유전자 탐색에 의한 웹문서 검색

서영우°, 장병탁
서울대학교 컴퓨터공학과

Web-Document Retrieval Using Genetic Search

Young-Woo Seo and Byoung-Tak Zhang
Dept. of Computer Engineering, Seoul National University

요 약

본 논문에서는 웹을 기반으로 한 인터넷에서 유전자 알고리즘을 이용한 정보검색 방법을 제시한다. 특정 문제에 대한 가설 공간을 탐색하여 최적의 해를 찾을 때 지역성과 전역성을 함께 고려하는 유전자 알고리즘의 특성을 웹에서의 정보검색에 이용한다. 여기에서 고려할 점은 탐색속도와 탐색방향인데 본 논문에서는 탐색속도를 고려하지 않았다. 탐색방향은 사용자의 정보 요구와 검색된 문서와의 유사도 평가함수로 조절하였다. 본 논문에서 제안한 유사도 평가함수로 실험을 한 결과, 사용자의 초기 정보요구에 대한 검색 결과의 적합성 여부에 대한 사용자의 평가가 기존의 검색엔진을 사용했을 때보다 개선된 결과를 얻을 수 있었다. 그리고 HTML 문서의 특성을 고려해서 검색하는 경우에는 검색어에 대해 보다 특정적인 결과를 제시했으며, 문서 내에서 검색어의 지역 중요도만을 고려하는 경우는 보다 일반적인 결과를 제시하는 것을 확인할 수 있었다.

1. 서 론

과거 인터넷은 학자, 학생들과 전문가들 사이에서만 정보교류를 위해 이용되었다. 그러나, 90년대 초반 이후 넷스케이프와 같은 인터넷 브라우저의 개발로 그 접근 범위가 일반인들에게 확대되었고, 정보통신기술의 발전, 기반 통신망의 확충, 관련 하드웨어 기술의 발전과 더불어 웹에 기반을 둔 인터넷의 사용이 하루가 다르게 증가하고 있는 추세이다[1].

대다수 사람들은 이러한 환경변화로 인하여 자신이 원하는 정보를 찾아 볼 수 있는 엄청난 크기의 정보창고를 가지게 되었다. 그러나, 실세계의 정보체계와 마찬가지로 정리가 되어 있지 않는 이 창고에서 자신에게 실제로 필요한 정보를 찾는 것이 상당히 어려운 문제로 새로이 대두되고 있다. 현재, 이러한 문제에 대한 여러 가지 대안들이 나와 있는데, 그 하나는 사용자의 정보요구를 표현하는 하나 이상의 검색어에 대해 이미 색인해 놓은 문서집합에서 연관성 있는 문서를 검색어와 문서들간의 유사도 값에 따라 검색하여 순서대로 보여주는 실시간 검색서비스인 검색엔진들이고, 또 다른 하나는 사용자의 동적, 정적 정보요구를 담은 프로파일을 기초로 하여 주기적으로 적절한 웹 페이지를 검색하여 제공하는 개인화된 캐스트 서비스이다. 전자의 경우는 웹에서 검색한 문서를 수집하여 미리 색인화 기법을 통해 정적인 문서집합을 구축하고 그 문서집합에서 사용자의 동적인 정보요구를 다룬다는 점에서 기존의 정보검색(information retrieval) 시스템의 변형으로 볼 수 있고, 후자의 경우는 색인된 문서집합 없이 동적인 웹을 그 정보집합으로 하여 정적인 사용자의 정보요구를 다루는 정보여과(information filtering) 시스템으로 볼 수 있다.

본 논문에서 제시하는 정보검색 방법은 정보여과 시스템과 유사하다. 따라서, 웹에서의 정보검색이란 동적인 정보집합에 대해 정적 혹은 동적인 사용자의 정보요구를 다루는 문제로 정의할 수 있다. 즉, 웹이라는 탐색공간에서 특정 사용자의 정보요구에 대한 최적의 해를 구하는 것을 목표로 한다. 여기에서 최적의 해는 사용자의 검색기준, 즉 검색된 정보를 선택할 것인지의 여부를 결정하는 기준에 따라 달라지므로, 본 연구의 검색방식을 구현한 시스템의 평가를 위해서 하나의 검색어에 대해 특정적인(more specific) 것을 검색했을 경우에 그것을 최적의 해에 가까운 해로 간주하였다

2. 유전자 알고리즘을 이용한 웹 정보검색

유전자 알고리즘을 이용한 정보 검색에서는 사용자의 질의문이나 프로파일을 구성하는 각각의 검색어를 유전자로 간주하여 유전 연산자를 적용하고, 웹 문서 공간을 탐색하면서 유전자, 즉 키워드와 가장 특정적으로 연관되는 문서를 검색한다. 따라서, 가설집합에서 최적의 해를 찾을 때 지역성과 전역성을 함께 고려하면서 탐색하는 유전자 알고리즘의 특성을 웹 정보검색에 이용할 수 있다.

