유전자 알고리즘을 이용한 전자메일분류 시스템에서의 사용자선호도 추출모델링

안희국[O] 노희영
강 원 대 학 교 컴 퓨 터 과 학 과
creadelp@kwnu.kangwon.ac.kr, young@cc.kangwon.ac.kr

User Modeling in E-Mail Classification System with Genetic Algorithm

Heui-Kook Ahn[O] Hi-Young Roh
Dept. of Computer Science, Kangwon National University

요 약

본 논문에서는 전자메일을 사용자 적합도(선호도)를 기준으로 분류함에 있어 좀더 사용자 선호도를 반영할 수 있는 시스템 구조를 제안한다. 사용자 선호도는 2단계에 걸쳐서 반영되는데, 1단계에서는 사용자 관련메일로 판단된 메일정보추출어구(MIWs)들로부터 사용자 동적 시소러스(DS)의 갱신을 통해 이뤄지며, 2단계에서는 DS로부터 추출된 키워드들을 갖고 유전자 알고리즘을 작동시킬 때, 사용자선호도 feedback을 받음으로서 이뤄진다. 테스트는 kaist뉴스그룹으로부터 임의로 추출된 5개 분야 10개씩의 메일을 sample로 사용하였으며, DS로부터 추출된 키워드가 유전자알고리즘 모듈을 통해 사용자 feedback을 받았을 때, 세대가 거듭함에 따라 사용자가 요구하는 threshold 값에 근사하게 관련키워드들이 수집되었다. 그 결과 사용자 전자메일분류시스템(PECS)의 성능도 폴더정보키워드(FIWs)의 변화에 따라 향상될 수 있음을 확인하였다.

1. 서 론

전자메일은 정보의 빠른 교환이라는 장점 때문에 이미 그 사용범위가 확대되고 사용량이 증가되고 있다. 하지만, 전자메일은 아무런 여과없이 인터넷상에서 사용자에게 전달되기 때문에 수신자(recepient)는 원치않는 내용의 메일(광고성메일, spam메일, etc)을 받게되며, 이를 확인하고 분류하는 작업을 수행한다. 특히, 다량의 메일을 수신하는 사용자의 경우는 메일을 여러개의 주제(category)로 분류(classification)관리하는 작업을 수행하게되며, 이를 자동으로 걸러서(filtering) 분류해주는 시스템의 개발이 요구된다. 다량의 메일을 받아 바로 답변을 해줘야하는 기업체의 경우 또한 자동 전자메일분류기를 더욱 필요로 한다.

현재 이에 관한 연구는 문서분류(text classification)의 한 분야로서 연구가 진행되어지고 있으며, 그 대상이 단순 주제분류에서 사용자선호도분류로, off-line에서 on-line문서로 확대 이동되고 있다.[1, 4] 현재 사용자메일을 분류하기 위한 실용화는 단순히 사용자가 지정한 특정 키워드(URL, sender, word etc)의 포함유무를 갖고 filtering하는 단어검색법에 의존하고있으며, 때문에 제한된 범위에서만 분류가 이뤄지며, 변화하는 사용자 선호도를 동적으로 반영하지 못하게 된다. 문서분류에는 규칙학습에 의한 linear model들과 vector machine, Naive Bayesian분류, 신경망과 같은 알고리즘들이 적용되고 있으며, 정제된 data(문서)를 이용해 그 효율을 측정하는 데에 비중을 두고 있다. 본 논문에서는 자동적으로 사용자 메일을 분류하기 위해 기존에 제시된 동적 시소러스와 유전자 알고리즘을 사용하였으며[1], 사용자 선호도를 높이 반영하기 위해 기존에 제안된 시스템의 구조에 유전자알고리즘 모듈을 추가함으로서 기존시스템보다 높은 사용자 선호도를 반영하는 FIWs들을 추출하고자 한다.

