유전자 알고리즘을 이용한 WWW 정보 검색

서 영 우, 장 병 탁
서울대학교 컴퓨터공학과 인공지능 연구실

WWW Information Retrieval Using a Genetic Algorithm

Young-Woo Seo°, Byoung-Tak Zhang
Artificial Intelligence Lab (SCAI)
Dept. of Computer Engineering, Seoul National University
{ywseo, btzhang}@scai.snu.ac.kr

요 약

최근 웹 상에서 여러 가지 정보에 대한 접근이 용이하여 많은 사람들이 다양한 검색 시스템을 이용하여 원하는 정보를 얻고 있다. 그러나 웹의 크기가 점점 커지고 그에 따른 사용량 또한 증가함에 따라 원하는 시간 안에 원하는 수준의 정보를 얻기가 매우 어렵다. 본 논문에서는 유전자 알고리즘을 이용하여 사용자의 요구수준에 보다 가까운 정보를 검색하는 학습 방법에 대해 고찰한다. 검색 엔진의 초기 검색 결과로부터 만들어진 색인어들이 하나의 염색체로 구성한다. 염색체를 구성하고 있는 각 유전자는 사용자의 기호에 맞는 URL을 추천하기 위해 검색된 문서들과 연관성 값을 비교하여 유전 연산자에 의해 변형된다. 제시된 정보 검색 방식은 기존의 검색 엔진으로부터 반환되는 검색 결과로부터 사용자가 원하는 정보에 연관된 하나 이상의 색인어를 생성한 다음 재검색하여 연관성이 높은 소수의 정보만을 사용자에게 제공한다. 제안된 학습 방식과 기존 검색 엔진으로부터 검색된 결과를 초기의 사용자 정보 요구와의 연관성에 있어서 비교 분석하였다.

1. 서론

Know-how보다는 know-where의 중요성이 날로 높아 가는 현재, 실행 환경과 관련 없이 구현될 수 있는 HTML 문서를 기반으로 한 WWW(World Wide Web)은 가장 중요한 정보의 근원이 된다. 국내에는 약 40만개 이상이, 전 세계적으로는 수 천만개 이상의 웹 문서가 있는 것으로 추산된다. 물론, 지금 이 시간에도 수많은 웹서버가 설치되고, 그 수에 배가되는 웹 페이지들이 새로 작성, 수정, 삭제되고 있을 것이다.

이러한 다변적이고 동적인 웹 환경에서 사용자의 정보요구(information needs)에 정확하게 들어맞는 웹 페이지를 찾는다는 것은 거의 불가능하다. 이에 사용자가 일일이 문서를 찾아 가지 않고, 이미 존재하고 있는 자료들의 색인 역할을 하여 사용자의 정보요구를 보다 쉽고, 빨리 제공하는 검색엔진이 등장하게 되었고, 웹의 크기가 방대해져 갈수록 그 역할이 점점 더 중요해져가고 있다.

현재 국내외적으로 많은 검색엔진들이 존재하고 있다. 이 검색엔진들 대부분의 이론적 기반은 정보검색(Information Retrieval)에 두고 있다. 각 정보검색 시스템들은 사용자로부터 정보요구를 받아서 최종 결과를 사용자에게 보여주는 각 단계별의 차이에 따라 여러 가지 종류로 나뉘어 질 수 있는데, 그 기준점은 다음과 같다[3].

① 문서와 질의문의 표현 방법
② 사용자의 질의에 대한 각 문서들의 유사성 평가를 위한 매칭 기법
③ 질의에 대한 결과를 사용자의 정보 요구에 대한 순위부여 방법
④ 사용자가 검색결과에 대해 부여하는 연관성의 정도를 어떻게 획득할 것인가에 대한 방법

