연역적 유전자 알고리즘을 이용한 연관 단어 지식베이스의 최적화

(Optimization of Associative Word Knowledge Base using Apriori-Genetic Algorithm)

고 수 정 † 최 준 혁 †† 이 정 현 †††

(Soo-Jung Ko) (JunHyeog Choi) (Jung-Hyen Lee)

요 약 지식 기반 정보검색 시스템에서의 질의 확장은 단어간의 의미 관계를 고려한 지식베이스를 필요로 한다. 기존의 단순 마이닝 기법은 사용자의 선호도를 고려하지 않은 채 연관 단어를 추출하므로 재현율은 향상되나 정확도는 저하된다. 본 논문에서는 단어간의 의미 관계를 고려한 연관 단어 중에서 사용자가 선호하는 연관 단어만을 포함하는 정확도가 향상된 최적화된 연관 단어 지식베이스 구축을 위한 방법을 제안한다. 이를 위해 컴퓨터 분야의 웹문서를 8개의 클래스로 분류하고, 각 클래스별 웹문서에서 명사를 추출한다. 추출된 명사를 대상으로 Apriori 알고리즘을 이용하여 연관 단어를 추출하고, 유전자 알고리즘을 이용하여 사용자가 선호하지 않는 연관 단어를 지식베이스의 구축 대상에서 제외시킨다. 본 논문에서 제안된 Apriori 알고리즘과 유전자 알고리즘의 성능을 평가하기 위하여 Apriori 알고리즘은 상호정보량과 Rocchio 알고리즘과 비교하며, 유전자 알고리즘은 TF·IDF를 이용한 단어 정제 방법과 비교한다.

Abstract The query expansion in the KBQP(Knowledge Based Query Processor) needs a knowledge base being considered semantic relation among words. Because established simple mining technique extracts the association words without considering user preference, it shows higher recall but lower accuracy. In this paper, we propose a method to construct optimized association word knowledge base that improves the accuracy of categorization and includes only the association words of user preference being considered semantic relation among words. For the purpose, web documents on the field of computer are categorized into 8 classes and nouns are extracted from each classified web document. Association words are mined from these nouns by Apriori algorithm and association words which user doesn't prefer are pruned by genetic algorithm. For the purpose of evaluating the performance of Apriori and Genetic algorithm designed in this paper, Apriori algorithm is compared with Rocchio algorithm and mutual information, and genetic algorithm is compared with word refining method using TF·IDF.

1. 서 론

지식 기반 정보검색(Knowledge Based Query Processor) 시스템에서는 질의어와 완전히 일치하는 색인어가 존재하지 않을 경우, 단어 분류 지식베이스를 이용한 질의어 확장을 통하여 검색을 수행한다[9]. 기존의 단어 분류 지식베이스는 전문가에 의해 수작업으로 구축하는 방법[10]과 말뭉치를 대상으로 단어간 공기정보를 이용한 연관 단어 군집에 의한 방법이 있다[4,7]. 이들 방법에 의한 단어 분류 지식베이스 구축은 많은 시간과 노력이 요구되며, 동의어나 상위 개념어 혹은 하위 개념어 위주로 구축되어 단어간의 의미 관계를 정확히 반영하지 못하는 단점이 있다. 이를 해결하기 위해서 단어간의 의미 관계를 고려하여 웹문서로부터 연관 단어를 추출하는 방법을 사용한다[3,15]. 추출된 연관 단어를 이용하여 사용자 질의어를 확장할 경우 재현율은 향상되나

† 정 회 원 : 유니버셜 소프트 정보통신(주) 연구원
sujung@nisun.inha.ac.kr
†† 종신회원 : 김포대학 컴퓨터계열 교수
jhchoi@kimpo.ac.kr
††† 종신회원 : 인하대학교 전자계산공학과 교수
jhlee@inha.ac.kr
논문접수 : 2000년 11월 17일
심사완료 : 2001년 6월 4일

정확도는 저하되는 문제점이 발생한다. 이러한 문제점를 해결하기 위해서는 지식베이스를 최적화하여야 한다. 지식베이스를 최적화하기 위해서 사례 기반이나 사용자의 연관 피드백을 이용하는 방법이 있으나 이러한 방법들은 기존의 사례가 필요하거나 사용자의 수작업을 필요로 한다[11,13].