각 키워드에 대한 초기 염색체는 기존의 검색엔진에 메타검색을 하여 얻는 n개의 유전자, 즉 단어로 구성된다. 초기 염색체를 구성하는 유전자는 메타-검색된 웹 문서로부터 추출하게 되는데, 그 기준은 단어의 지역 중요도로 결정한다. 단어의 중요도란 단어가 문서내에서 가지게 되는 상대적 중요도를 수치로 나타낸 것으로, 검색된 문서에서 HTML 태그를 제거하고 영어의 불용어-리스트(stop-list)를 제거한 후 계산한다.

$$w_{ij} = K\frac{freq_{ij}}{maxfreq_j} \qquad (1)$$

K . document size (the total number of words in document j)
$freq_{ij}$ the frequency of term i in document j
$maxfreq_j$ the maximum frequency term in document j

위의 과정을 거치면 검색된 각 문서들이 지역 중요도값을 갖는 단어들의 벡터로 변환된다.

다음 과정은 이 문서 벡터와 사용자의 정보요구를 표현한 단어로 구성된 사용자의 질의문 혹은 프로파일과의 유사도(similarity)를 측정하는 것이다. 각 유전자에 대한 다음 세대로의 진화 여부는 그 유전자의 적합도(fitness) 값에 의해 결정되는데, 여기에서 적합도 값은 그 유전자를 검색어로 해서 검색된 문서와 사용자의 질의 혹은 프로파일과의 유사도 값이다.

$$Fitness(gene_i) = relevance\ estimate\ function(D_i)$$

where,

D_i is a document retrieved by gene.

웹처럼 문서집합이 동적인 경우에는 문서집합이 정적인 일반적인 정보검색 시스템에서 사용되는 유사도 평가함수를 적용하기 어렵다. 따라서, 본 연구에서는 HTML 문서 구조를 최대한 이용하기 위해 새로운 유사도 평가방식을 고안했다

$$Match(D_i) = \sum_{j=1}^{q} w_{ij} \quad for\ i = 1, 2, .., n \qquad (2)$$

$$Match\text{-}Title(D_i) = \alpha(\sum_{j=1}^{q} w_{ij}) + \beta(\sum_{j=1}^{q} tw_{ij}) \qquad (2)\text{-}1$$

$$Match\text{+}Header(D_i) = \alpha(\sum_{j=1}^{q} w_{ij}) + \gamma(\sum_{j=1}^{q} hw_{ij}) \qquad (2)\text{-}2$$

$$Match\text{+}anchortext(D_i) = \alpha(\sum_{j=1}^{q} w_{ij}) + \delta(\sum_{j=1}^{q} aw_{ij}) \qquad (2)\text{-}3$$

$$WIRA(D_i) = \alpha(Match(D_i)) + \beta(Match\text{+}Title(D_i)) + \gamma(Match\text{+}Header(D_i)) + \delta(Match\text{+}anchortext(D_i)) \qquad (3)$$

식 (2)는 질의문의 검색어가 검색된 문서 내에서 가지는 중요도(가중치)를 합하여 문서의 유사도 값을 계산하는 것이고 (2)-1에서 (2)-3까지의 유사도 평가함수는 HTML 태그의 <TITLE>, <HEADER> 그리고 <anchor text> 내의 검색어의 중요도를 식 (2)에 더한 형태로, 검색어의 중요도를 HTML 문서구조 측면에서 측정한 것이다. 즉, <TITLE>, <HEADER>, <anchor text> 부분에 나오는 검색어가 <BODY> 부분에 나오는 검색어보다 중요하다고 전제하는 것이다. 식 (3)은 식 (2)의 모든 요소에 대한 검색어의 중요도를 고려한 것으로, 검색된 문서에서 질의문의 검색어가 가지는 중요도를 여러 가지 측면에서 고려한 것이다.