일반적으로 각 검색엔진들도 위의 네 가지 기준점을 근거로 구축된 정보검색시스템이라는 점과 사용자의 질의를 검색엔진이 자체적으로 가지고 있는 데이터베이스의 인덱스와 비교하여 검색결과를 반환한다는 점에서 도서관이나 기업 등의 문자기반의 정보검색시스템과 유사하다. 전체 문서집합이 유동적이고, 분산되어 있고, 보다 크다는 것과 전체 문서집합과 그 인덱스의 생성자가 사람이 아닌 web robot(혹은 spiders, crawler, [3]) 등으로 불려지는 소프트웨어에 의해 행해진다는 점이 기존의 정보검색 시스템과 다른 점이다. 그러나, 웹이 기하급수적으로 변화하는 속도를 볼 때[4], 궁극적으로 웹 전체를 색인화하여 엄청나게 큰 색인을 만든다는 것은 결국 검색효율을 떨어뜨리는 결과를 초래하게 될 것이다.

본 논문에서는 하나 이상의 기존 검색엔진을 이용하여 사용자의 정보요구에 가장 적합한 웹 정보를 검색하는 방법과 사용자의 프로파일을 기초로 하여 사용자의 정보요구에 적합한 정보를 정해진 시간에 따라 제안하는 방법을 제시한다.

2. 유전자 알고리즘을 이용한 정보 검색

유전자 알고리즘을 이용한 정보 검색의 핵심은 사용자의 관심 분야에 대해 그 분야를 적절히 표현할 수 있는 하나 이상의 키워드로 구성된 단어의 집합을 만드는데 있다. 여기에서 단어의 집합이 염색체가 되고, 각 집합의 구성원이 유전자가 된다.

각 유전자들은 한번 이상의 사용자 연관성 판단이나 유사성 검사(돌연변이)를 거치면서 자신의 적합도 값에 따라 다음 세대로 유전이 되는지의 여부를 판정받게 되고, 해당 분야에서 적절하다고 판단되는 유전자는 같은 분야에 관심도를 표현한 사용자에게 자료를 추천할 때 사용되기도 한다.

사용자는 맨 처음 자신이 관심 있는 분야를 나타내는 하나 이상의 키워드로 정의된 프로파일을 제공하여 지속적으로 자신이 관심 있는 분야의 정보를 자신의 e-mail로 받아보거나(off-line 검색), 직접 자신이 찾고자 하는 정보를 나타내는 하나 이상의 keyword를 입력하여 검색을 한다(on-line 검색).
위의 두 경우 모두 사용자가 정의한 각 분야에 대해서 기존의 검색 엔진에 메타 검색을 하여 각 검색엔진으로부터 반환된 문서집합을 가져온다. 가져온 결과에 대해서 그 분야를 표현할 수 있는 염색체가 생성되는데 그 과정은 다음과 같다.

초기 검색에서 반환된 문서집합 중 상위에 위치한 문서 몇 개를 분석하여 가장 빈도 높은 단어를 수집한다. 식 (1)을 이용하여 지역 중요도를 계산하여 각 단어에 가중치(weight)를 부여하고 가중치가 높게 부여된 단어를 선택한다. 이때, 전체 문장 구성상 자주 쓰이는 전치사, 부사, 관사 등은 stop-list 사용하여, 제거하고 문서의 크기가 보다 큰문서의 중요하지 않은 단어가 선택될 가능성을 배제하기 위해서 문서의 크기로 정규화를 한다.

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

K : 문서 j 의 크기 (j 문서내의 단어의 총수)
$freq_{ij}$: 문서 j 내에서 단어 i 의 출현빈도수
$maxfreq_j$: 문서 j 내에서 출현 빈도수가 가장 높은 단어

선택된 각 단어는 하나의 유전자가 되고, 그 유전자들이 모여서 하나의 분야에 대한 염색체(chromosome, gene pool)를 형성하게 된다. 각 유전자에 대해 메타 검색을 하여(혹은 각 gene을 and 연산자로 질의를 하여) 그 중 상위에 놓여진 문서들(약 0.1%)과 맨 처음 사용자가 명시한 관심분야와 유사도를 계산하여 유사도 함수값이 높은 document들만을 사용자에게 제시한다.

