동적 시소러스와 GA을 이용한 개별화된 E-Mail 분류시스템(PECS)

안희국[0] 노희영
강 원 대 학 교 컴 퓨 터 과 학 과
creadelp@mail.kangwon.ac.kr, young@cc.kangwon.ac.kr

Personalized E-Mail Classification System Using Dynamic Thesaurus and Genetic Algorithm

Heui-Kook Ahn[0] Hi-Young Roh
Dept. of Computer Science, Kangwon National University

요 약

본 논문에서는 전자메일을 사용자 적합도(선호도)를 기준으로 분류하기 위한 구조를 제안한다. 분류는 1차 분류와 2차 분류로 나눠지는데, 1차 분류에서는 사용자 적합도를 판단하기 위해 사용자 관련 정보로부터 동적 시소러스를 구축하고, 구축된 시소러스와의 비교를 통해 사용자에게 유용한 메일인지 아닌지를 결정하고, 2차 분류에서는 사용자가 지정한 폴더키워드를 중심으로 사용자 시소러스로부터 유전자 알고리즘을 이용해 추출한 키워드들과의 적합도 비교를 통해서 특정 폴더로의 분류가 이뤄지게 된다. 테스트에는 메일 정보값(*Mail Information Word*)을 추출하기 위해 HAM(Hangul Analysys Module)을 포함하는 메일정보추출 에이전트를 사용하였고, mail의 subject와 본문(body)로부터 추출된 16개의 word정보와 시소러스 적합도 정보, 분류 적합도 정보를 하나의 데이터구조로 사용하였다. 이러한 통합된 시스템 구조와 data structure를 이용해 mail을 사용자의 선호도에 따라, 1차와 2차에 걸친 분류시 분류가 사용자 선호도에 근접하게 이루어 질 수 있음을 확인하였다.

1. 서 론

전자메일은 사용자들간의 필수적인 정보의 빠른 교환을 가능하게 한다. 하지만, 전자메일은 사용자(recipient)가 원하지 않는 광고성메일, spam mail등의 관심분야와 동떨어진 mail들을 포함하게 된다. 따라서, 사용자는 이러한 메일들을 확인하고 분류하는 작업들을 수행하게 된다. 메일의 사용이 급증함에 따라 소량의 메일을 주고받는 일반 개인 사용자뿐만 아니라, 다량의 메일을 수신하는 사용자는 메일을 받아서 원하는 폴더로 분류(classification)하고 관리하는 추가적인 작업의 필요하게 된다. 따라서, 사용자에게 도착한 메일을 선호도에 따라 자동으로 filtering하고, 분류해주는 시스템의 개발이 필요하다. 특히, 수신 후에 바로 답변을 요구하는 mail의 경우는 자동으로 메일을 분류해주는 시스템의 필요성이 더욱 요구된다.

현재 메일을 분류하기 위해 여러 분야에서 연구가 진행되고 있으며, 이 분야는 특히 문서 필터링(text filtering)의 한 분야로서 연구가 활발히 진행되고 있다.[2][3][4] 현재 사용자의 선호도에 따라 mail을 분류하는 시스템의 상용화는 사용자가 지정한 특정 단어의 포함유무를 이용해 mail을 분류하거나(ex. hotmail.com) 보낸 사람의 이름을 이용해 mail을 분류하는(ex. hanmail.net) 단계까지 진행되어 있으며, 그로 인해 정보필터링 시에 제한된 범위내에서만 필터링이 이뤄지게 된다. 따라서, 사용자 선호도를 나타낼 수 있는 좀더 다양하고 광범위한 관련어구를 자동적으로 추출해 내고 갱신하며 분류할 수 있는 시스템의 요구가 필요하다.

