바이오 인포메틱스를 이용한 웹 페이지 분석 기법에 관한 연구*

윤효근*, 이상용**
공주대학교 컴퓨터 공학과*, 공주대학교 정보통신공학부**
E-mail : {kosher*, sylee**}@kongju.ac.kr

A Study on Web Pages Analysis Technique based on Bioinformatics

Hyo-gun Yun*, Sang-yong Lee**
Dept. of Computer Science & Eng, Kongju National University*,
Division of Information & Communication Engineering**

요 약

대부분의 정보검색 과정들은 웹 페이지의 분석에 따라 검색 로봇을 이용한 검색기법, 카테고리를 이용한 색인 DB를 검색기법, 메타 태그를 이용한 검색기법을 사용하고 있다. 그러나 이러한 기법을 통하여 원하는 정보를 얻을 경우 정확도가 떨어지는 정보가 검색되어 사용자는 다시 한번 검색된 목록들을 확인해야 하는 경우가 발생한다.

본 논문은 다양한 형태의 웹 페이지에 대하여 바이오 인포메틱스 기술을 적용하여 분석, 사용자에게 필요로 하는 정보를 보다 정확하게 제공하는 기법을 제안한다.

1. 서 론

정보의 홍수 속에서 살고 있는 현대인들에게 정보 검색은 단순히 많은 정보를 보여주는 것만이 아니라 검색된 정보가 사용자들에게 얼마나 적합한 정보를 제공하는가가 가장 중요하다[1].

본 논문은 생물학에서의 유전자 분자 구조를 분석하고 분자 정보를 가지고 데이터 베이스를 구축하는 기법 인 바이오 인포메틱스(BioInformatics)를 이용하여 사용자에게 보다 정확한 정보를 제공하는 기법을 제안한다[2][6][9].

2. 관련 연구

2.1 검색 기법

웹에서의 검색기법들은 크게 세 가지로 구분된다. 단어기반 검색기법(Word based Search)과 카테고리 기반 검색기법(Category based Search), 메타 태그 검색기법(Meta-tag based Search)이 있다.

단어기반 검색은 사용자가 찾고자 하는 사항에 가장 근접한 단어를 검색엔진에 넘겨주어 웹 로봇이 검색하여 자동화된 색인기법을 사용함으로 보유 자료가 많다는 장점을 가지고 있다. 대표적인 기법은 TF-IDF와 같은 통계적 기법을 사용한다. TF(Term Frequency)는 전체 페이지에 대한 단어의 출연 빈도수를 나타내고, IDF (Inverse Document Frequency)는 전체 결과 중에 가장 적게 나타난 역문헌 빈도수를 나타낸다[4][10]. 각각 TF와 IDF의 값을 서로 곱하여 나온 가중치를 가지고 정렬처리하여 검색자료를 사용자에게 제공한다. 그러나 단어기반 검색 기법은 초보자들이 쉽게 자료를 검색하지 못한다는 단점을 가지고 있다.

카테고리 기반 검색은 유사 성격의 사이트끼리 일정한 카테고리에 두고 사용자는 이런 카테고리를 찾아 들어감으로써 목적하는 자료 또는 사이트를 찾게 되는 기법이다. 따라서 매우 엄선된 자료만을 색인하므로 질적인 면에서 우수하다[5][10]. 사람이 직접 웹 페이지를 유형별로 분류하여 사용자에게 제공하므로 시간과 노력이 많이 소모되는 단점을 가지고 있고 현재 이러한 단점을 없애기 위해 많은 연구가 진행중이다.

메타 태그 기반 검색은 자신의 고유DB는 보유하지 않으며, 다른 검색엔진의 DB를 이용하여 사용자의 질의에 대해 제휴하고 있는 각 검색엔진의 추출결과를 볼 수 있는 검색 기법으로서 에이전트를 활용한다. 다양한 검색엔진의 자료를 한자리에서 확인할 수 있다는 장점을 가지고 있지만 각 검색엔진의 검색기법의 차이로 인해 정밀 검색은 불가능하다[10].

* BK21 대전·충남 정보통신 인력 양성 사업단의 RA 수혜를 받았음.

2.2 유전자 알고리즘

유전자 알고리즘(GA)은 생태계에서 생물이 자신의 유전정보를 다음 세대로 전해주고 세대간의 진화를 통해 생태환경에 대한 적응능력에 대한 메커니즘을 공학적으로 모델화한 기법으로 초기 모집단에서 다음 세대를 구성하기 위해 적합도 평가를 거쳐 유전자를 선택하고, 교배, 돌연변이 과정을 종결조건에 만족할 때까지 반복하게 된다[3].

이러한 유전자 알고리즘은 전역적 탐색 기법중의 하나이면서 자연 도태와 진화의 원리에 기반을 둔 확률적인 탐색 알고리즘으로도 표현된다. 특히 탐색 및 최적화, 기계학습의 도구로 많이 이용되고 있다.