본 논문에서는 폴더정보 키워드(FIWs)들을 추출할 때, 1차와 2차에 걸쳐 사용자 선호도를 반영할 수 있는 시스템의 구조를 제안한다. 방법으로는 사용자 관련메일로부터 추출된 키워드들(MIWs)로 사용자 시소러스를 유지 갱신함으로서 사용자선호도를 1차 반영하고, 사용자 동적 시소러스로부터 추출된 *keyword*들을 바탕으로 유전자알고리즘으로 적용하는 과정에서 사용자 feedback을 받음으로서 사용자 선호도를 2차로 반영하게 된다. 즉, 2단계에 걸친 사용자 선호도의 반영을 통해 기존의 동적 시소러스만을 이용할 때보다 더 좋은 신뢰도를 갖는 사용자 관련어구(FIWs)를 추출하게 된다. 2단계의 방법은 동적 시소러스의 특정 노드(starting node)를 중심으로 2 distance내에 있으면서 가장 거리가 짧은 노드들을 추출해 유전자 군을 형성하고, 그로부터 유전자배열(chromosome)을 형성하고, elite보존방식에 의해 가장 우수한 개체2개를 선택한다. 그리고, 나머지 개체들 중 우수한 사용자 feedback을 받은 2개의 개체를 select해, 최장거리 교배법, mutation을 통해 다음세대를 형성한다. 조건을 만족할 때까지 세대를 반복하면서 적응도가 우수한, 즉, threshold값을 만족하는 개체가 나타났을 때를 최적의 FIWs로 선택하고 수렴한다. 따라서, FIWs는 경험에 의해 사용자에게 유용한 메일로 판단된 유전자들을 갖게되며, 추가적으로 유전자알고리즘을 통해 사용자 선호도를 습득함으로서 해당 폴더(주제)와 관련된 키워드들을 유지하게 된다.

2. 관련 연구

시소러스는 용어와 용어들 사이의 관계 집합으로 구성된 일종의 용어사전이며, 동적 시소러스는 데이터가 추가되거나 삭제될 때마다 변화하는 용어들의 관계를 반영하는 시소러스를 말한다. 동적 시소러스는 용어간의 관계를 어디로부터 추출하느냐에 따라 구분할 수 있는데, 리스포터(L. K Rees-Potter)의 인용 및 동시 인용분석, 인용문맥분석을 이용하는 방법, 권쩌둥(U. Guntzer, et al)의 실

제 탐색행위에서 조합되는 용어들과 그 조합방식으로부터 이용자의 전문 지식을 추출하는 방법, 키모토(H.Kimoto)와 이와데라(T. Iwadera)의 사용자 적합 문헌으로부터 추출된 용어정보로부터 개인별 동적 시소러스를 구축하는 방법이 있다.[2] 이 세 시스템은 용어의 획득원이 각각 인용문맥, 탐색과정, 적합 문헌이라는 특징이 있으나 상황변화에 동적으로 대처하고자 하는 목적은 일치한다. 시소러스의 구성은 용어(node)와 용어가중치, 링크와 링크가중치로 구성된다. 본 논문에서는 사용자에게 선호도정보를 담기 위해 사용자적합 메일로부터 시소러스를 작성하였으며, 관련어구를 추출할 때에 기존의 유전자알고리즘을 적용하는 방법[1]보다는 결정론적 알고리즘을 사용하여 최대값을 갖는 키워드들을 추출하였다.

유전자알고리즘은 선택적 도태나 돌연변이 같은 생물진화의 원리로부터 착안된 알고리즘으로서 초기 유전자군(gene pool)로부터 형성된 후보해에 적합도함수 (fitness function)및 유전연산(mutation, selection, cross-over)을 적용함으로서 세대를 거듭하는 동안 최적의 확률을 갖는 개체만을 선택하는 확률알고리즘의 일종이다. 확률적 탐색이나 학습 그리고 최적화를 위한 한가지 기법으로 사용되고 있다.[4, 5] 본 논문에서는 시소러스로부터 추출된 최적의 키워드들과 사용자 선호도를 반영 받은 키워드들간의 조합을 통해 2단계에 걸친 사용자 선호도를 반영하게 되며, 기존의 결정론적 방법에 의해 추출된 키워드들만을 사용하면 국지 최적해(local minima)에 빠질 위험이 있으므로 이를 해결 하기위해 유전자 알고리즘의 변이(mutation)를 통해 탐색공간을 넓혀 적합도를 계산하도록 하였다.

