유전자 알고리즘을 사용한 메타검색엔진에서의 사용자 모델링

김 수영, 조 성배
연세대학교 컴퓨터과학과

User Modeling in Meta-Search Engine with Genetic Algorithm

Soo-Young Kim and Sung-Bae Cho
Computer Science Department, Yonsei University

요 약

정보의 바다라고 일컬어지는 인터넷에서 원하는 정보를 찾기란 쉽지 않다. 웹 검색엔진이 나날이 발전하고 있기는 하지만, 이들 검색엔진은 개인에게 특화된 것이 아니기 때문에 검색 결과의 양이 엄청나게 많을 뿐만 아니라 원하지 않는 정보인 경우가 많다. 따라서 사용자의 기호를 반영하여 원하는 정보만을 찾아줄 수 있는 시스템이 필요하다. 본 논문에서는 사용자의 기호를 반영하여 개인에게 특화된 웹 검색 시스템을 개발하기 위하여 유전자 알고리즘으로 최적화된 다중에이전트 시스템을 제안한다. 실험결과 사용자 프로파일 벡터가 사용자기호에 따라 변화하여 제안한 시스템이 유용함을 알 수 있었다.

1. 서 론

정보의 양과 다양성으로 인해 정보의 바다라고 불리는 인터넷에서 사용자가 원하는 정보를 찾기가 점점 어려워지고 있다. 이러한 문제를 해결하기위하여 WWW에서의 정보검색을 도와주는 웹검색엔진들이 개발되고 있다. 대표적인 시스템인 Yahoo[1]나 Lycos[2]와 같이 사람이 직접 웹사이트를 등록해서 디렉토리서비스를 하는 것도 있으며, MetaCrawler[3]와 같이 여러 검색엔진들로부터 결과를 받아오는 메타검색엔진 등도 개발되었다. 그러나 이들 검색엔진은 일반적인 사용자를 위한 것이어서, 대개의 경우 검색결과가 너무 많아 일일이 페이지들을 읽어서 원하는 정보를 찾는 것이 매우 불편하다.

여러개의 검색엔진을 사용자의 기호에 맞게 결합하는 시스템으로는 NewT[1]와 Amalthaea[2][3]가 대표적이다. NewT는 뉴스그룹에서 효과적인 정보검색을 위해 진화와 적합성 피드백을 사용하는 다중에이전트 시스템이다. 반면에, Amalthaea는 NewT와 방식은 비슷하지만 영역을 WWW으로 확장시킨 검색시스템이다. 특히, Amalthaea는 사용자의 기호를 반영하는 개인화된 시스템으로 가능성이 매우 크지만, 초기 프로파일 구축단계에서 사용자가 관심있어 하는 URL의 리스트를 입력해야 한다는 점에서 숙달되지 않은 사용자는 이용하기 힘들다.

따라서 본 논문에서는 사용자가 원하는 분야의 키워드로 초기 프로파일을 구축하고, 웹문서에 대한 사용자의 기호를 가중치 벡터로 표현하며, 사용자의 피드백을 받아 지속적으로 사용자 기호를 반영하여 정보를 웹에서 쉽게 검색할 수 있도록 하는 개인화된 정보검색 시스템을 다중에이전트 모델과 유전자 알고리즘[4]을 이용해 구현하고자 한다. 이 시스템은 사용자가 빈번하게 검색하는 특정 분야의 정보를 검색하고자 할 때 유용하게 사용될 수 있다.

2. 방 법

제안하는 방법은 시스템이 제시한 문서에 대한 사용자의 피드백을 받아서 프로파일을 갱신시켜 나감으로써, 사용자가 원하는 정보를 지속적으로 제공한다. 시스템은 그림 1과 같이 크게 키워드수집 에이전트(Keyword Collection Agent)와 정보필터링 에이전트(Information Filtering Agent)로 구성된다. 키워드수집 에이전트는 사용자가 명시한 분야에 대한 핵심키워드로 초기 프로파일 벡터를 생성하는 역할을 한다. 정보필터링 에이전트는 사용자가 흥미있어할 만한 문서들을 찾아서 다이제스트 형태로 제시해준다. 이 밖에 프로파일 생성기는 사용자가 주는 피드백 정보로 사용자의 프로파일을 갱신시켜서, 사용자의 기호를 학습하게 한다.