사용자의 질의 Q에 대한 j 번째 문서 D_j의 유사도 (relevance)는 다음 식으로 평가한다.

$$R(D_j) = \sqrt[4]{\frac{\sum_{k \in D_j} match(k,Q)}{|K|}} \qquad (2)$$

$$match(k,Q) : \begin{cases} \text{만약 } k \in Q \text{ 이면 } 1 \\ \text{그렇지않으면 } 0 \end{cases}$$

$|K|$: 문서 j 의 전체 *keyword* 수

식 (2)의 유사도 함수값에 따라 순위를 부여하여 사용자에게 제시한 문서들에 대한 사용자의 relevance feedback을 기준으로 하여 query modification, 즉 mutation을 한다. 초기 검색된 문서들 중 순위가 높은 문서 1%를 각 검색엔진별로 선택하여, 이 문서들을 구성하는 단어들 중 빈도수가 높은 단어를 선택함으로써 염색체의 다음 세대가 구성된다.

즉 특정 분야는 그 분야에 연관되면서 출현 빈도수가 높은 단어(유전자)로 구성된 염색체로 표현된다. 여러 세대가 지난 후에 각 염색체는 다른 사용자가 같은 분야에 관심이 있는 경우 사용자에게 제시하게 되는 근거가 되고, 여기서 만들어진 각 domain에 대한 정보는 전체적으로 관리된다.

예) 사용자 프로파일 1
관심분야 : agent
keyword : agent, multi-agent, agent ecology
URL : http://www.crystaliz.com/logicware/mubot.html

사용자 프로파일 2
관심분야 : cognitive science
keyword : memory, learning, reasoning
URL : http://www.aimnet.com/~wattsl/

전체 domain 파일(알파벳순으로 정렬)
keyword list : agent(agent), agent ecology(agent), learning(cognitive science), memory(cognitive science),multi-agent(agent),reasoning(cognitive science)

위의 과정을 거치면서, 오프 라인으로 등록된 사용자수 만큼의 domain이 정해지게 되는데 이를 전체적으로 관리하여 이를 근거로 하여 cross over를 한다[5]. 새로운 사용자가 관심을 표명한 어떤 분야가 전체 domain내의 keyword에 있다면 그 keyword가 속한 분야의 URL을 추천하게 된다.

3. 설계 및 구현

메타 검색은 존재하는 하나 이상의 검색 엔진에 질의를 하여 그 결과를 결합하여 사용자에게 보여주는 검색 방법이다. 메타 검색은 일반적으로 질의문처리 부분, 제어 부분, 표시부분으로 나뉘어 질 수 있다.

① 질의문 처리 부분 : 사용자의 초기 질의문을 분석하고 각 검색엔진에 적절한 질의문을 생성한다.
② 제어 부분 : 질의문 처리 부분에서 생성된 질의문으로 각 검색엔진에 질의하고 반환된 결과를 분석하여 사용자의 초기 질의와 결과간의 유사성 검사를 하여 사용자에게 보여줄 웹 문서를 선택한다.
③ 표시 부분 : 제어부분으로부터 넘겨받은 하나 이상의 웹 문서를 적절한 형태로 표시한다.

본 논문에서 메타 검색의 대상으로 한 기존 검색엔진은 Alta-vista, Excite, Lycos이다. 사용자로부터 검색 제의를 받으면 각 검색엔진에 해당하는 에이전트가 사용자의 질의문을 해당 검색엔진에 맞게 변경하여 질의를 한다. 그 질의문에 대해 반환된 결과를 적절한 형태로 가공하여 제어 부분으로 넘겨주게 된다. 제어 부분은 넘겨받은 문서집합을 대상으로 앞에서 제시한 방법으로 각 분야를 적절히 표현할 수 있는 염색체를 만든다. 이 염색체를 이용하여 재검색한 문서집합에 대해 유사도를 평가하

고 그 유사도값에 따라 순위를 부여하여 사용자에게 제시한다.

