정보검색에서 다유전자군 관리에 의한 사용자 프로파일 시스템#

김남일, 박영찬*, 남기춘, 최기선
한국과학기술원 전산학과 컴퓨터 시스템 연구실
*시스템공학 연구소 자연어처리 연구부 정보검색실

User Profile System Using Gene Group Management in Information Retrieval

Namil Kim, Young Chan Park*, Kichun Nam, and Key-Sun Choi
Korea Advanced Institute of Science and Technology,
Computer Science Dept. Computer System Lab.
***System Engineering Research Institute,**
Dept. of Natural Language Information Processing Information Retrieval Lab

사용자 프로파일(user profile)은 사용자 개개인의 관심 분야를 기술한 것이다. 정보 시스템이 사용자 개개인의 관심 분야를 알고 있다면, 사용자의 관심 분야에 속하는 문서만을 사용자에게 제시해 줌으로써 정보검색의 효율을 높일 수 있다. 사용자 프로파일 관리 시스템은 시간이 지날수록 개인 사용자의 관심분야를 더욱 잘 나타내어야 하고(Specialization), 사용자의 관심 분야가 변화해도 그에 대응해야 하며(Adaptation), 사용자가 흥미있어할 만한 관련 분야도 제시할 수 있어야 한다(Exploration). 본 논문은 기존에 연구되어 왔던 정보여과 시스템에서 사용되는 사용자 프로파일 관리 시스템들의 단점을 보완하고 정보검색 시스템에서도 사용될 수 있는 범용성을 지닌 사용자 프로파일 관리 시스템을 제안한다. 제안하는 사용자 프로파일 관리 시스템은 사용자의 관심분야의 변화를 유전자 알고리즘을 이용하여 모델링하고, 여러 유전자군(群)으로 사용자의 다양한 관심분야를 기술한다.

제 1 장 서론

프로파일의 필요성

사용자 프로파일(User Profile)은 사용자 개인의 관심 분야를 기술한 것을 말한다. 정보검색 시스템이 사용자 개개인의 관심 분야를 알고 있다면, 문서를 검색할 때 사용자의 관심 분야에 속하는 문서만을 사용자에게 제시해 주거나 관심 분야의 문서를 그렇지 않은 문서들보다 목록의 상위에 놓음으로써 사용자에게 편의를 제공할 수 있다. 한편, 정보검색 시스템의 사용자는 한 개나 두 개의 단어로 이루어진 짧고 의미가 넓은 질의를 주로 사용한다. 정보검색 시스템이 사용자의 관심분야를 고려하여 짧은 질의를 좀 더 자세하고 명확한 질의로 확장해 준다면 사용자는 모호한 질의의 결과로 검색된

본 논문은 1997 년도 정보통신부의 지원을 받아 수행된 "다국어 정보검색 아키텍춰"의 일환으로 이루어졌다.

지나치게 많은 문서들을 모두 확인해 보지 않아도 될 것이다.

즉, 사용자는 자신의 관심분야가 아닌 문서는 아예 읽지 않거나 다른 문서보다 나중에 접하게 됨으로써 쓸모없는 문서를 읽는데 낭비되는 시간을 줄일 수 있고, 자신이 원하는 정보에만 집중할 수 있다.

정보검색과 정보여과에서의 프로파일

정보여과(information filtering)는 끊임없이 새롭게 유입되는 문서들 중에서 사용자의 관심에 적합한 문서만을 사용자에게 전달하는 과정이다. 이 과정에서 사용자의 관심과 관련없는 문서는 걸러진다. 따라서, 사용자의 관심분야를 기술한 사용자 프로파일은 정보여과 과정에서 여과기라는 핵심적인 위치를 차지한다.

정보검색 시스템(Information retrieval system)에서 사용자 프로파일은 시스템 기능의 핵심을 이루는 구성요소는 아니다. 그러나 사용자 프로파일을 도입함으로써 정보검색 시스템은 사용자에게 보다 지능적이고 능동적인 서비스를 제공할 수 있다. 가령 사용자가 여러 의미를 갖는 단어를 질의어로 사용했을 때, 사용자의 관심분야에 해당하지 않는 의미로 질의어가 사용된 문서들을 자동적으로 제거해 줄 수 있다. 또한, 문서 집합에 새로운 문서가 추가되었을 경우, 사용자의 관심과 일치한다면 사용자가 요구하기 전에 능동적으로 사용자에게 새로운 문서를 제시해 줄 수 있다. 그리고, 사용자가 정보검색 시스템의 전체 문서에서 관심분야의 문서들을 브라우즈할 수도 있다.