본 논문에서는 개인사용자의 선호도를 중심으로 메일을 분류할 때, 1차와 2차 필터링 시에 통합적으로 사용될 수 있는 시스템의 구조를 제안하고자 한다. 방법으로는 1차 필터링과 2차 필터링을 위해서 동적으로 사용자 시소러스를 구성하고, 이로부터 유전자 알고리즘을 사용해 사용자의 선호도와 가장 근접한 word들을 추출하게 된다. 즉, 사용자 시소러스라는 구조에 사용자의 관심 word들을 담고, 그로부터 사용자에게 도착한 메일이 사용자에게 유용한 메일인지 아닌지를 구분해 내는 1차 필터링의 과정을 거치게 된다. 1차 필터링을 통과한 메일정보는 시소러스로부터 유전자알고리즘을 통해 추출된 사용자 정보 word(UIW)와 비교를 통해 2차 filtering (classification)을 하게 된다. 이러한 구조는 메일과 사용자와의 적합도를 판단하며, 사용자가 원하는 내용을 따로 구분시켜 둔 폴더로 메일을 2차 분류하는데 있어서 동일한 구조를 제공하게 되며, 2차 분류가 완성되었을 시에 추가된 word들을 시소러스에 추가시킴으로서 시소러스가 갖고있는 정보값들(노드, 노드가중치, 링크, 링크가중치)을 동적으로 갱신시켜 준다. 따라서, 사용자는 초기에 시소러스작성에 필요한 keyword 하나만을 제공하면, 그로부터 사용자에게 유용한 정보들을 추출해내고, 추출된 정보는 mail filtering시에 사용된다. 또한 갱신된 시소러스로부터 사용자가 지정한 폴더의 키워드들을 중심으로 FIW를 자동적으로 갱신시킴으로서 최적의 사용자 정보를 폴더들이 소유하게 된다.

2. 관련 연구

동적 시소러스를 구성하는 방법에 관한 연구는 용어간의 관계를 어디로부터 추출하는가에 따라 구분할 수 있는데, 리스포터 (L.K. Rees-Potter)의 인용 및 동시 인용분석, 인용문맥분석을 이용하는 방법과 귄쩌 등(U. Güntzer, et al)의 실제 탐색행위에서 조합되는 용어들과 그 조합방식으로부터 이용자 전문지식을 추출하는 방법 키모토(H. Kimoto)와 이데와라(T. Iwadera)의 사용자 적합문헌으로부터 추출한 용어정보로부터 개인별 동적시소러스를 구축하는 방법이 있다.[1] 이 세가지 시스템은 용어의 획득원이 각각 인용문맥, 탐색과정, 적합문헌이라는 특징이 있으나 상황변화에 동적으로 대처하고자 하는 목적은 일치한다. 이중에서 키모토와 이데와라의 방법은 동적시소러스의 구성을 사용자의 관심있는 관련문서로 작성함으로서 그로부터 사용자의 개별적인 관심사를 반영하는 연관

키워드들을 생성할 수 있었다. 하지만 이러한 방법의 경우는 링크로 연결된 노드들 중에서 threshold값을 기준으로 그 이상의 노드 가중치를 갖는 키워드들을 추출하게 되므로, 추출되는 키워드들의 개수가 가변적이다. 또한 하나의 문서에서 동시에 노드가 나타날 경우 생성되는 노드가중치는 관련 키워드 추출 시에 직접적인 영향을 주지 못하게 된다. 따라서, 본 논문에서는 유전자 알고리즘을 통해 확률적으로 Folder Keyword(FK)와 가장 근사한 값을 갖는 일정한 개수의 keyword들을 추출함으로서 좀더 사용자 선호도와 의미적으로 가까운 FIW들을 추출하게 된다.

유전자 알고리즘은 1970년대 중반에 Holland에 의해 체계화된 것으로서 진화프로그래밍의 골격이 되며, 초기 후보해로부터 세대를 거듭하는 동안 적응도가 우수한 후보해를 선택하고 이를 유전연산 (crossover, mutation)을 통해 적응도가 우수한 개체만을 선택함으로서 최적의 확률을 갖는 개체만을 선택하는 확률알고리즘의 일종이다.[2][5] 본 논문에서는 새로운 MIW가 시소러스에 추가될 때마다 갱신되는 시소러스내의 무수한 노드들과 노드가중치, 노드들간의 링크와 링크 가중치 값을 바탕으로 최적의 근사값을 갖는 키워드들을 추출하는 것이 이론적으로 상당히 많은 비용(시간과 처리과정)을 소요하므로 유전자 알고리즘을 이용하여 관련 키워드들을 추출하게 된다.