3. PECSII(Personalized E-mail classification System II)의 구조

본 논문에서 제안하는 개별화된 전자메일 분류시스템(PECSII)의 구조는 사용자 시소러스, 유전자 알고리즘 모듈, 사용자폴더 정보추출 에이전트, 메일정보 추출 에이전트로 구성된다. 사용자 선호도를 반영하는 모듈은 사용자 시소러스와, 사용자 선호도 feedback을 포함하는 유전자 알고리즘 모듈이 되며 이를 통해 2단계에 걸친 사용자 선호도를 담을 수 있게 된다.

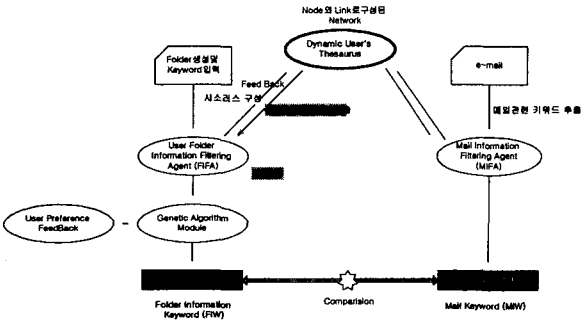

그림 1. 전체 시스템의 구조

3.1 MIW와 FIW의 구조

메일의 정보를 대표할수있는 키워드집합인 MIWs는 mail의 subject로부터 추출된 명사전부, 나머지, body로 부터 추출된 명사중 출현빈도가 가장 높은 총 16개의 keyword로 구성되며, 1차분류 적합도, 2차분류적합도 필드를 갖는다.

1차분류적합도	MIW_1	....	MIW_16	2차분류 적합도

그림 2. MIWs Data Structure

$$\text{1차분류 적합도} = \frac{\text{일치하는}MIW\text{갯수}}{\text{전체}MIW\text{갯수}} \quad (1)$$

$$\text{2차분류 적합도} = \frac{FIW\text{와일치하는}MIW\text{갯수}}{\text{전체}MIW\text{갯수}} \quad (2)$$

FIW는 사용자폴더 정보추출 에이전트가 시소러스로부터 추출한 키워드들에 대해 유전자알고리즘을 모듈을 적용시킴으로서 사용자 인터페이스에 의해 제공된 threshold값을 만족하는 키워드집합(유전자배열)으로서 16개의 키워드필드와 threshold값필드를 포함한다.

Threshold값	FIW_1	F	....	FIW_16	F

(F :사용자 선호도 False, T: 사용자 선호도 True)

그림 3. FIWs Data Structure

3.2 시소러스 구성도및 변환식 (노드및 링크가중치)

시소러스는 사용자 관련메일로부터 추출된 용어정보(word, weighted word)와 keyword들간의 관계정보(link, weighted link)를 갖는 사전으로 방향성이 없는 망구조를 하고 있다. keyword가중치는 1차 필터링을 통과한 word들이 시소러스에서 반복되어 나타나는 횟수의 누적값이다. link는 동일한 문서에서 동시에 word가 나타날 경우에 형성되며, link가중치는 시소러스에서 동시에 나타나는 횟수의 누적값이다.[2]

사용자폴더정보 추출 에이전트는 keyword가중치와 link가중치를 정량화하기 위해 다음의 1차 변환공식을 사용하며,

$$\text{1차변환}KW\text{가중치} = \frac{keyword\text{가중치}}{\text{유전자풀내의}keyword\text{가중치들의총합}} \quad (3)$$

$$\text{1차변환}link\text{가중치} = \frac{link\text{가중치}}{\text{유전자풀내의}link\text{가중치들의총합}} \quad (4)$$

starting node를 중심으로 상대적인 거리에 대한 용어관련성을 적용하기 위해 다음과 같은 2차 변환공식을 사용한다.

$$\text{2차변환}link\text{가중치} = \text{1차변환}link\text{가중치} * Cos((distance-1)*30) \quad (5)$$

$$\text{선호도가중치} = \frac{1}{\text{전체가능선호도갯수}} \quad (6)$$

이로부터 에이전트는 starting node로부터 가장 근접한 keyword 100개를 유전자 pool로 추출하여 제공하게 된다.