각 상수 α, β, γ, δ는 검색된 각 문서의 유사도 값에서 각 요소들이 차지하는 비율을 조절하기 위한 것이다 문서의 유사도 값에서 α의 비율이 커지면 검색어에 대해서 보다 일반적인(general) 내용이 검색되고, β, γ, δ의 비율이 커지면 보다 특정적인(specific) 내용이 검색된다.

다음 세대의 염색체는 임계치(threshold) 이상의 유사도 값을 부여받은 유전자 n/2개와 다른 검색엔진으로부터 동일한 키워드에 대해 생성된 염색체 중 유사도 값이 임계치 이상인 유전자를 변이(crossover)연산하여 얻게 되는 n/2개로 구성된다

위 과정들을 사용자가 처음에 정한 정지조건(stop criteria), 즉 세대(generation) 수만큼 반복한다. 정지조건이 되면 각 세대별로 유사도 값이 가장 좋은 웹 문서 하나를 선택하여 세대수만큼 사용자에게 제시한다[4]

3. 정보검색 시스템 WIRAs의 설계 및 구현

WIRAs(Web Information Retrieval Agent System)는 본 연구에서 제시한 웹 정보검색 방법을 구현한 시스템이다. 검색방식은 기존의 검색엔진에 질의를 하여 그 결과에 유전자 알고리즘을 적용하는 메타-검색(meta-search)이다. 즉, 사용자의 정보요구에 적합한 HTML 문서를 찾기 위한 웹 공간 탐색 알고리즘은 유전자 알고리즘이고, 공간 탐색방향은 유사도 평가함수로 조절한다. 검색을 마치는 시기는 사용자가 초기에 입력한 정지조건을 만족하는 때인데, 본 연구에서는 진화하는 세대수를 정지조건으로 하고 있다. 검색을 마치면 세대수만큼의 결과가 사용자에게 제시되고 제시된 결과에 대한 사용자의 반응을 학습하고 사용자의 프로파일을 수정한다. 이 후 검색은 수정된 사용자의 프로파일을 기준으로 계속된다.

WIRAs의 구조는 다음과 같이 크게 두 부분으로 나누어진다.

① 사용자 인터페이스 부분 : 사용자와 반응하는 부분으로 사용자의 정보요구를 입력받고 정보검색 에이전트로부터 검색된 결과를 사용자에게 보여주고 검색된 결과에 대한 사용자의 평가를 입력받는 부분이다

② 정보검색 에이전트부분 : 정보검색 에이전트는 염색체당 하나가 생성되어 검색을 수행한다. 각 에이전트는 유전자 알고리즘을 적용하여 검색횟수(세대수 × 메타-검색엔진의수 × 최종결과로 제시되는 문서의 수)와 검색 방향(유사도 평가함수)을 제어하게 된다.

②-1. 질의문 처리부분 : 사용자의 질의를 메타-검색할 검색엔진에 적절한 질의문으로 바꾸는 부분

②-2. 문서처리부분 . 검색된 각 웹 문서를 분석하는 부분으로 먼저 HTML 태그를 제거하고 영어의 불용어-리스트에 해당하는 단어를 제거한다. 그런 다음 그 문서에 나타난 단어들의 지역중요도를 계산하고 유사도 평가방법을 사용하여 사용자의 정보요구와 유사도를 측정한다.

②-3. 학습부분 : 검색결과에 대해서 사용자가 내린 적합성 평가를 이용하여 사용자의 프로파일을 수정하기 위해서 사용자의 평가를 학습하는 부분이다.