이렇게 사용자 프로파일은 정보여과 시스템에서는 여과기로서, 정보검색 시스템에서는 지능적이고 능동적인 기능을 제공하는데 필요한 정보로서 사용된다.

해결하고자 하는 문제의 범위

본 논문은 기존의 정보여과 시스템에서 사용되는 사용자 프로파일 관리자가 가진 단점들을 보완하고 정보검색 시스템에서도 사용될 수 있는 범용성을 지닌 사용자 프로파일 관리 시스템을 제안한다. 사용자 프로파일 관리자는 크게 다음과 같은 세 가지 기능을 가지고 있어야 한다.

- 특화성(Specialization) : 개인 사용자의 관심 분야를 기술한다.
- 적응성(Adaptation) : 사용자의 관심 분야가 변화해도 그에 맞게 관심분야를 기술한다.
- 탐색성(Exploration) : 사용자가 흥미있어 할 만한 비슷한 다른 관심분야도 기술한다.

제 2 장 기존연구 및 문제점

이 장에서는 지금까지 연구된 여러 사용자 프로파일 시스템과 이들의 장단점에 대하여 살펴본다.

SIFT[Yan, 1995]

SIFT는 정보여과 시스템이며 단순한 단어목록(keyword list)으로 이루어진 사용자 프로파일을 이용한다. 사용자는 관심 분야 각각에 대하여 한 개씩의 프로파일을 직접 만든다. 프로파일의 수정은 사용자에 의해 직접 이루어지거나 적합성 피드백(relevance feedback)을 통하여 자동적으로 이루어진다.

이 시스템은 프로파일의 생성 및 수정작업을 대부분 사용자가 직접 해야 하고 적합성 피드백 기법을 이용하여 프로파일을 자동으로 수정한다 하더라도 단어의 가중치만 수정될 뿐 단어 자체가 추가되거나 삭제되지는 않는다. 즉, 프로파일 관리 책임의 대부분을 사용자에게 맡겨 두고 있다. 따라서 사용자가 직접 자신의 관심을 프로파일로 나타내야 하고 관심이 바뀌었을 경우에도 직접 프로파일을 수정해야 한다. 하지만, 때로는 사용자가 자신의 관심을 제대로 나타내지 못하는 경우도 있고, 관심이 변할 때마다 프로파일을 수정하는 것도 귀찮은 일이다.

단어목록과 적합도 피드백을 이용한 시스템

JASPER[Davies & Weeks, 1997]나 FAB[Balabanovic, 1997]등 웹(Web)에서 HTML 문서를 여과하거나 추천하는 시스템들[Pazzani et al.][Balabanovic & Shoham, 1995]의 대부분이 사용자의 관심을 기술하는데 벡터공간모델(vector space model)에 기초한 가중치 단어목록(weighted keyword list)를 사용한다. 프로파일 생성은 사용자의 웹 페이지 접근 기록(web page access history)을 바탕으로 하여 빈번히 접근된 웹 페이지를 프로파일의 자료로 삼는다. 프로파일의 갱신에는 적합도 피드백이 이용된다.

이러한 시스템은 단순히 적합도 피드백을 이용하여 프로파일을 갱신하므로 사용자의 관심이 바뀌는 것에 적응할 수는 있지만 흥미있어할 만한 새로운 관심분야를 제시하지는 못한다. JASPER가 단어군집화(term clustering)를 이용하여 단어간의 유사도(similarity)에 바탕한 프로파일 확장 기법을 사용하지만 단어간의 유사도가 사용자의 관심이 아닌 문서집합내의 문서들을 분석함으로써 얻어진 것이기 때문에 사용자의 잠재적 관심을 나타내기에는 부적절하다.

사용자 모델(User Model)을 이용한 시스템

UMIE(User Model for Information Extraction) [Benaki et al., 1997]는 사용자 모델을 이용하였다. 하지만, 사용자 모델을 만들려면 적용 분야에 대한 완전한 지식이 필요하다. 따라서, 넓은 분야를 포함하는 일반적인 시스템을 구축하기 어렵고 사용자가 여러 전형(stereotype)에 걸쳐있을 경우 처리가 복잡하다는 단점이 있다.