2.3 바이오 인포메틱스

바이오 인포메틱스는 유전자 알고리즘에서 분파되어 생물의 분자 구조인 mRNA를 통계적 이론, 전산기술 등을 이용하여 생물의 분자 정보들을 저장, 분석 및 해석에 대해 주목적으로 하는 탐색 기법 중의 하나이다. 현재 바이오 인포메틱스는 데이터 마이닝, 네트워크, 기계학습등의 도구로 많이 이용되고 있다[2].

바이오 인포메틱스의 기본적인 전제는 염기 서열이 구조(structure)를 결정하고, 결정된 구조에 의해서 그 기능(function)을 수행, 반복하여 염기 서열의 유사성을 분석하여 그 구조와 기능을 유추하는데 있다[6]. 유전 배열(alignment)은 두 염기 서열들을 상호 전개함으로서, 배열을 통해 두 서열의 유사성의 정량적인 정도를 알 수 있다. 적절한 배열(optimal alignment)는 두 서열이 가장 높은 유사성을 가질 수 있게 배열하는 것을 의미한다[7].

3. 페이지 분석 기법

본 연구에서는 단어 기반 검색 기법을 대상으로 바이오 인포메틱스 기법을 사용하였다. 그리고 사용자 프로파일과 PCR(Polymerase Chain Reaction)를 이용하여 제공되는 정보들 중 사용자의 관심을 가질만한 키워드를 생성하고, 사용자에게 적합한 웹 페이지를 제공하는 기법을 제안한다.

3.1 사용자 프로파일

사용자 프로파일은 사용자의 관심과 기호를 표현한 것으로 <그림 1>과 같이 하나 이상의 키워드와 관심도로 구성되며 이를 위해 사용자가 입력한 키워드를 벡터 방식으로 변환하여 [키워드, 관심도]의 집합으로 구성된 초기 사용자 프로파일을 생성한다.

사용자의 관심도는 0에서 1사이의 값으로 표시한다. 초기 사용자의 키워드는 1로 하고 새로운 키워드에 대해서 통계학상의 0.95의 오차를 두고 웹 페이지를 분석하도록 한다.

키 워 드	k_1 k_2 k_3 . . . . k_n
관 심 도	I_1 I_2 I_3 . . . . I_n

그림 1. 사용자 프로파일

3.2 웹 페이지 분석기법

페이지 분석기법은 사용자에게 제공되는 컨텐츠를 바탕으로 웹 페이지의 각 단어와 태그, 기호 및 공백(space)을 각기 하나의 유전자로 보고 불필요한 단어를 제거, 분석한다.

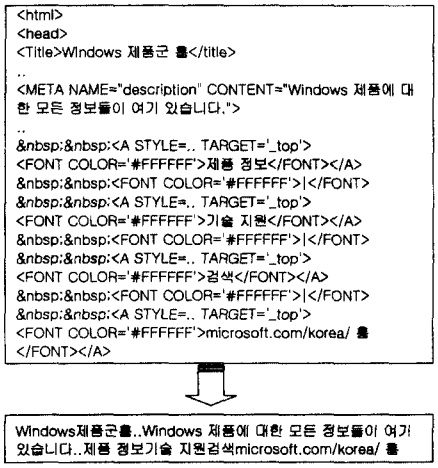

그림 2. 웹 페이지 염기 서열화

사용자 프로파일에 등록된 관심 키워드를 초기 염기 서열(키워드)로 하고 <그림 2>와 같이 분석된 웹 페이지의 염기 서열을 비교하여 동일 염기의 출연 빈도를 0에서 100사이의 값으로 표시하고, 출연 빈도가 높은 염기를 찾아내어 새로운 염기로 확장(PCR), 표시하고 새로운 염기로부터 다시 다음 염기 서열과 비교, 평가한다[9].

사용자의 관심 키워드에 따른 웹 페이지 분석 및 추출 유사성(S)은 다음과 같다[8].

$$f(S)=\begin{cases}1, & \text{if } \sum_{x\in k_i}\sum_{i=1}^{n} w_i^x I(t_i,x) > \theta \\ 0, & \text{otherwise}\end{cases}$$

t_i : 웹 페이지의 단어 집합 T에 대한 i 번째 원소

n : 웹 페이지의 단어 집합 T의 총 개수

w_i^x : 키워드에 따른 단어의 유사성

θ : 동일 단어의 출연 빈도

웹 페이지의 정확도는 사용자의 관심 키워드와 동일한 단어 출연 빈도수 θ값을 조정함으로써 사용자가 원하는

웹 페이지로 평가하고 표현 값 1로 변환한다. 1의 값이 많으면 해당 웹 페이지는 사용자에게 가장 적합한 웹 페이지가 된다. 해당 단어에 대해 사용자의 프로파일에 추가하고, 또한 θ값을 만족하지 못하는 웹 페이지는 사용자에 적합하지 않은 문서로 표현되어 값을 0으로 변환하여 사용자의 관심도에서 표현하지 않는 문서로 분류하여 제공하지 않는다.