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

본 논문에서 제안하는 개별화된 전자메일 분류시스템(PECS)의 *framework*는 자동으로 갱신되는 사용자 시소러스와 메일의 정보(MIW)를 추출하는 메일정보추출 에이전트, 폴더키워드를 중심으로 관련키워드들(FIW)을 추출하는 사용자폴더 정보추출에이전트로 구성되며, 이들간의 정보교환을 통해 사용자에게 필요한 메일들을 다단계로 분류하게 된다.

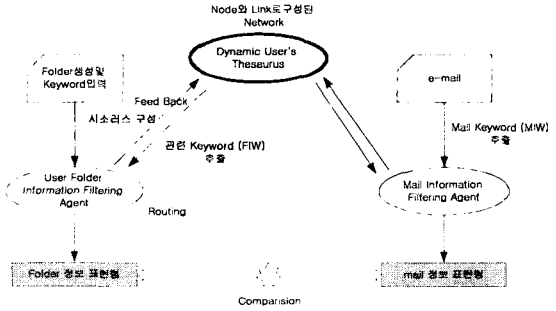


그림 1. 전체 시스템의 구조

3.1 MIW와 FIW의 구조

*MIWs(Mail Information Words)*는 메일의 내용을 대표할 수 있는 keyword들의 집합으로, 본 논문에서는 HAM(Hangul Analysis Module)을 포함하는 MIFA(Mail Information Filtering Agent)가 그 역할을 하게 된다. HAM 처리과정에서는 e-mail문서로부터 어절분리, 기존용어처리, 불용어 제거, 조사어미 제거, 접사 제거, 선어말어미 제거, 경동사의 활용형일 경우 제거, 명사등록, 복합명사처리, 약어처리의 과정을 거쳐, 명사를 추출한다. MIFA는 추출된 명사들 중에서 subject로부터 2개, body로부터 14개의 명사를 추출한다. 이때, body로부터 추출된 명사가 14개가 넘을 경우는 출현빈도가 높은 순으로 가중치를 두어 우선 추출하고, 나머지 출현빈도가 낮은(1회) 명사들은 random방법으로 추출하여 data structure에 넣는다. datastructure의 앞과 뒤에는 사용자와의 적합도 정보를 갖는 시소러스 적합도 필드와 폴더와의 적합도 정보를 갖는 분류적합도 필드를 갖는다.

FIWs(Folder Information Words)는 사용자 폴더정보 추출에이전트가 폴더에게 지정된 keyword를 가지고 사용자 시소러스로부터 추출한 관련word들의 집합이다. datastructure는 16개의 관련word와 추출된 word들이 keyword를 중심으로 관련정도(적응도)를 나타내는 정보값을 갖는 키워드적응도 필드를 포함한다.

분류적합도	MIW_1	····	MIW_16	시소러스적합도

키워드적응도	FIW_1	····	FIW_16

3.2 사용자 시소러스

사용자 시소러스는 개인 사용자의 관심분야와 관련된 용어정보를 갖고 있는 사전으로서 방향성이 없는 망구조를 하고 있으며, 구성은 keyword정보와 keyword관계정보로서 구성된다. keyword정보는 다시 keyword와 keyword가중치로, keyword관계정보는 link와 link가중치로 구성된다. keyword가중치는 1차 필터링을 통과한 word들이 시소러스에서 반복되어 나타나는 횟수의 누적값이며, link는 동일한 문서에서 동시에 word가 나타날 경우에 만들어지며, link가중치는 link가 시소러스에서 반복되어 나타나는 횟수의 누적값이다.