유전자 알고리즘을 이용한 시스템

Amalthaea[Moukas, 1996]나 NewT[Sheth, 1994]는 벡터공간모델에 기초한 유전자 알고리즘을 사용하여 프로파일을 관리한다. 각각의 유전자는 문서벡터이고 사용자의 관심과 유사한 유전자는 살아남고 그렇지 못한 유전자는 도태한다. 하지만 이들은 사용자의 다양한 관심분야를 제대로 처리하지 못했다. Amalthaea는 사용자의 여러 관심분야를 하나의 유전자 집합으로 나타냄으로써 유전자 알고리즘의 효율이 떨어진다. 실험결과에 나타나 있듯이 관심분야의 수가 증가할수록 프로파일의 성능이 떨어진다.

반면, NewT는 관심분야를 엄격히 나누었다. 관심분야 끼리는 영향을 거의 미치지 않고 각 관심분야가 독자적으로 관리된다. 따라서, 하나하나의 관심분야는 프로파일 시스템이 관리하지만 관심분야가 새로 생기거나 없어지는등 관심분야 자체의 변화는 사용자가 직접 관리해야 한다.

시소러스(thesaurus)를 이용한 시스템

[Bloedorn et al., 1996]가 제안한 시스템은 주제분야 코드라는 일종의 시소러스 분류와 고유명사와 관련된 인명, 기관명, 지역명 정보, 그리고 단어의 빈도 정보등 3개의 특징(feature)을 이용하여 사용자 프로파일을 구성한다.

시소러스를 이용하면 단어의 일반화(generalization)가 이루어져 단어끼리 정확히 일치(exact match)하지 않아도 된다는 장점이 있지만 시소러스는 만들기 어렵고 유지보수도 힘들다. 또한, 시소러스를 만드는데 제작자의 주관이 개입되기도 하고 구조가 경직되어 있기 때문에 개개인에게 적합하도록 구조를 바꾸는 것도 어렵다. [Bloedorn et al., 1996]의 실험에서 사용한 시소러스도 실험에 관련된 분야를 수동으로 확장해서 사용했다.

단어의 공기정보(共起情報)를 이용한 시스템

IfWeb[Asnicar & Tasso, 1997] 시스템은 가중치 의미 네트워크(weighted semantic network)로 사용자 프로파일을 표현한다. 의미 네트워크의 노드(node)는 단어이고 에지(edge)는 문서에서 함께 나타난 단어를 연결한다. 이 모델은 사용자가 그 분야에 관심이 없다는 사실을 표현할 때, 오랫동안 에지에 해당하는 피드백이 없으면 에지의 가중치를 낮춘다. 그러나, 피드백이 없다는 사실은 사용자가 그 분야에 관심이 없다는 것이 아니라 현재의 관심분야가 아니라는 사실을 의미할 뿐이다.

정보검색과 정보여과의 차이점에 대한 고려

위에서 살펴본 사용자 프로파일에 대한 기존연구의 대부분은 정보여과 시스템과 관련된 것들이다. 정보검색 시스템은 정보여과 시스템과 개념적으로 유사하다. 하지만 두 시스템을 사용할 때 사용자의 관심 범위가 서로 다르다.

정보검색 시스템에서 사용자 질의의 역할을 정보여과 시스템에서는 사용자 프로파일이 담당한다. 정보검색에서 사용자 질의는 그 순간의 특정한 정보요구(information need)를 표현하기 때문에 구체적이다. 하지만, 정보여과 시스템의 사용자 프로파일은 새로운 문서가 어떤 것이 될 지 모르기 때문에 하나의 관심분야에 대하여 비교적 넓은 범위를 나타낸다.

가령, 사용자가 야구에 관심을 가지고 있을 때, 정보여과 시스템의 프로파일에는 야구 전반에 걸친 관심이 기술되어야 하지만 정보검색 시스템에서 실제로 입력되는 것은 어떤 야구선수에 대한 기록을 원하는 것과

같은 구체적인 질의이다.

따라서, 정보검색 시스템에서 사용되는 사용자 프로파일 관리자는 질의들간의 공통적인 특성을 찾아서 이 질의들을 하나의 그룹으로 만들어 주어야 한다.

제 3 장 시스템 설계

본 논문에서 제안하는 사용자 프로파일 관리자는 유전자 알고리즘으로 개개의 관심분야 변화를 모델링하고, 여러 유전자군을 사용하여 사용자의 다양한 관심분야를 나타낸다. 각각의 유전자군은 사용자의 관심정도를 반영하며 수시로 크기가 변화한다.

각각의 유전자군은 유전자 알고리즘에 따라 진화함으로써 시간이 지날수록 사용자의 관심분야와 일치하는 유전자만이 살아남아 사용자의 관심분야를 더 정확하게 기술하고(특화성), 관심분야의 변화에도 잘 대처한다(적응성). 한편, 유전자 연산 중 하나인 돌연변이는 기존 관심분야로부터 새로운 관심분야를 끊임없이 생성하여 사용자에게 제시한다(탐색성).