2.1 키워드수집 에이전트

사용자의 초기 프로파일을 얻는 방법으로는 사용자에게서 직접 입력받는 방법과 사용자가 자주 방문하는 사이트의 URL, 즉 웹 브라우징한 이력을 입력받음으로써 간접적으로 사용자의 기호를 습득하는 방법이 있다. 본 시스템은 사용자가 원하는 분야를 복수로 유지하면서 각 분야에 대해 사용자

1) http://www.yahoo.com
2) http://www.lycos.com
3) http://www.metacrawler.com

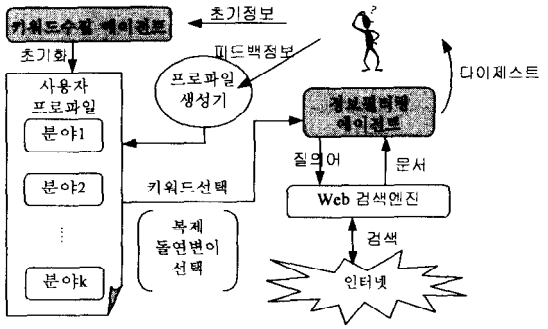

그림 1. 시스템 구조

의 기호를 반영하여 정보를 발굴하기 위해, 초기에 사용자가 원하는 분야를 직접 입력받는 방식을 사용한다.

사용자는 시스템 초기에 자신이 관심있는 분야를 나타내는 키워드를 "java, xml"과 같이 입력한다. 사용자가 입력한 키워드외에도 사용자의 기호를 나타낼 수 있는 단어가 있을 수 있으므로 사용자의 기호를 나타낼 수 있는 단어를 추가로 수집하는 과정을 수행한다. Yahoo 검색엔진에 사용자가 입력한 키워드를 질의어로 해서 나온 결과의 URL 집합에서 시스템내의 키워드수집 에이전트가 키워드들을 추출한다. Yahoo에 등록되는 대부분의 웹페이지는 html의 메타태그로 키워드를 명시하므로, 이 태그로 명시된 단어들을 키워드로 선택한다.

초기 프로파일 벡터내의 키워드 k의 가중치는 식 (1)과 같이 계산한다.

$$W_k = \log(f(k)/N \times C) \qquad (1)$$

$f(k)$는 키워드 k의 출현빈도로서 총 검색된 웹페이지 중에서 키워드 k가 나타난 회수, N은 Yahoo에서 검색된 웹페이지의 메타태그에서 지정된 키워드들의 총 개수, C는 숫자가 너무 작아지는 것을 보정해 주는 상수이다.

2.2 정보필터링 에이전트

정보필터링 에이전트는 사용자의 프로파일 벡터를 바탕으로, 사용자가 흥미있어할 만한 문서들을 찾아 다이제스트형태로 제시해주는 역할을 한다. 사용자의 프로파일 벡터에서 일정 개수의 키워드를 추출하여, 웹 검색엔진에 질의어로 입력한 후 나온 문서들을 필터링하여 사용자에게 추천한다. 이때 기존의 웹 검색엔진을 활용하므로 메타검색엔진이라 할 수 있다. 정보필터링 에이전트는 일정 개수의 키워드와 가중치로 이루어진 키워드 벡터를 유전자형으로 가지며, 키워드의 추출방법으로 유전자 알고리즘의 돌연변이와 선택, 복제의 방법을 사용한다.

복제는 사용자로부터 좋은 피드백을 받은 에이전트의 유전자형을 그대로 사용하여 다음 세대를 생성하는 것이고, 선택은 사용자 프로파일 벡터의 가중치를 기반으로 Roulette Wheel Selection을 사용하여 키워드를 추출해내는 방법이며, 돌연변이는 사용자 프로파일 벡터에서 랜덤하게 키워드를 추출해내어 좋은 키워드이나 아직까지 가중치가 낮을 수 있는 키워드를 추출하게 한다.

다중에이전트기법을 도입하여 한 세대에서 여러 개의 정보필터링 에이전트를 동시에 실행시킴으로써, 사용자가 원하는 정보를 빠르고 다양하게 찾아올 수 있다. 그림 2는 사용자 프로파일과 필터링 에이전트와의 관계를 나타낸 것이다.