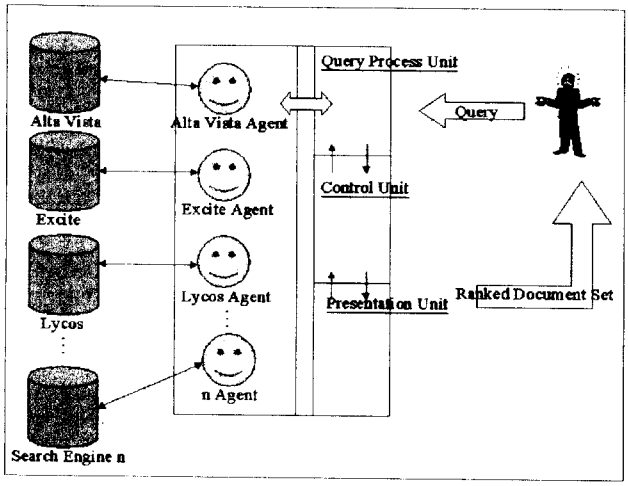

그림 1 메타 검색 엔진의 구성도

사용자가 관심이 있다고 제시한 각 분야에 대해 하나의 염색체가 존재한다. 각 염색체는 하나 이상의 유전자로 구성되는데, 이 유전자는 각 분야를 표현할 수 있는 단어가 된다. 각 염색체를 구성하는 유전자인 단어를 선택하려면 대상 집합이 있어야 한다. 초기 검색으로부터 반환된 문서들 중에서 대상 집합을 만들게 되는데, 3개의 검색 엔진으로부터 반환된 결과중 상위에 위치한 각 10개씩의 문서를 선택하여 초기 세대를 구성한다.

그 초기 세대로부터 식 (1)에 의해 유전자를 선택하게 된다.각 염색체를 구성하는 유전자는 사용자로부터 받은 연관성 판단과 유사성 검사로 적합도값을 받아서 다음세대에 그 유전자의 생존 여부를 평가하게 되는데, 적합도값이 임계치이상이 되어야 계속 생존하게 된다.

$$Fit_{(t+1)ij} = Fit_{(t)ij} + w_{ij} \qquad (4)$$

Fit_{ij} : 시간 t에 유전자 i의 적합도 함수 값
w_{ij} : j번째 염색체의 i 번째 유전자의 지역 중요도

사용자의 유사도 평가가 없는 경우에는 식 (4)에 의해 유전자의 생존 여부가 결정된다. 적합도 함수값이 임계치 이상으로 좋은 유전자는 다음의 질의와 비슷한 기호를 가진 사용자에게 추천할 때 사용되며(다음 세대로 상속되며) 적합도 함수값이 임계치 이하인 유전자는 사라지게 된다.

4. 실험 및 결과

사용자는 자신의 기본 정보를 등록하여 지속적인 검색을 할 수도 있고(그림 2), 단순한 메타 검색을 할 수도 있다.

표 1은 5명의 사용자의 프로파일의 일부로서, 하나 이상의 관심분야와 그 관심분야를 표현하는 염색체로 구성되어 있다. 초기 상태에는 각 분야와 그 분야를 표현하는 염색체의 단어 수가 일치하게 된다.

실험방법은 사용자가 검색 엔진을 사용하여 질의했을 때 검색된 문서와 본 논문에서 제시한 방법을 사용하여 검색된 문서들 중 상위순위의 10개에 대한 최초 사용자의 정보 요구로 제