제 1 절 개관

유전자 알고리즘(Genetic Algorithm)

유전자 알고리즘이란 자연선택, 돌연변이와 같은 생물진화의 원리에 착상을 얻은 알고리즘으로, 확률적 탐색, 학습, 최적화의 한 방법이다[北野, 1993]. 유전자 알고리즘은 각 세대마다 이전 세대에서 가장 적합한 유전자들을 선택하여 교차(crossover)와 돌연변이(mutation)와 같은 연산을 적용하여 새로운 유전자들을 만들어 낸다. 이러한 과정은 임의적으로 보이지만 유전자 알고리즘은 과거 정보로부터 효율적으로 향상된 새로운 탐색 지점을 찾아낸다. 유전자 알고리즘은 강건(robust)하고[Goldberg, 1989] 사용자 프로파일을 관리할 때 적응성과 탐색성이 뛰어나다는 장점이 있다. 또한, 사전지식이 거의 필요없다는 장점도 가지고 있다.

유전자군

사용자 프로파일 관리 시스템은 사용자의 다양한 관심분야를 나타내기 위하여 복수의 유전자군(群)을 이용한다. 각 유전자군은 사용자의 관심분야 중 하나를 나타낸다. 그런데 사용자의 관심분야는 고정되어 있지 않고 시간에 따라 변화한다. 이 현상을 모델링하려면 유전자군이 사용자 관심분야의 변화에 따라 역동적으로 변화해야 한다.

사용자가 새로운 분야에 관심을 보이면 그 분야를 나타내는 유전자군이 새롭게 생성된다. 그 분야에 관심이 많아지면 대응하는 유전자군은 점점 커져서 보다 넓은 관심분야를 표현한다. 유전자군이 지나치게 넓은 관심분야를 표현하게 되면 유전자 알고리즘의 효율이 나빠지므로 커다란 유전자군을 여러 개의 작은 유전자군으로 나뉘어진다. 사용자가 더 이상 이 분야에 관심이 없어지면 유전자군은 소멸한다.

새로운 유전자군을 생성하는 작업은 주의깊게 이루어져야 한다. 새로운 유전자군이 자주 생성되어 유전자군이 지나치게 많아질 경우, 사용자의 관심분야는 여러 유전자군에 일부분씩 걸쳐있게 된다. 이 때 관심분야가 변하면 하나의 유전자군이 변화에 적응하는 방향으로 진화하는 대신 관심분야의 변화에 대응하는 인접한 또다른 유전자군이 발달한다. 따라서, 하나의 유전자군이 하나의 사용자 관심분야를 줄곧 따라가며 진화하지 않으므로 시간이 지남에 따라 사용자의 관심분야를 점점 정확하게 기술하는 특성화가 불가능하다. 그러므로 지나치게 많은 유전자군이 존재하면 프로파일 시스템의 성능이 떨어진다.

피드백의 영향

피드백은 정보 시스템의 출력에 대한 사용자의 반응이다. 사용자 프로파일 관리 시스템에서의 피드백은 사용자가 자신의 질의에 적합하다고 판단한 문서들의 색인이다. 관련 문서들은 사용자가 명시적으로 정보 시스템에 알려주거나, 암시적으로 정보 시스템이 사용자의 행위를 관찰하여 알아낼 수 있다.

사용자 프로파일 관리 시스템은 사용자가 질의를 할 때마다 질의가 속한 관련분야를 예측한다. 이 예측은 질의와 유전자간의 비교에 기반하며, 각 유전자군이 독자적으로 관련분야(관련 단어 목록, 질의와의 유사도)를 예측한다. 이 예측값과 사용자의 피드백을 비교하여 각 유전자군을 어떻게 진화시킬 것인지 결정한다.