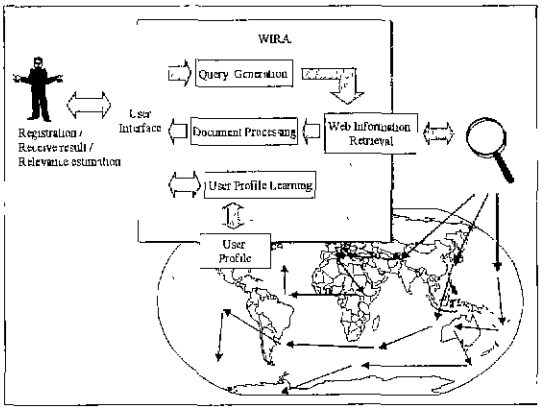

그림 1. WIRAs 구조

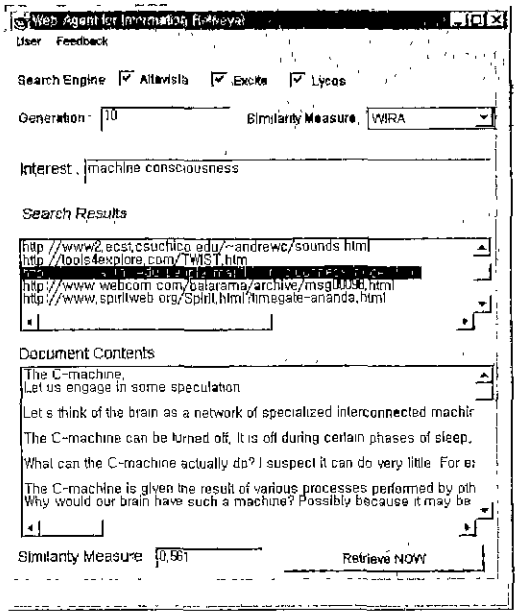

그림 2 WIRAs 주 화면

4. WWW 정보검색 실험 및 결과

실험은 6명의 사용자가 관심있는 분야에 대한 검색어를 기록한 프로파일을 기초로 웹을 검색하여 제시된 결과를 비교한다. 그리고, 사용자는 검색어를 등록할 때 자신의 검색기준을 정한다 예를 들면, British museum을 검색어로 제시한 사용자의 경우, 원하는 최종결과는 영국 런던에 있는 대영 박물관에 관련된 전반적인 내용이다.

사용자	검 색 어
가	spiking neuron
나	HONG KONG Ocean Park
다	vedanta, machine consciousness, ecological optics
라	Leonardo da Vinci, British Museum, semantic networks
마	NetBEUI
바	Information Retrieval Agent

표 1. 실험 데이터

최종결과는 유사도 평가함수 별로 세대수(정지 조건)만큼의 HTML문서를 지정하는 URL로 제시된다. 사용자는 제시된 문서에 대해서 자신의 정보요구와의 적합성 평가를 하게 된다. 이 적합성 평가의 대상은 최종 제시된 세대수만큼의 HTML문서로, 본 논문에서 제시한 유사도 평가함수 식 (2)와 (3)으로 검색된 문서와 메타-검색을 한 검색엔진이 검색한 문서이다.

적합성 평가는 100, 75, 50, 25, 0의 다섯 단계로 나뉘어 지는데, 적합성 평가 분포를 살펴보면 검색의 결과가 검색어에 대해 특징적인지 일반적인지를 알 수 있다[표 2]

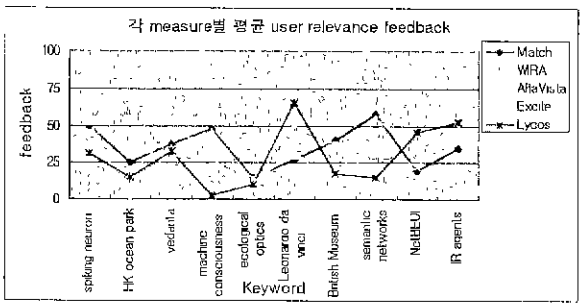

그림 3 각 measure 별 평균 user relevance feedback

사용자로부터 받은 적합성 평가의 평균값은 본 연구에서 제시한 유사도 평가함수가 약간 우수한 것으로 나와 있지만[그림 3]. 이것으로 전체 시스템의 검색효율을 판단할 수는 없다 그러나, 본 논문에서 제시한 바와 같이 개인화된 웹 정보검색 도구에서는 그 결과가 사용자의 정보요구에 얼마나 특정적인가 하는 것은 의미가 있다. 그러므로 각 검색어에 대한 적합성 평가의 분포를 살펴볼 필요가 있다.