본 논문에서는 컴퓨터 분야의 웹문서를 대상으로 각 문서를 8개의 클래스로 분류하였고 Apriori 알고리즘을 사용하여 연관 단어를 추출한 후 최적화를 목적으로 하는 유전자 알고리즘[12]을 적용한다. 유전자 알고리즘은 마이닝 기법에 의해 추출된 연관 단어 중에서 사용자가 선호하지 않는 연관 단어를 지식베이스 구축대상에서 제외시킴으로써 지식베이스를 최적화한다.

2. 최적화된 연관 단어 지식베이스 설계

그림 1은 Apriori-Genetic 알고리즘을 이용하여 연관 단어를 추출하기 위한 전체 시스템이다. 그림 1의 시스템은 웹문서 수집기(HTTP Down Loader), 텍스트 엔진(Text Engine), 단어 연관 규칙 마이닝, 단어 연관 규칙 필터링, 연관 단어 집합(Word Rule Set)으로 구성된다.

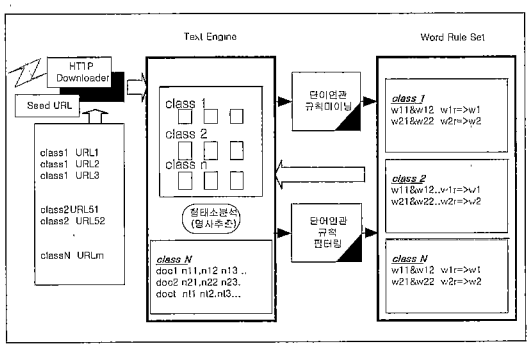


그림 1 연관 단어 지식베이스 구현 시스템

웹문서 수집기는 각 클래스별로 웹문서를 수집한다. 각 클래스는 {class1,class2,...,classN}으로 표현한다. 텍스트 엔진은 각 클래스에 수집된 웹문서를 형태소 분석한 후 명사만을 추출하는 역할을 한다. 텍스트 엔진에서 각 클래스에 포함된 문서는 {doc1,doc2,...,doct}로 표현하며, 각 문서로부터 추출된 명사는 {n11,n12,n13,...}의 형태로 표현한다. 연관 단어 집합은 {class1,class2,...,classN}의 각 클래스에 마이닝된 연관 단어를 의미한다. 연관 단어는 {w11&w12,..,&w1r=>w1, w21&w22,...,&w2r=>w2,...}의 형태로 표현한다. 연관 단어에 포함된 {w11,w12,w14,w21,w22,w2r}의 각 구성요소는 연관 단어를 구성하는 단어를 의미하며, "&"의 의미는 단어와 단어가 연관되었음을 나타내는 기호이다. 또한, {w1,w2,...}는 연관 단어를 대표하는 단어이다. 이러한 단어는 각 웹문서로부터 추출된 명사들을 기반으로 한다. 그러나 {n11,...}과 {w11,...}과 같이 다르게 표현한 이유는 {w11,...}이 연관 단어에 속한 단어임을 나타내기 위해서이다.

그림 1을 구현하기 위해 그림 2의 흐름도를 제시하였다. 그림 2는 웹문서에서 연관 단어를 추출하여 지식베이스를 구성하는 단계와 지식베이스가 최적화되도록 연관 단어를 정제하는 단계로 분류된다. 그림 2의 Block1과 Block2는 지식베이스를 구성하는 단계이며, Block3과 Block4는 지식베이스를 최적화하는 단계이다. 각 Block 안의 *표1*~*표8*은 본 논문의 4장에서 기술한 예를 나타낸다.

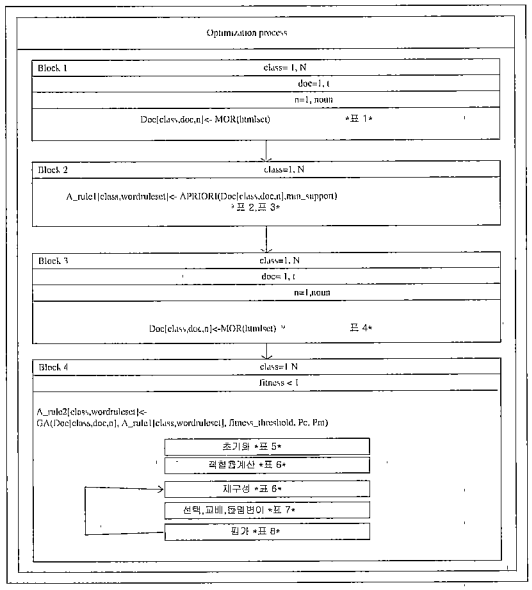


그림 2 연관 단어 지식베이스의 구축 및 최적화 흐름도

Block1의 단계에서는 형태소 분석을 이용하여 웹문서(htmlset)로부터 명사를 추출하는 전처리 과정을 수행한다. 이러한 전처리 과정을 MOR( )이라는 함수로 표현한다. Doc[class,doc,n]는 각 클래스에 수집된 문서를 형태소 분석한 후 명사만을 추출한 결과를 저장한 배열이다. Doc[class,doc,n]의 class는 {class1,...,classN}의 클래스를 나타낸다. doc는 {doc1,..,doct}의 문서를 나타

내며 {class1,..,classN}의 각 클래스로 수집된 문서를 의미한다. n은 {n11,n12,...,ndocnoun}를 나타내며, {doc1,...,doct}의 각 문서에서 추출된 명사를 의미한다.