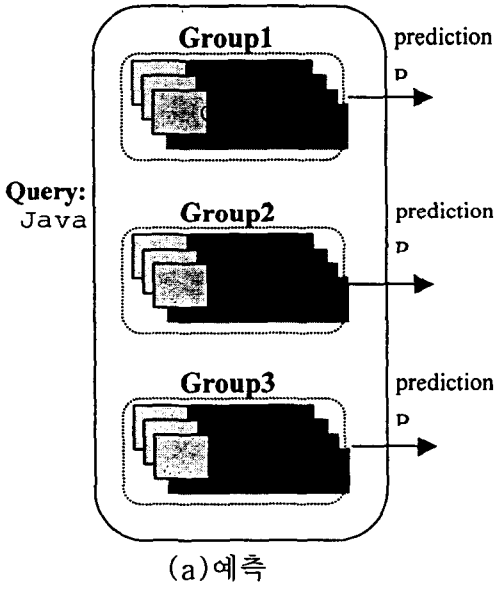
둘 이상의 유전자군의 예측이 비슷하다면 이 유전자군들은 비슷한 사용자 관심분야를 표현하고 있는 경쟁관계이다. 하나의 관심

분야는 하나의 유전자군으로 표현되어야 하므로 가장 유사도가 높은 유전자군만 피드백과 같은 방향으로 유전자의 수를 늘려 진화하도록 하고 나머지 유전자군은 피드백과 다른 방향으로 진화하도록 하면서 유전자군의 크기를 줄인다.

피드백과 가장 유사도가 높은 유전자군 안에서 피드백과 유사도가 높은 유전자는 살아남아 자손을 만들고, 유사도가 낮은 유전자는 도태된다. 그 외의 유전자군은 피드백과 유사도가 높은 유전자를 도태시키고 피드백과 유사도가 낮은 유전자 수가 늘어나도록 하여 피드백과 다른 방향으로 진화하도록 한다.

전체적인 알고리즘

1. 사용자의 질의에 대하여 각 유전자군이 질의가 속한 관심분야를 예측하여 관련단어 목록을 출력한다.(그림 1(a))



(a)예측

그림 1 전체적인 알고리즘

2. 사용자로부터 입력받은 피드백과 1 에서 예측한 관련단어 목록을 비교하여 유사도를 구하고 이로부터 각 유전자군의 다음세대 크기를 계산한다.

3. 각각의 유전자군을 유전자 알고리즘에 따라 진화시킨다.(그림 1(b))

4. 유전자군의 크기에 따라 유전자군을 새로 생성하거나, 하나의 유전자군을 여러 개로 분리하거나, 두 유전자군을 하나로 융합시킨다.(그림 1(c))

제 2 절 시스템 구성

1. 적합도 피드백과 유전자의 표현

피드백과 유전자는 질의부분(query part)과 관련부분(relevant part)으로 이루어진다.

$$F = (Q, R)$$
$$Q = (q_1, q_2, q_3, \mathrm{K}, q_m)$$
$$R = (r_1, r_2, r_3, \mathrm{K}, r_n)$$
$$q_i = (term_i, weight_i)$$
$$r_j = (term_j, weight_j)$$

질의부분은 질의어들로 구성되고, 관련부분은 질의와 관련된 문서들의 색인의 총합이다. 각 부분은 (단어, 가중치) 쌍들의 목록이다. 관련문서의 색인은 가중치가 부여된 단어목록으로, 벡터형태이다. 이 문서 벡터는 자신의 길이로 각 성분의 가중치를 나누어 정규화(normalization)한 후 다른 문서벡터와 더해진다. 피드백의 각 성분의 가중치 weight_i는 아래 수식과 같다.

$$weight_i = \sum_{k=1}^{n} \frac{w_{ki}}{|d_k|}$$
$$|d_k| = \sqrt{\sum_{i=1}^{t} w_{ki}^2}$$

$|d_k|$는 문서 k를 나타내는 벡터의 길이이고, w_{ki}는 문서 k의 i번째 단어의 가중치를 나타낸다.

2. 질의로부터 관심분야 예측

사용자로부터 질의가 들어오면 각각의 유전자군은 질의가 속한 관심분야를 예측한다. 관심분야 예측 p_g는 관심분야를 나타내는 단어목록이고 c_g는 이 예측에 대한 신뢰도(confidence)이다.

$$p_g = term\ list\ of\ the\ gene\ of \max c_k$$

$$c_k = \sum_{i=1}^{q} (q_i \cdot w_i)$$

$$where\ q_i = \begin{cases} qpriority & if\ term_i\ is\ in\ query\ part \\ 1 & otherwise \end{cases}$$