결과에서 보듯이 본 연구에서 제시한 유사도 평가함수가 사용자의 적합성 판단에서 약간 우수한 경우의 high와 low의 분포를 주의해서 살펴보면, 검색어에 대해서 Match 함수는 보다 일반적 검색결과를, WIRA 함수는 보다 특정적인 검색결과를 제시하는 것을 알 수 있다

5. 결 론

본 연구에서 제시한 웹 정보검색은 탐색공간의 지역성과 전역성을 동시에 고려하면서 탐색을 하여 최적에 해를 찾아내는 유전자 알고리즘의 특성을 이용하였다. 웹에서 정보를 탐색하는 방향은 검색된 문서의 유사도 평가함수로 조절했다. 실험에서 알 수 있듯이 웹상의 정보 중 대부분을 차지하는 HTML문서를 보다 잘 분석하기 위해서는 HTML 문서의 특성을 유사도 평가시

feedback / 질의어	평균(%)/표준편차(%)/high(100~75)/low (25~0)				
	Match	WIRA	Altavista	Excite	Lycos
spiking neuron	50/33.43 5 / 3	40/32.16 2 / 6	**52.5**/32 91 5 / 2	32.5/43.22 5 / 4	31 5/24 15 1 / 9
HK Ocean park	25/7.90 0 / 10	**60**/41 16 5 / 3	0 / 0 0 / 10	17.5/26 48 1 / 8	15/18.5 1 / 8
vedanta	37.5/33.85 2 / 5	28.7/33.74 2 / 7	35/45.94 3 / 7	**45**/34.96 4 / 5	32 5/35.74 3 / 7
machine consciousness	**47.5**/29.93 2 / 2	27.5/36 22 2 / 7	46 5/0 0 / 0	37.5/33 43 1 / 5	2.5/7.905 0 / 10
ecological optics	**15**/26 874 1 / 8	12.5/31.73 1 / 9	13.5/31 62 2 / 8	14.5/24.15 1 / 9	10/24.15 1 / 9
Leonardo da Vinci	27.5/37.36 3 / 7	47.5/41.16 6 / 4	30/32 16 2 / 6	32.5/47.79 4 / 6	**65**/45.06 3 / 6
British Museum	**40**/26.87 2 / 2	37.5/27.00 2 / 2	15/26.87 1 / 8	10/24.15 1 / 9	17.5/20.58 0 / 8
semantic networks	**57.5**/33.43 5 / 3	27 5/32 16 2 / 6	55/32.91 5 / 2	37 5/43.22 5 / 3	15/24 15 1 / 9
NetBEUI	20/22.97 1 / 9	30/30.73 1 / 7	32 5/23 71 1 / 9	22.5/21.88 0 / 7	**45.3**/39.87 2 / 8
Information Retrieval Agent	35/24.15 1 / 5	**55**/30.73 5 / 3	45/12 90 10 / 0	50/33.33 3 / 3	52.5/35.35 5 / 3

표 2. 사용자 적합성 판단별 분포

반영하는 것이 중요하다.

본 논문에서는 HTML 문서 구성요소 중 <TITLE>, <HEADER>, <Anchor> 세 가지만을 고려하였지만 문서의 유사도에 영향을 끼치는 또 다른 요소가 있는지 계속 연구중이다.

감사의 글 : 본 연구는 대학기초 연구기술연구 지원사업 (C1-98-0068-00)에 의해 지원되었음

참고문헌

[1] Web growth summary.
http://www.mit.edu/people/mkgray/net/
[2] F Menczer and R. Belew, Adaptive information agents in distributed textual environments, In *Autonomous Agents'98*, 1998
[3] A.Falk and Ing-Marie Jonsson. PAWS. an agent for WWW-retrieval and filtering,In *PAAM '97*. pp.169-179.
[4] J.Yang and V.Honavar, Feature subset selection using a genetic algorithm, In *IEEE Intelligent System*, pp.44-49. 1998
[5] Y.Li. Beyond relevance ranking:hyperlink vector voting, In *RIAO'97. Computer-Assisted Information Searching on Internet*, CA, 1997, pp.227-235.
[6] B. Sheth, NEWT: A learning approach to personalized information filtering, M.S. Thesis, Massachusetts Institute of Technology. 1994.
[7] J.Yang and V.Honavar, Feature subset selection using a genetic algorithm, In *IEEE Intelligent System*, pp.44-49. 1998.
[8] Frakes, W., and Baeza-Yates, R , *Information Retrieval Data Structure and Algorithms*, Prentice hall, NJ. 1992.
[9] Salton, G., *Automatic Text Processing*, Addison Wesley Publishing Company, 1989.
[10] Maes, P., and Kozierok, R, Learning interface agents, In *Proc. AAAI'93 Conference*, 1993.