본 논문에서는 실험시에 사용자 선호도에 맞게 수작업으로 작성된 정적시소러스를 사용하였다.

3.3 사용자 feedback을 포함한 유전자 알고리즘

starting node를 중심으로 가장 근접한 키워드와 사용자 선호도를 갖는 keyword들을 추출하기 위해 사용한 유전자 알고리즘은 다음과 같다.

1. t세대에서 임의의 16개의 개체로 구성된 군집을 유지한다. (16개) P(t)={x1t, x2t, ,xnt}
2. 적응도 함수를 사용하여 적응도가 우수한 후보해를 선택한다. (2개) 즉, 적응도값(노드가중치합 +링크가중치합)이 가장 높은(2미만) 후보해를 선택한다.
3. 나머지 14개의 후보에 대해 사용자 선호도 feedback을 받은 2개 개체 선택 (나머지는 도태)
4. 1. point교배 시행(최대거리교배)
   *point random ---> 부모4, 자식4
5. 8개의 후보해에 대해 mutation연산 (mutation인자는 유전자군(pool)을 포함한 임의의 시소러스 노드)

6. t+1세대 후보해 배열 16개 생성
7. 적응도 계산
8. threshold값을 만족할때까지 2-5까지를 반복한다.
   결과 사용자 선호도와 시소러스로부터 가장 가까운 키워드들로 구성된 FIW추출하였다.

4. Genetic Algorithm 모듈 기능

4.1 사용자 선호도 feedback을 통한 FIW의 변화

GA모듈이 사용자에게 제공한 Interface를 통해 사용자 feedback을 적용하였을 경우, 다음과 같이 시작 노드를 중심으로 사용자 선호도를 갖는 FIWs들을 추출할 수가 있었다. test에는 "방송"이라는 키워드를 중심으로 2distance에 있는 98개의 keyword를 초기 유전자 군으로 사용하였으며, 처음 2세대만 사용자 feedback을 받았다.

* 초기 모집단(gene pool) : 98개 keyword
총 출현횟수 : 193, 총 링크횟수 : 121, 총 선호도횟수 : 98

GA를 적용유무에 따른 키워드벡터의 거리값 변화.
(Starting node : "방송")

KW	WN	TWV	WL	1' TWLV	Distance	2' TWLV	PV	TPV	VD	Rank of DS	PV	VD-GA	Rank of GA
건가	1	0.0052	2	0.0165	1	0.0165	0	0	0.0217	8	0	0.0217	12
그날	1	0.0052	2	0.0165	1	0.0165	0	0	0.0217	8	0	0.0217	12
근데	1	0.0052	3	0.0248	1	0.0248	0	0	0.0300	7	0	0.0300	8
긴급방송	7	0.0363	4	0.0331	2	0.0051	0	0	0.0414	4	0	0.0414	5
노래	3	0.0155	2	0.0165	2	0.0025	0	0	0.0181	11	0	0.0181	14
VOK	21	0.1088	5	0.0413	1	0.0413	0	0	0.1501	1	0	0.1501	1
입장	1	0.0052	1	0.0083	2	0.0013	0	0	0.0065	15	0	0.0065	16
스피커	10	0.0518	2	0.0165	2	0.0025	0	0	0.0544	3	0	0.0544	3
기속사	5	0.0259	1	0.0083	1	0.0083	0	0	0.0342	6	0	0.0342	6
생각	11	0.0570	4	0.0331	1	0.0331	0	0	0.0901	2	0	0.0901	2
						Fitness Value =			0.5775			0.6387	

그림 4. Keyword Vector거리값 변화

* KW= Keyword, WN=Weighted Node,
TWNV=Translated Weight Node Value, PV= Preference Value,
TPV=Translated Preference Value, VD=Vector Distance

본 결과는 threshold수준을 0.6으로 하였을 때, 7세대를 거쳐 얻어진 변화된 키워드 거리 벡터 값을 나타낸다. 즉, 해당 키워드들이 세대를 거듭하면서, GA의 feedback을 통해 벡터거리의 변화를 얻게 되었다.