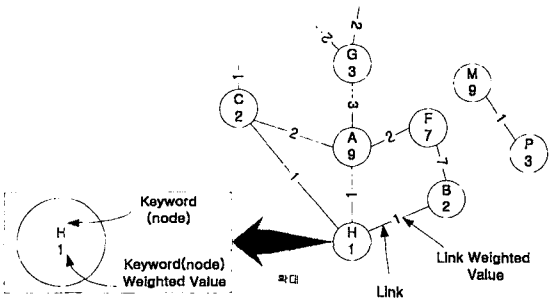


그림 2. 시소러스 구성도

동적으로 시소러스의 노드링크를 생성하는 알고리즘은 1차필터링을 통과한 MIW들이 기존의 사용자 시소러스에서 링크되어 있지 않다면, 다음과 같이 노드를 생성한다.[1]

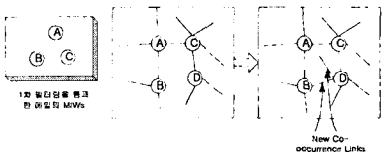


그림 3. 링크생성 알고리즘

본 논문에서는 실험시에 사용자 선호도에 맞게 수작업으로 작성된 정적시소러스를 사용하였다.

3.3 유전자 알고리즘을 이용한 FIW의 추출

starting node를 중심으로 가장 적응도가 높은 링크가중치를 갖는 keyword들을 추출하기 위해 본 논문에서는 다음과 같은 유전자 알고리즘을 사용하였다.

1. starting node를 중심으로 3 level내에 있는 keywords들을 추출해, 그들을 gene pool로 사용한다.
2. t세대에서 임의의 16개의 개체로 구성된 군집을 유지한다.
   $P(t)=\{x_1^t, x_2^t, ,x_n^t\}$
3. 적응도 함수를 사용하여 적응도가 우수한 후보해를 선택한다

즉, 적응도값(노드가중치합 + 링크가중치합)이 가장 높은 (2미만) 후보해를 선택한다.
4. 유전연산(crossover, mutation)을 통하여 t+1세대를 만든다.
5. 적응도를 계산한다.
6. 조건이 만족할때까지 2-5까지를 반복한다.

적응도를 계산 할 때는 먼저, keyword가중치와 link가중치들을 정량화하기 위해 각각 1값을 기준으로 변환한다. 변환공식은 아래와 같다.

$$\text{변환}keyword\text{가중치} = \frac{keyword\text{가중치}}{\text{유전자풀내에있는모든}keyword\text{가중치들의총합}}$$

$$1\text{차변환}link\text{가중치} = \frac{link\text{가중치}}{\text{유전자풀내에있는모든}link\text{가중치들의총합}}$$

이때, starting node를 중심으로 distance가 멀어질수록 용어관련성이 멀어지므로 각 변환 link가중치에 대해 y=cos(x)함수를 적용하여 관련도를 수정하였다. x는 0과 90사이에 있으며, 편의상 6level을 초과하는 링크는 관련성이 없는 것으로 제한하였다. 즉, 시작노드로부터 level에 따른 변환 link값은 다음과 같다.

$$1\text{차변환}link\text{가중치} = 1\text{차변환}link\text{가중치} * Cos((level\text{수} - 1) * 15)$$

4. PECS의 기능

4.1 1차 필터링

MIF와 시소러스 구성 word간의 비교와 threshold값의 변화를 통해 사용자에게 적합한 메일인지, 아닌지를 판단한다.

MIW

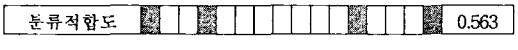
시소러스 구성 keyword와의 적합도

$$\frac{\text{일치하는}MIW\text{갯수}}{\text{전체}MIW\text{갯수}} = \frac{9}{16} = 0.563$$

threshold값을 0.5로 하였을 때, 본 메일은 threshold값을 만족하므로 사용자에게 적합한 메일로 분류된다.

4.2 2차 필터링

MIF와 FIW들간의 비교와 threshold값의 변화를 통해 해당 폴더들간의 유사도를 판단한다.