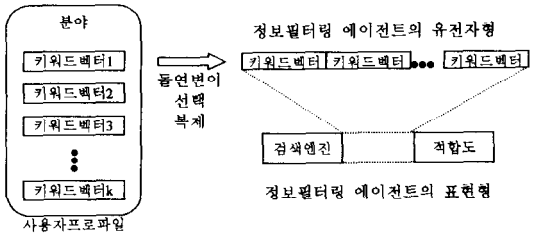

그림 2. 사용자 프로파일과 필터링 에이전트와의 관계

2.3 요약 및 제시

정보필터링 에이전트가 찾아온 문서들 중에서 사용자에게 제시해 줄 문서를 걸러내는 작업이 필요하다. 각 문서의 중요도를 구해서 높은 중요도를 얻은 문서들을 사용자에게 요약하여 제시한다.

정보필터링 에이전트가 찾아온 i번째 문서의 중요도 C_i 는 다음과 같이 구한다[3].

$$C_i = D_{IFA} \times F_i \qquad (2)$$

F_i는 정보필터링 에이전트의 적합도이며, D_{IFA}는 정보필터링 에이전트의 유전자형과 문서벡터사이의 거리이다. 두 개의 가중치 키워드벡터 a와 b사이의 거리는 다음과 같이 구한다[3].

$$D_{IDF_{a,b}} = \frac{\sum_{k=1}^{i} W_{ak} \cdot W_{bk}}{\sqrt{\sum_{k=1}^{i}(W_{ak})^2 \cdot \sum_{k=1}^{i}(W_{bk})^2}} \qquad (3)$$

다이제스트에서 보여줄 문서의 개수는 사용자가 설정한 환경변수로 결정된다.

2.4 사용자 기호 학습

시스템이 제시하는 문서는 사용자에 의해 −3에서 3까지로 평가받게 되는데, 3이 가장 선호하는 문서임을 나타낸다. 사용자의 평가에 따라 식 (4)를 이용하여 사용자 프로파일내 키워드의 가중치를 갱신한다.

$$W_n = W_p + R_l \times feedback \times W_{dk} \qquad (4)$$

W_n 은 키워드의 새로운 가중치, W_p는 키워드의 현재 가중치 R_l은 학습률인데 이 숫자를 크게 할 수록 가중치의 변화가 심해진다. $feedback$은 −3에서 3까지의 사용자 피드백이다. W_{dk}는 문서내에서 키워드의 가중치이다. 문서내에서의 키워드 가중치는 TFIDF방법을 이용해 구한다[5]. 이렇게 변화된

프로파일 벡터를 이용하여 새로운 세대를 생성한다. 세대가 변화할수록 사용자가 선호하는 키워드들이 사용자의 프로파일내에 높은 가중치를 가지게 된다.

그러나 프로파일 벡터내의 키워드만 가지고 위의 과정을 반복하면, 초기 프로파일내의 가중치만 변화되므로 사용자의 새로운 기호를 반영하기 힘들다. 프로파일내의 키워드들을 변화시키기 위해서, 높은 점수를 얻은 문서에서 프로파일 벡터내에 없는 키워드를 추출해서 프로파일 벡터에 추가하고, 추가된 키워드 개수만큼 프로파일 벡터에서 가중치가 낮은 키워드를 제거하여, 프로파일 벡터의 크기는 유지시킨다.

3. 실험결과

본 시스템의 목적은 웹에서 사용자의 기호에 맞는 정보를 찾아 주는 것으로, 피드백을 통해 사용자의 기호를 학습해 나가게된다. 사용자의 프로파일 벡터는 사용자의 피드백에 따라 가중치가 변화되어, 사용자가 원하는 키워드가 점점 더 높은 가중치를 얻게 되는지 실험하였다. 실험에 참가한 사용자는 java를 xml에 어떻게 적용할 것인지, java와 xml을 이용한 소프트웨어가 있는지와 같은 기술적인 문제 등 폭넓게 정보를 얻으려는 컴퓨터 프로그래머이다.