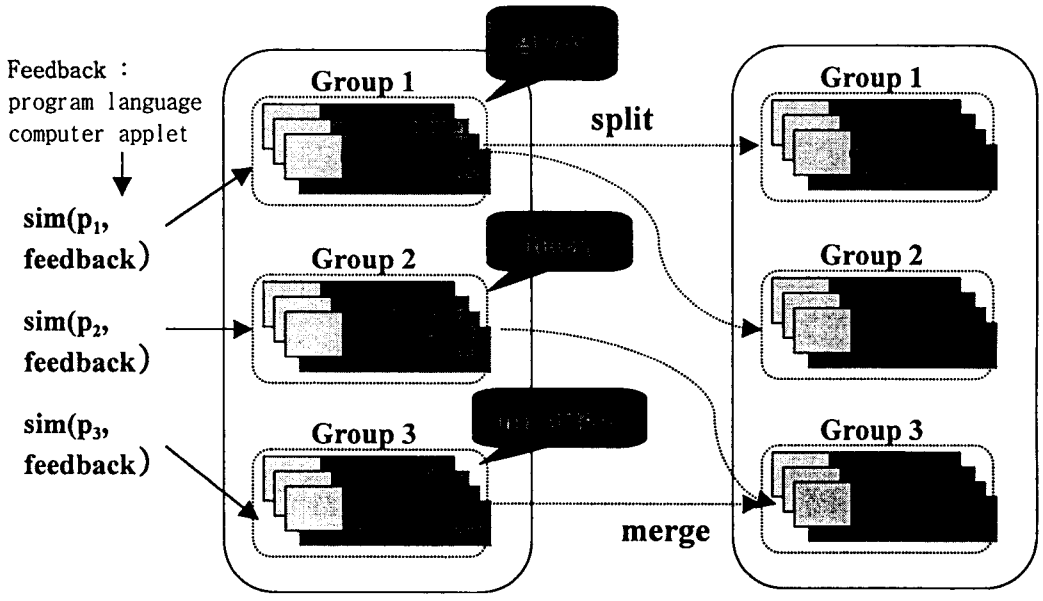


그림 1 전체적인 알고리즘

유전자군의 예측 p_g는 신뢰도가 가장 높은 유전자의 단어목록이고, 유전자군의 신뢰도 c_g는 이 유전자의 신뢰도이다. 유전자의 신뢰도는 유전자의 단어목록 중, 질의에 속한 단어들의 가중치를 모두 더한 값이다. 가중치를 더할 때 유전자의 질의 부분에 속한 단어는 가중치를 높인다.

각 유전자군의 관련분야 예측으로부터 가장 신뢰도가 높은 것을 질의가 속한 관련분야 예측으로 선택한다. 만약 둘 이상의 유전자군의 예측값에 대한 신뢰도가 같다면, 현재 사용자의 관심분야의 경향을 고려하여 유전자군의 크기가 큰 것을 선택한다.

3. 유전자 알고리즘

사용자 관심분야를 효율적으로 기술하고 관심분야의 변화에 대처하기 위하여 유전자 알고리즘을 사용한다. 사용자의 관심과 일치하는 유전자는 번성하고 그렇지 않은 유전자는 도태됨으로써 사용자의 관심분야를 기술한다.

초기상태

사용자로부터 관심분야와 일치하는 문서들을 입력받아 초기 유전자군을 생성한다.

자연선택(Selection)

적합도가 높은 유전자를 선택하여 다음 세대를 탄생시키도록 하는 작업이다. 각 유전자의 적합도 f'은 유전자와 피드백의 유사도 f를 스케일링(scaling)함으로써 얻어진다.

$$f' = (f'_{max} - f'_{min}) f + f'_{min} \quad for \quad feedback \quad direction$$

$$f' = (f'_{min} - f'_{max}) f + f'_{max} \quad for \quad other \quad direction$$

$$f = sim(feedback, gene)$$

유전자와 피드백의 유사도는 두 벡터의 코사인 유사도(cosine similarity)로 정의된다 [Frakes & Baeza-Yates, 1992].

$$sim(v_1, v_2) = \frac{\sum_{i=1}^{t} (q_{1i} w_{1i}) \cdot (q_{2i} w_{2i})}{|v_1| \cdot |v_2|}$$

$$0 \le sim(v_1, v_2) \le qpriority^{\,2}$$

q_i는 단어 i의 질의어 가중치로서 단어 i가 유전자의 질의 부분(qeury part)에 속하면 qpriority 값을, 유전자의 관련 부분에 속하면 1의 값을 갖는다.

유전자와 피드백과의 유사도를 직접 유전자의 적합도로 사용하지 않고 적합도 스케일링(fitness scaling)을 한 이유는 유전자군이 초기에 국부최적해에 도달하는 것과 진화말

기에 유전자들이 서로 지나치게 비슷해지는 것을 막고[Goldberg, 1989] 진화의 방향을 조절하기 위해서이다.