MIW

FIW

폴더 keyword들과의 적합도(분류적합도)

$$\frac{FIW\text{와일치하는}MIW\text{갯수}}{\text{전체}MIW\text{갯수}} = \frac{4}{16} = 0.25$$

ex) Folder 1 =0.063
Folder 2 =0.250
Folder 3 =0.188

예에서, threshold값을 0.15라고 하였을 때, 역치값을 만족하는 Folder2와 Folder3중에서 가장 유사한 폴더 적합도를 갖는 Folder 2로 direction을 한다.

4.3 시소러스 및 FIW 갱신

2차필터링이 이뤄지면, FIW를 사용자 시소러스에 추가한다. 방법은 그림 3.과 같은 방법으로 진행하며, 관련 가중치들을 갱신한다. 또한 갱신된 시소러스로부터 3.3절에서 제시한 유전자 알고리즘을 통해 FIW들을 다시 추출한다.

4.4 FIW의 갱신결과

-초기 FIW data

-갱신된 FIW data모습

따라서, 동일한 threshold값을 유지시켰을 때, 시소러스 갱신 후에 1차필터링에서 사용자에게 유용하지 않던 것으로 판명되었던 메일이 유용한 것으로 판단되며, 2차 필터링시에 folder와의 적합도에서 적합하지 않던 것이 적합한 것으로, 적합한 것이 적합하지 않은 것으로 분류가 된다. 즉, 사용자 관련 어구의 미세한 변화로 인해 메일 필터링의 변화가 유도된다.

5. 결론 및 향후 연구과제

본 논문은 개인사용자의 전자메일을 분류하기 위해 동적으로 갱신이 가능한 사용자 시소러스와 그로부터 유전자 알고리즘을 통해 관련어구를 추출하는 과정을 개별화된 메일분류시스템 모델로서 제안한다. 이렇게 추출된 MIW정보는 사용자 시소러스와의 적합도를 통해 forward/discard를 결정하고, FIW정보와의 적합도를 통해 어느 폴더로 가야할지를 결정한다. 즉, 사용자 동적 시소러스와 유전자 알고리즘을 이용한 키워드정보추출을 통해 사용자 적합여부에 따라 다단계의 메일분류가 이뤄짐을 확인하였다. 하지만, 본 논문에서 제시한 PECS은 메일정보 필터링 에이전트가 얼마나 메일의 정보를 대표할 수 있는 word를 잘 추출해 낼 수 있느냐에 그 정확도가 많이 좌우된다. 향후에 mail을 좀더 사용자 선호도와 근사하도록 분류가 이루어지게 하려면, 정보추출 Agent의 성능향상을 꾀하는 연구가 진행되어야 한다. 즉, 정확한 어휘분석기, 구분분석기 등을 통해 정확한 MIF를 추출하는 기법의 연구가 필요하다. 또한 사용자 관련메일이 증가함에 따라 사용자 시소러스가 무한히 증가하게 된다. 그 결과 시소러스로부터 정보를 추출하는데 많은 자원(기억공간, 처리시간)이 필요하게 된다. 따라서, 처리의 효율을 증대시키기 위해 동적으로 node를 삭제하고 링크를 삭제하는 방법을 이용함으로서 일정수준의 노드를 갖는 사용자 시소러스를 구축하는 방법이 연구되어야 할 부분이다. 추가로 시소러스로부터 사용자 관련어구들을 좀더 정확하게 추출하기 위해 시소러스의 구성에 관한 연구가 필요하다.

참고문헌

[1] 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
[2] Michael J. Pazzani , "Representation of electronic mail filtering profiles", Proceedings of the 2000 international conference on Intelligent user interfaces January 2000
[3] S.-Y. Kim and S.-B Cho, "User modeling in meta-search engine with genetic algorithm," Proc. Korea Information Science Society, 2000 (In Korean)
[4] J. Ryu and S.-B. Cho, "Automatic categorization of real world FAQs using hierarchical document clustering," Proc. Korea Fuzzy and Intelligent Systems, Seoul, May 2001. (In Korean)
[5] 조유근 외, [알고리즘] 이한출판사 2000년