Block2의 단계에서는 추출된 명사로부터 연관 단어를 마이닝한다. 이 단계에서의 A_rule1[class,wordruleset]는 각 클래스별로 마이닝된 연관 단어 집합을 의미한다. 여기서, class는 {class1,...,classN}의 각 클래스를 나타내고, wordruleset은 {w11&w12,..,&w1r=>w1,w21&w22,...,&w2r=>w2,...}의 형태로 표현할 수 있다.

APRIORI( ) 함수는 연관 단어를 마이닝하기 위한 Apriori 알고리즘을 나타내고, min_support는 Apriori 알고리즘에 사용되는 최소 지지도를 나타낸다.

Block3의 단계에서는 Block1의 단계에서와 같이 형태소 분석을 이용하여 웹문서(htmlset)로부터 명사를 추출하는 전처리 과정을 하게 된다. 그러나 Block1과의 차이점은 Block1에서는 사용자의 선호도를 고려하지 않은 웹문서로부터 명사를 추출하며, Block3에서는 사용자가 선호하는 웹문서로부터 명사를 추출하는 차이점을 갖는다.

Block4의 단계에서는 Block3에서 추출된 명사를 이용하여 Block2에서 마이닝된 연관 단어 A_rule1[class,wordruleset]에서 사용자가 선호하지 않는 연관 단어는 지식베이스에서 제거한다. GA( )는 사용자가 선호하지 않는 연관 단어를 지식베이스에서 제거하기 위한 유전자 알고리즘을 의미하며, fitness_threshold와 Pc, Pm은 유전자 알고리즘에서 사용되는 적합율 한계값과 교배율, 돌연변이율을 의미한다. A_rule2[class, wordruleset]는 사용자가 선호하지 않는 연관 단어를 제거한 최적의 연관 단어를 저장하는 배열을 의미한다.

3. 연역적 유전자 알고리즘

3.1 Apriori 알고리즘을 이용한 연관 단어의 추출

단어간의 연관 규칙을 마이닝하는 Apriori 알고리즘[1,2]은 두 단계를 통하여 연관 규칙을 추출한다. 첫 번째 단계에서는 최소 지지도(min_support) 이상의 발생 지지도(transaction support)를 가지는 조합을 찾아 고빈도 단어 항목을 구성한다. 두 번째 단계에서는 데이타베이스로부터 연관 규칙을 생성하기 위하여 고빈도 단어 항목을 사용한다. 모든 고빈도 단어 항목(L)에 대해서 고빈도 단어 항목의 모든 공집합이 아닌 부분집합들을 찾는다. 각각의 부분집합(A)에 대하여, 만약 support(A)에 대한 support(L)의 비율이 적어도 최소 신뢰도(min_confidence) 이상이면, A->(L-A)의 형태의 규칙을 출력한다. 이 규칙의 지지도는 support(L)이고, 신뢰도는 support(L)/support(A)이다.

연관 단어 쌍을 구성하기 위해서는 신뢰도(confidence)와 지지도(support)를 결정해야 한다. 신뢰도를 결정하기 위한 식(1)은 다음과 같이 구해진다. 식(1)은 단어 W1과 W2의 모든 항목을 포함하고 있는 트랜잭션의 수를 단어 W1의 항목을 포함하고 있는 트랜잭션의 수로 나눈 결과 값을 나타낸다.