위와 같은 방법으로 구한 적합도로부터 유전자가 자손을 남길 확률을 계산한다. 유전자 k가 자손을 남길 확률 p_k는 유전자군 전체의 적합도에서 자신의 적합도가 차지하는 만큼이다.

$$p_k = \frac{f'_k}{\sum_{i=1}^{pop\_size} f'_i}$$

교차(Crossover)

유전자 연산의 하나로서, 같은 유전자군에 속한 임의의 두 유전자 G_1, G_2로부터 유전자 내용의 일부를 서로 교환하여 새로운 두 유전자 G_3, G_4를 생성한다. 먼저 유전자내에서 임의의 두 위치 p_1, p_2를 결정한 후, p_1과 p_2 사이의 유전자 내용을 서로 교환한다.

$$G_1 \otimes G_2 \rightarrow G_3, G_4$$

$$p_1 = rand(1, sizeof(G)-2)$$

$$p_2 = rand(p_1, sizeof(G)-1)$$

$$G_3 = \begin{Bmatrix} G_{1i}, 0 < i < p_1 \; and \; p_2 < i \le sizeof(G)-1 \\ G_{2i}, p_1 \le i \le p_2 \end{Bmatrix}$$

$$G_4 = \begin{Bmatrix} G_{1i}, p_1 \le i \le p_2 \\ G_{2i}, 0 < i < p_1 \; and \; p_2 < i \le sizeof(G) \end{Bmatrix}$$

돌연변이(Mutation)

유전자의 관련 부분(relevant part)으로 임의의 색인어를 삽입한다. 이때 색인어의 가중치는 유전자내 단어들의 평균 가중치로 한다. 또한 유전자의 관련 부분에 속한 단어 중 가중치가 높은 것을 임의로 질의부분(query part)으로 보낸다.

4. 유전자군 관리

유전자 알고리즘에 의해 진화하는 유전자군들이 관심분야가 겹치지 않고 서로 다른 방향으로 진화하도록 관리한다.

유전자군의 크기

피드백과 유전자군의 예측을 비교하여 다음 세대의 유전자의 수, 즉 유전자군의 크기를 계산한다. 다음 세대에 증감할 유전자의 수 Δn은 진화의 방향과 현재 유전자군의 상태, 피드백과 관심분야 예측과의 유사도에 의하여 결정된다.

$$\Delta n = I \cdot c \cdot h(g,n) \cdot sim(feedback, prediction)$$

$$h(g,n) = \left(\frac{G}{g} \cdot \frac{N}{n} \right)^I \qquad \text{I}$$

$I = 1 \; or \; -1$

G : *proper # of groups*

N : *proper # of genes in a group*

c : *constant*

는 진화의 방향을 결정하는 인자로서 유전자군이 피드백에 대하여 가장 정확하게 예측한 군인 경우 피드백과 유사한 방향으로 진화하도록 1이 되고, 그렇지 않은 경우 피드백과 다른 방향으로 진화하도록 -1이 된다.

h(g,n)와 sim(f,p)는 진화의 강도를 결정하는 인자로서 유전자군의 수와 유전자군 내의 유전자 수가 적절한 수준과 차이가 클수록 진화를 가속시킨다. 또한 피드백과 관심분야 예측이 유사할수록 진화를 가속시킨다.

새로운 유전자군 생성

각 유전자군의 예측에 대한 확신이 임계치보다 낮고 피드백과 예측과의 유사도가 역시 임계치보다 낮을 때, 즉 기존의 어느 유전자군과도 유사성이 없을 때 새로운 유전자군을 생성한다.

유전자군 분할

하나의 유전자군이 지나치게 커져서 다른 유전자군을 지배(dominate)할 때 유전자 알고리즘의 성능을 향상시키고 세분화된 관심분야 예측이 가능하도록 유전자군을 여럿으로 나눈다.

먼저 유전자 사이의 유사도에 따라 유전자를 군집화하고 각 군집이 개별적인 유전자군이 되도록 분할한다.

유전자군 병합

유전자군이 지나치게 작아지고, 전체 유전자군의 수가 많을 때 유사도가 높은 다른 유전자군과 결합시켜 유전자군의 수를 줄인다.