사용자가 입력한 키워드는 "java xml"이다. 사용자가 입력한 키워드를 바탕으로 초기 프로파일 벡터를 구성한다. Yahoo 검색엔진에 "java xml"로 질의어를 주고, 그 결과 URL들을 받아온다. 사용자가 선택한 키워드외에도 사용자의 기호를 반영할 만한 키워드들을 선정하기 위해, 검색엔진의 결과로 나온 URL로부터 키워드를 추출한다. 이때 html 태그 중 메타태그를 검색하여 키워드로 명시한 것들만 선택한다. 이렇게 하여 구축된 초기 프로파일 벡터는 표 1과 같다. 초기 가중치는 (1)식을 이용해 구하고, 프로파일 벡터의 총 개수는 25개로 제한하였다.

키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치
java	0.91	xml	0.64	database	0.27	design	0.14
web	0.35	html	0.32	developer	0.09	script	0.05
program	0.05	perl	0.09	windows	-0.2		

표 1. 초기 프로파일 벡터

유전자 알고리즘을 이용하여, 검색엔진에 질의어로 줄 키워드들을 초기 프로파일 벡터의 키워드 중에서 선택하여 질의어 집합을 만든다. 만들어진 질의어 집합을 필터링 에이전트의 유전자형으로 하여, 필터링 에이전트가 사이트를 검색하도록 한다. 필터링 에이전트의 개수는 5개로 제한하여 시스템에 무리가 가지 않도록 하였다.

표 2는 각 세대에 생성된 필터링 에이전트의 유전자형과 각 에이전트가 추천한 문서에 대한 사용자의 피드백을 보여준다. 사용자가 환경변수로 설정한 다이제스트에서 보여줄 문서의 개수는 3개였다.

그림 3은 세대가 변화하면서 프로파일내의 키워드들의 가중치 변화를 그래프로 나타낸 것이다. 실험에서 시스템이 세대가 변함에 따라 사용자가 관심있어 하는 분야의 키워드를 지속적으로 찾아내고, 키워드의 가중치를 변화시켜 감을 알 수 있었다.

세대	유전자형	피드백	유전자형	피드백
1	java, xml, web	+1	java, web, html	0
	java, xml, database	+2		0
2	java, xml, script	-3	java, xml, web	+3
	script, perl, database			+2
3	developer, swing, program	-3	java, xml, web	+3
			java,xml,database	+1
4	java, swing, program	-2	java,xml, database	+2
	java, xml, web	+2		
5	java, internet, javabean	+1	windows, program, javascript	+1
	java, javabean	+1		+1
6	developer, program, corba	-3	windows, program, javascript	-3
	java, xml, web	+2		
7	javabean, web, internet	-1	java, xml, perl	+2
			developer, xml, web	+1

표 2. 세대별 에이전트의 유전자형과 사용자피드백

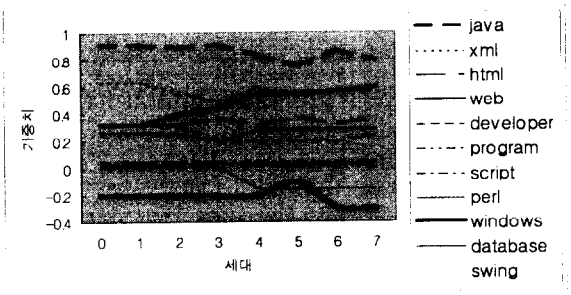

그림 3. 세대별 사용자 프로파일내 키워드의변화

4. 결 론

본 논문에서는 유전자 알고리즘과 다중 에이전트기법을 이용해서, 사용자의 기호를 반영하여 지속적인 정보검색을 할 수 있는 시스템을 구현해 보았다. 실험결과에서와 같이 사용자의 피드백이 증가할수록 사용자의 기호가 높은 단어들이 높은 가중치를 받음을 알 수 있었다.

참고문헌

[1] B. Sheth, "A Learning Approach to Personalized Information Filtering," *M.S. Thesis*, MIT, 1994.

[2] A. G. Moukas, "Amalthaea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem," *Proc. of Conf. on Practical Application of Intelligent Agents and Multiagent Technology*, London, 1996.

[3] A. G. Moukas and G. Zacharia, "Evolving a Multi-agent Information Filtering Solution in Amalthaea," *Proc. of Int. Conf. on Autonomous Agent*, Marina Del Ray, California, 1997.

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

[5] G. Salton and C. Buckley, "Text Weighting Approaches in Automatic Text Retrieval," *Cornell University Technical Report* 87-881, 1987.