또한 이러한 변화를 통해 FIWs의 변화도 다음과 같이 이뤄졌다. 이러한 변화는 유전자알고리즘의 feedback을 통해 사용자 선호도를 받은 개체는 선택되고, 상대적으로 향상되지 못한 개체는 도태됨으로서 얻어지게 된다.

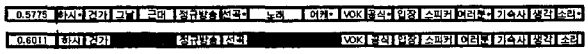

그림 5. DS(上)와 GA(下)를 통해 추출한 FIWs들의 변화

즉, 세대를 거듭하면서, 사용자 선호도 feedback을 주었을 때의 FIWs구성원소들이 적응도가 향상되는 쪽으로 조합이 이뤄짐을 확인하였다.

4.2 GA의 세대에 따른 FIWs의 적응도값 변화

수렴시기를 0.6으로 하였을 때, 적응도를 만족하는 FIWs의 조합이 GA의 7세대에서 얻어졌으며 탐색범위의 확대 및 사용자선호도의 조합으로 통해 기존의 DS만을 사용했을 때보다 높은 적응도를 갖는 FIWs들을 추출하였다.

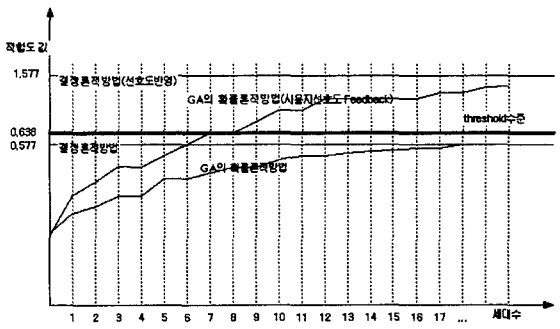

그림 6. 세대에 따른 FIWs의 적응도값 변화

결과적으로 사용자 선호도를 반영할 수 있는 키워드들을 많이 포함한 새로운 개체가 선택되었으며 전체 사용자 메일분류 시스템에 적용시에도 더 높은 신뢰도를 갖을 수 있다.

5. 결론 및 향후 연구과제

본 논문에서는 전자메일을 사용자 선호도에 따라 분류하기 위한 시스템모델[PECS참조논문]에서 동적 시소러스와 유전자알고리즘모듈을 독립적으로 적용함으로서 동적 시소러스만을 이용했을 때 보다 높은 사용자선호도를 갖는 관련어구를 추출할 수 있었다. 즉, 동적 시소러스로부터 추출된 keyword들에 대해 사용자 선호도를 포함하는 유전자알고리즘을 적용시킴으로서 2차에 걸쳐 사용자 선호도를 담아내고 추출함으로서 PECS의 사용자 관련메일분류에 신뢰도를 높일 수 있음을 확인할 수 있었다. 하지만, 본 논문에서는 기본적인 유전자 알고리즘을 적용하였고, 향후 좀더 다양한 유전연산의 적용 및 초기 모집단의 범위에 따른 관련어구 추출의 관계 및 효율성에 관한 연구가 필요하다. 또한 본 논문에서 제안하는 시스템이 사용자 선호도에 따라 전자메일을 분류하는데 있어서의 효과를 실제 사용자 test를 통하여 검증하는 연구가 필요하다.

참고문헌

[1] H.-K. Ahn and H. -Y. Rho, "Personalized E-Mail Classification System Using Dynamic Thesaurus and Genetic Algorithm", Proc. Korea Information Socience Society, 2002 (In Korean)

[2]H. Kimoto , T. Iwadera. "Construction of a dynamic Thesaurus and its use for associated information retrieval" , Proceedings of the thirteenth international conference on Research and development in information retrieval December 1989

[3] Michael J. Pazzani , "Representation of electronic mail filtering profiles", Proceedings of the 2000 international conference on Intelligent user interfaces January 2000

[4] S.-Y. Kim and S.-B Cho, "User modeling in meta-search engine with genetic algorithm," Proc. Korea Information Science Society, 2000 (In Korean)

[5] 조유근 외, [알고리즘] 이한출판사 2000년