3.3 PCR

PCR은 DNA의 영역을 수 시간 동안 수십만 배로 증폭하는 방법으로 통상 유전자는 이중 가닥으로 존재하지만 RNA는 단일 가닥으로 구성되었다.

본 연구에서는 검색된 웹 페이지를 하나의 긴 스트링 염기 서열로 전환하고, 다음에 검색된 웹 페이지와 합성하기 위하여 키워드를 중심으로 한 목적 영역을 설정하고, 이것을 초기 염기서열과 함께 이중 가닥으로 표현, 서열 비교를 한다. 이중 가닥의 염기 서열은 목적 영역과 합한 새로운 단일 스트링 염기 서열로 표현하고 유사도를 평가한다.

표현된 단일 스트링의 염기서열과 새로운 웹 페이지의 염기 서열을 다시 이중 가닥으로 표현, 키워드를 찾아 새로운 목적영역을 설정하고, 통합된 염기서열을 생성, 평가한다. 이와 같은 기법을 계속적으로 반복하여 사용자에게 적합한 키워드와 웹 페이지를 제공하도록 한다.

4. 실험 및 평가

<표 1>은 수집된 신문사의 일부 웹 페이지 중에서 키워드 출연 빈도수를 3으로 하고 분석했을 때 신규 키워드의 출연과 공통 키워드의 출연 빈도수를 통하여 웹 페이지를 비교한 결과이다.

표 1. 각 웹 페이지의 염기 서열 비교

k_i : 추출 키워드, Wi : Web Page

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	View Page
W_1	3	0	3	0	5	0	3	0	0	4	1
W_2	3	5	4	3	2	2	4	0	4	1	1
W_3	0	1	0	1	2	1	0	1	0	0	0
W_4	3	0	5	0	0	0	3	0	0	0	1
추가 키워드	1	0	1	0	0	0	1	0	0	0	

예를 들어 사용자가 "정치"라고 하는 단어를 검색을 했을 때, 각 웹 페이지 중 "정치"라는 단어가 일정 출연 빈도수를 넘는 웹 페이지들을 검색하고, 검색된 페이지들 중의 출연 빈도수를 넘는 염기 서열을 다른 웹 페이지들의 염기 서열과 비교, 공통 염기가 95%의 오차를 갖는 웹 페이지들을 색인하여 사용자에게 맞는 웹 페이지들을 찾아 제공한다.

5. 결론 및 향후 연구 과제

제안한 알고리즘은 사용자의 관심 주제어를 충분히 반영할 수 있고, 사용자가 제시한 내용에 따라 새로운 키워드를 추출하는데 있어 사용자 프로파일 각 계층의 변화를 파악하여 사용자에게 맞는 프로파일을 제시할 수 있는 장점을 가지고 있다. 또한 사용자가 새로운 주제어와 키워드를 삽입하여 프로파일 변화에 영향을 주어 사용자가 학습을 시킬 수도 있다.

단, 웹 페이지마다 키워드의 출연 빈도수를 달리하므로 계산 속도가 느리다는 단점이 있다. 향후 이러한 단점을 보완한다면 사용자들에게 보다 빠르고 정확한 웹 서비스를 할 수 있을 것이다.

참고문헌

[1] Greg, R., "Searching the Hidden Internet", Database, Vol. 20, No. 2, 1997.

[2] Kwonmoo Lee, J. H. Kim, T. S. Chung, B.S. Moon, H.S. Lee, Isaac S. Kohane, "Evolution Strategy Applied to Global Optimization of Clusters in Gene Expression Data of DNA Microarrays", CEC 2001, pp.845-850, 2001

[3] Reginald L. Walker, "Parallel Clustering system Using the Methodologies of Evolutionary Computations" , CEC 2001, pp.831-838, 2001

[4] Salton, G., and Buckley, C. , "Term-Weighting approaches in automatic text retrieval", Infromation Processing & Management, No 5, Vol 24, pp 513-523.

[5] Makoto Iwayama, Takenobu Tokunaga, " Cluster-Based Text Categorization: A Comparison of Category Search Strategies.", SIGIR , pp.273-280, 1995

[6] P. Dayan, G. E. Hinton, R. neal, and R. S. Zemel, " The Helmholtz Machine", Neural Computation, 7:1022-1037, 1995

[7] SHIN ANDO, HITOSHI IBA, "Inference of Gene Regulatory Model by Genetic Algorithms", CEC 2001, pp.712-719, 2001

[8] David Corne, Andrew Meade, richard Sibly, "Evolving Core Promoter Signal Motifs", CEC 2001, pp.1162-1169, 2001

[9] Altschul, S.F, Gish,W,Miller, W, Myers, E.W. & Lipmna, D.J , "Basic local alignment search tool.", J.Mol.Biol. No 3, Vol 215, pp.403-410, 1990

[10] http://apmlab.snu.ac.kr/ryuds/