIGIR'98 Conference*, pp. 182-189, 1998.

[Robertson and Sparck, 1976] Robertson, S. E. and Sparck Jones, K., Relevance Weighting of Search Terms, *Journal of the American Society for Information Science*, 27, pp. 129-146, 1976.

[Robertson et al., 1995] Robertson, S. E. et al., Okapi at TREC-3, *The Third Text REtrieval Conference (TREC-3)*, pp. 109-126, 1995.

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

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

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

[Shin and Zhang, 1998] 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)*, pp. 501-507, 1998.

[Spertus, 1997] Spertus, E., ParaSite: Mining Structual Information on the Web, *The Sixth International World Wide Web Conference (WWW6)*, pp. 1205-1215, 1997.

[Turtle and Croft, 1991] Turtle, H. and Croft, W. B., Evaluation of an Inference Network-based Retrieval Model, *ACM Transactions on Information Systems*, Vol. 9, No. 3, pp. 187-222, 1991.

[Voorhees and Harman, 1999] Voorhees, E. M. and Harman, D., Overview of the Eighth Text Retrieval Conference, *The Eighth Text REtrieval Conference (TREC-8)*, pp. 1-27, 1999.

[Weiss et al., 1996] Weiss, Ron., Vèlez, B., and Sheldon, M. A., HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering, *Proceedings of the Seventh ACM Conference on Hypertext*, pp. 180-193, 1996.

[Yang et al., 1993] Yang, J., Korfhage, R. R., and Rasmussen, E., Query Improvement in Information Retrieval using Genetic Algorithms: A Report on the Experiments of the TREC Project, *The First Text REtrieval Conference (TREC-1)*, pp. 31-58, 1993.

[Yang and Honavar, 1998] Yang, J. and Honavar, V., *Feature Extraction, Construction and Selection - A Data Mining Perspective*, Kluwer Academic Publishes, pp. 117-136, 1998.

[Zhang, 1995] Zhang, B.-T, Learning and Optimization by Artificial Evolution (in Korean), *The Institute of Control, Automation and Systems Engineers Magazine*, Vol. 1, No. 3, pp. 52-61, 1995.

[Zobel, 1998] Zobel, J., How Reliable are the Results of Large-Scale Information Retrieval Experiments?, *Proceedings of the ACM SIGIR'98 Conference*, pp. 307-314, 1998.

Abstract

This paper presents a method for web-document retrieval by learning importance factors for tags which are used for document structuring. Web documents are usually written in Hypertext Markup Language (HTML). HTML consists of tags which make a document into a specific form and a homepage is designed using the tags according to its object. In this paper, we propose a method for improving the retrieval performance using the information of HTML tags. The importance factors for the tags are learned using a genetic algorithm. A tag is mapped into a gene and a set of tags represented as a chromosome. The results obtained by genetic learning are the weights for tag importance, and provided the retrieval engine as the weights of documents.

Experiments have been performed on an artificial dataset and a large collection of TREC (Text REtrieval Conference) documents. Our empirical results show that this algorithm learns the weights by tag importance factors, and can improve the retrieval performance on top-ranked documents.

Keywords: Information Retrieval, Web-Document, Tag Weight, Genetic Algorithm

감사의 글

목적지가 어딘지 무엇인지 아직 모르지만 지나간 한 걸음 한 걸음이 모여 여기까지 오게 되었습니다. 지난 석사과정 생활 동안의 결실이라고 말하기에는 많이 부족하지만 작게 시도한 한가지 일에 대한 중간 결과물로 이 논문을 내어놓습니다. 감사의 글이 가장 쓰기 어려운 부분인 것 같습니다. 무엇인가 나의 흔적을 남긴 다는 것에 대한 두려움 때문일까요.

이 한 장의 작은 공간 안에 감사의 뜻을 모두 전할 수는 없겠지만 작은 정성을 담아 적습니다. 먼저 끌고 가는 것이 아니라 지켜보고 방향을 제시해 주신 장병탁 교수님께 진심으로 감사를 드립니다. 혼자서 해야 한다는 것이 섭섭하기도 했지만 공부에 대한 태도를 선생님을 통해 배우게 되었습니다. 그리고 가깝지만 여전히 멀었던 김영택 교수님, 믿음의 모습을 보여주신 박근수 교수님과 하순회 교수님께도 감사를 드립니다.

지난 2년 동안 같은 방에서 생활했던 영우형, 석준이형, 광주형, 동호, 재순이형, 컬트우먼 인영, 재롱?둥이 재홍, 태진이형, 현구, 승준, 희주... 연구실 동기인 사랑스런 형주와 유환, 박사장 상욱, 돌고래 호식... 그리고 NLP/AI 가족들...재원이형, 성동이형, 유섭이형, 제균이형, 양박사님, 정호형, 종우형, 성배형, 상윤이형, 동연, 규백이, 시은이 누나, 수용, 장민, 상준, 정집이, 정우형, 호진이형, 소영이누나, 선영이, 사랑스런 신입생들인 인희와 승우에게도 고마운 마음을 전합니다.

지금의 선이가 있을 수 있었던 이유이자 언제나 소중한 두 분, 어머니, 아버지. 스물 여덟을 셀 동안 제게 비춰졌던 많은 모습, 느낌들... 이루 말할 수 없는 눈물의 감사를 이 논문과 함께 부모님께 드립니다. 그리고 선이에게 없어서는 안 되는 나의 기댈 곳인 현지에게도 감사를 드립니다.

마지막으로 언제나 이끌어 주시고 보살펴 주시고 채찍질해 주시는 내 삶의 인도자이신 주님께 감사와 영광을 돌립니다.

Also, Thanks To:

My brother 우리 형 그리고 his wife... 임영환 교수님, 서창진 교수님... 믿음의 친구들이었던 지호, 형엽, 광호, 광혁, 근범... 희동, 종석... 학부 초년때 한참 잘 다녔던 성광, 휘재, 인이, 정화... 동기인 영표, 성구, 현수, 광현이형, 석춘이, 인숙이, 하영이, 민용이, 신... 친한 나의 친구 동선... 친근한? 누님인 윤희누나... 쑤동 친구들 신이형, 승은, 미영, 연주, 미경, 병기, 정현이, 혜정, 정호, 수연이 누나... 토론토 갈 때 많은 도움을 주었던 친구 선정이... 그곳에서 만난 한국인 친구들 종훈이, 성미, 연주, 윤정, 수영, 보영, 임경... 그리고 내게 깊은 인상을 준 도시 토론토... 수정동 가족들 재호형, 정우형, 미경, 경주, 송, 현선, 은진, 현수누나... 플러스 송이 남자친구 진형, 느지막하게 알게 된 효영... 같은 년도에 태어났다는 이유로 뭉쳤던 또 다른 믿음의 친구 은화, 정우, 보용, 성진이, 희경이... CCMGer 은영이누나, 경임누나, 복현이형, 재혁, 민정, 명수, 정희, 신영이, 윤정이 누나... 월드스피커 사장님, 박승옥씨... Stryper, Michael W. Smith, Impellitteri... dog라는 동물을 친근하게 만들어 준 Roxi... 그리고 나의 마스코트 Gromit... 그리고 지금 생각이 안나 여기 적지 못한 나의 기억 속에 여전히 남아있는 많은 분들...