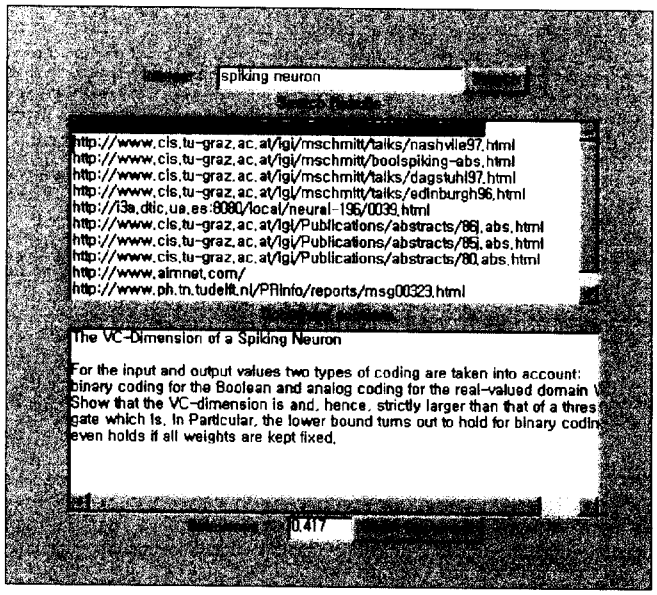

그림 2 메타 검색화면

시된 topic과의 유사도를 비교한다.

사 용 자	주요 관심분야	생성된 초기 염색체
가	world cup	: world cup
나	neural networks, machine learning	: neural networks : machine learning
다	baseball, CMAS Scuba	: baseball : CMAS Scuba
라	Spiking Neuron, DNA Computing	: Spiking Neuron : DNA Computing
마	Genetic Algorithm	: genetic algorithm

표 1. 실험에 사용된 사용자 프로파일

표 2는 (라) 사용자의 관심분야인 spiking neuron에 대하여 각 검색 엔진에 검색을 하여 반환된 문서들 중 상위에 위치한 5개씩의 문서를 선택하여 본 논문에서 제시한 방법으로 연관성을 검사한 결과이다. 그림 4는 표 1의 5명의 사용자의 여덟 개의 각 분야에 대해 연관성의 결과를 그래프로 도시한 것이다. 사용자의 정보 만족도가 연관성 값이 0.5이상일 때, 전반적인 결과가 좋지 않은 것은 의미론적 평가가 아니라 관심분야에 대한 단어가 검색된 문서에서 얼마나 많이 출현했는가를 평가했기 때문이다.

각 분야에 대해 하나의 염색체가 존재하는데, 맨 처음 염색체의 유전자는 그 분야를 나타내는 단어와 일치하게 되고, 다음 세대의 염색체는 이 문서들 중에서 가장 빈도가 높은 단어들 중 일부를 선택하여 구성하게 된다.

그림 4는 이 염색체들로 검색된 문서들의 연관성에 대한 것이다. 가 사용자의 관심분야인 world cup을 제외하고 나머지 사용자들의 관심분야에 대해서는 연관성이 높게 나타났다.