제 4 장 결론

본 논문은 정보여과 시스템에서 사용된 여러 프로파일 관리 시스템의 문제점을 해결하고 정보검색 시스템에서도 사용가능한 사용자 프로파일 관리 시스템을 제안하였다. 제안된 시스템은 사용자의 관심분야를 유전

자 알고리즘을 통하여 추적한다. 정보여과 시스템 뿐 아니라 정보검색 시스템에서도 사용할 수 있도록 질의 부분과 관련 부분으로 구성된 유전자 구조를 설계하였으며 이 구조에 적합하도록 유전자 알고리즘을 수정하였다. 한편 사용자의 다양한 관심분야를 표현할 수 있도록 여러 개의 유전자군을 두고, 각각의 유전자군이 관심분야를 효과적으로 추적할 수 있도록 관리하는 알고리즘을 제안하였다.

지금까지 정보검색 시스템은 사용자 개개인의 특성을 무시하고 모두 동일하게 취급했다. 이 때문에 사용자는 자신의 정보요구와 맞지 않는 정보까지 처리해야만 했다. 하지만 사용자 프로파일을 도입함으로써 사용자는 정보과부하(information overload)상태에서 벗어나 효율적으로 정보를 검색할 수 있다. 정보여과 시스템도 사용자 프로파일을 이용하지만 프로파일 관리에 문제점을 가지고 있다. 그러나 제안된 사용자 프로파일 관리 시스템은 사용자의 직접 개입 없이 사용자의 관심이 바뀌면 자동적으로 프로파일을 갱신하고 사용자가 다양한 관심분야를 가지고 있어도 효율적으로 관심분야를 기술한다. 아울러 이 시스템은 정보검색 시스템과 정보여과 시스템 모두에 사용할 수 있다.

본 논문에서 제시한 유전자군 관리 알고리즘은 유전자군을 생성, 분할, 병합할 때 임계치를 사용한다. 현재 이들 임계치는 실험에 의하여 적절한 값을 사용하고 있으나 향후 연구를 통하여 사용자 관심분야의 변화에 따라 자동으로 적절한 임계치를 구하는 방법을 개발해야 할 것이다.

참고문헌

[Armstrong et al., 1995] Robert Armstrong, Dayne Reitag, Thorsten Joachims and Tom Mitchell, WebWatcher : A Learning , Apprentice for the World Wide Web, AAAI spring symposium '95 on information gathering from heterogeneous, distributed environments

[Asnicar & Tasso, 1997] Fabio A. Asnicar and Carlo Tasso, ifWeb : a Prototype of User Model-Based Intelligent Agent for Document Filtering and Navigation in the World Wide Web, in Proceeding of the workshop "Adaptive System and User Modeling on the World Wide Web", June 1997

[Balabanovic & Shoham, 1995] Marko Balabanovic and Yoav Shoham, Learning Information Retrieval Agents: Experiments with Automated Web Browsing, AAAI spring symposium '95 on information gathering from heterogeneous, distributed environments

[Balabanovic, 1997] Marko Balabanovic, An Adaptive Web Page Recommendation Service, in the First international Conference on Autonomous Agents, Feb 1997

[Benaki et al., 1997] Eftihia Benaki, Vangelis A. Karkaletsis and Constantine D. Spyropoulos, User Modeling in WWW : the UMIE Prototype, in Proceeding of the workshop "Adaptive System and User Modeling on the World Wide Web", June 1997

[Bloedorn et al., 1996] Eric Bloedorn, Inderjeet Mani, T. Richard McMillan, Machine Learning of User Profiles : Representational Issues, In proceedings of AAAI '96 /IAAI '96

[Davies & Weeks, 1997] N. J. Davies and R. Weeks, Information Agents for the World Wide Web, M. C. Revett, Software Agents and Soft Computing, Springen, 1997

[Frakes & Baeza-Yates, 1992] William B. Frakes and Ricardo Baeza-Yates, Information Retrieval : Data Structures & Algorithms, Prentice Hall, 1992

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

[Morita & Shinoda, 1994] Masahiro Morita and Yoichi Shinoda, Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval, SIGIR '94

[Moukas, 1996] Alexandros G. Moukas, MS thesis, Amalthaea:Information Filtering and discovery in an Evolving Multiagent Ecosystem, MIT Nov, 1996

[Pazzani et al.] Michael Pazzani, Jack Muramatsu and Daniel Billsus, Syskill & Webert : Identifying Interesting Web Sites

[Sheth, 1994]Beerud Dilip Sheth, A Learning Approach to Personalized Informatioin Filtering, MS thesis, MIT, Feb 1994

[Yan, 1995] Tak W. Yan, Hector Garcia-Molina, SIFT-A Tool for Wide-Area Information Dissemination, Proceeding of the 1995 USENIX Technical Conference, 1995

[北野, 1993] 北野宏明, 遺傳的アルゴリズム, 産業圖書, 1993