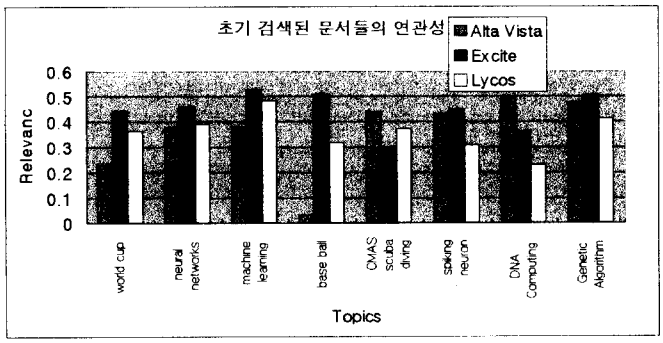

그림 3. 여덟개의 항목에 대한 연관성

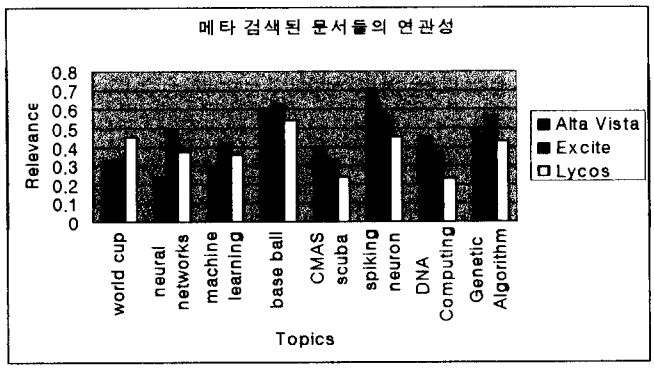

그림 4. 메타 검색된 문서들의 연관성

5. 결론 및 향후 연구 과제

본 논문에서는 하나 이상의 관심분야에 대하여 기존의 검색 엔진과 메타 검색을 이용한 검색의 연관성을 비교하였다. 기존의 검색 엔진의 결과들을 여과하여 검색한 메타 검색이 사용자의 초기의 정보요구에 대해 연관성이 크게 나왔다. 그러나 모든 분야에 대해 항상 위와 같은 결과가 나오는 것은 아니다. 연관성 검사 함수를 사용자의 정보 요구에 대한 단어가 검색된 문서에서 얼마나 많이 반복되었는지에 초점을 맞추었기 때문에 초기의 정보 요구와는 전혀 다른 분야의 문서가 검색되는 경우가 있었다. 정적이고 한정된 문서집합을 대상으로 하는 기존의 정보검색과는 달리 본 논문에서 제시한 검색 방법은 동적이고 그 범위가 제한이 없는 인터넷을 대상으로 하고 있기 때문에 기존의 연관성 검사 방법으로는 정확한 평가를 내리기가 곤란한 경우가 있다. 앞으로 네트웍 자체의 문제로 인해 문서 자체가 검색되지 않는 경우에 대한 고려 사항과 사용자의 연관성 평가를 보다 잘 활용하는 방법에 대한 연구가 필요하다.

감사의 글 : 본 연구는 일주학술문화재단에 의하여 일부 지원되었음.

6. 참고문헌

[1] A. Falk and I. Jonsson, PAWS: An agent for WWW-retrieval and filtering, In *PAAM'96*, pp.169-179, 1996.

[2] F. Menczer, ARACHNID: Adaptive retrieval agents choosing heuristic neighborhoods for information discovery, In *ICML'97*, pp. 227-235, 1997.

[3] M. Gray, Web growth summary, http://www.mit.edu/people/mkgray/net/web-growth-summary.html.

[4] K. Nygren, I. Jonsson and O. Carlvik, An agent system for media on demand services, In *PAAM'96*, pp. 437-454, 1996.

[5] V. Gudivada, V.Raghavan, W. Grosky, and R. Kasanagottu, Information retrieval on the world wide web, In *IEEE Internet Computing*, Sep-Oct. 1997, pp. 58-68.

검색된 문서	검색엔진	유사도(%)
www.bell-labs.com/new/gallery/neuron.html	Alta vista	37
portal.reserch.bell-labs.com/leisure/souvenirs/gallery/neuron.html		37
www.dag.uni-sb.de/DATA/Reports/9702/node30.html		41
www.cis.tu-graz.ac.at/igi/mschmitt/spikingneuron-abs.html		47
www.cis.tu-graz.ac.at/igi/mschmitt/talks/nashville97.html		47
www.dag.uni-sb.de/DATA/Reports/9702/node30.html	Excite	41
www.cis.tu-graz.ac.at/igi/mschmitt/talks/nashville97.htm		41
www.cis.tu-graz.ac.at/igi/mschmitt/boolspiking-abs.html		52
www.cis.tu-graz.ac.at/igi/mschmitt/talks/dagstuhl97.html		42
www.cis.tu-graz.ac.at/igi/mschmitt/talks/edinburgh96.html		46
www.aimnet.com/~wattsl/	Lycos	27
www.ph.tn.tudelft.nl/PRInfo/reports/msg00323.html		43
biogfx.bgsm.wfu.edu/literature/chapter/chapter.html		30
itb.biologie.hu-berlin.de/herz-group.html		fail
pooh.physik.uni-bremen.de/download.html		fail