$$Confidence(W1->W2)=Pr(W2|W1) \quad (1)$$

그림 3은 100개의 웹문서를 대상으로 신뢰도를 다양하게 변화시켰을 때, 추출된 연관 단어에 대한 정확도와 재현율을 나타낸다. 100개의 웹문서는 본 논문이 실험을 위해 컴퓨터 분야의 웹문서를 8개의 클래스로 분류한 클래스 중에서 게임 클래스에 수집된 웹문서이다. 웹문서 수집은 웹문서 수집기를 이용한다. 마이닝된 결과에 대해 재현율과 정확도를 평가하는 기준은 영어 단어에 대한 시소러스인 WordNet[9]을 사용하여 평가하였다. 평가를 위해WordNet에서 게임과 관련된 영어 단어의 동의어, 상의어, 하의어를 추출하였다. 추출된 단어를 한글로 번역하여 300개의 연관 단어를 구성하였다. 마이닝된 연관 단어가 이들 300개의 연관 단어에 포함되지 않을 경우 오류로 처리하였다. 정확도는 마이닝된 연관 단어 중에서 오류로 처리된 연관 단어의 비율을 나타낸다. 재현율은 마이닝된 연관 단어가 평가를 위해 구성된 연관 단어에 포함된 비율이다.

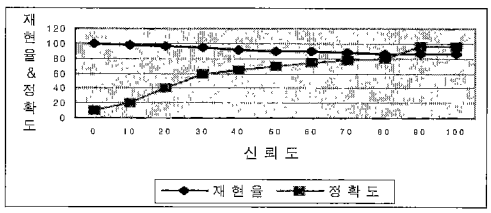


그림 3 신뢰도의 변화에 따른 재현율과 정확도

그림 3은 신뢰도가 클수록 추출되는 연관 단어의 정확도는 높아지나 재현율은 낮아짐을 나타낸다. 그러나 85이상의 신뢰도에서는 재현율이 거의 일정하고 정확도는 높은 수치를 나타낸다. 따라서 가장 적합한 연관 단어를 추출하기 위해서는 신뢰도를 85이상으로 지정해야 한다.

지지도를 결정하기 위한 식(2)는 전체 단어들의 쌍 중에 각 연관 단어의 출현 빈도를 나타낸다. 식 (2)는 단어 W1과 W2의 모든 항목을 포함하고 있는 트랜잭션의 수를 데이타베이스 내의 전체 트랜잭션의 수로 나눈 결과 값을 나타낸다.

$$Support(W1->W2)=Pr(W1 \cup W2) \quad (2)$$

지지도가 크다면 빈도수는 작으나 중요한 연관 단어가 생략될 수 있고, {기본&방식&이용&지정=>실행}와 같이 빈도는 높지만 중요하지 않은 연관 단어가 추출된다. 그림 4는 100개의 웹문서를 대상으로 지지도를 다양하게 변경시킴에 따른 정확도와 재현율의 변화를 나타낸다. 정확도와 재현율의 측정 기준은 신뢰도와 같다.

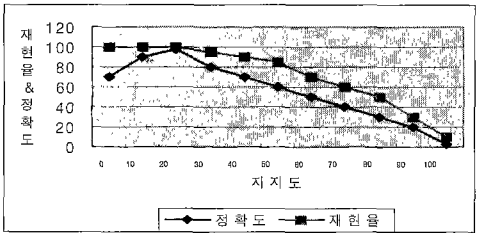


그림 4 지지도의 변화에 따른 재현율과 정확도

정확도와 재현율의 곡선이 일치하는 지점은 지지도가 22인 경우로, 이 지점에서 가장 적합한 연관 단어가 추출된다. 그러나, 지지도가 22이상인 경우에는 정확도와 재현율이 모두 낮아진다. 따라서 가장 신뢰할 만한 연관 단어를 추출하기 위해서는 22이하의 지지도를 지정해야 한다. 그러나 지지도를 0으로 한다면 클래스와 전혀 관계 없는 문서에서 연관 단어가 추출되므로 지지도는 0보다 크도록 설정하여야 한다.

3.2 유전자 알고리즘을 이용한 연관 단어 정제

유전자 알고리즘은 마이닝 기법에 의해 추출된 연관 단어 중에서 사용자가 선호하지 않는 연관 단어를 지식베이스 구축대상에서 제외시킴으로써 지식베이스를 최적화한다. 사용자가 선호하는 연관 단어는 사용자가 선호하는 웹문서를 기준으로 한다. 이러한 경우, 사용자가 선호하는 웹문서는 정보검색 엔진에서 사용자들의 검색 횟수가 높은 문서이다. 마이닝 기법에 의해 추출된 연관 단어 중에서 사용자가 선호하는 웹문서에 나타나는 빈도가 낮은 연관 단어는 연관 단어 지식베이스의 포함 대상에서 제외시킴으로써 연관 단어 지식베이스는 사용자의 선호도가 높은 연관 단어로 구성된다.

유전자 알고리즘은 유전인자(gene), 염색체(chromosome), 모집단(population)을 사용하여 사용자의 선호도가 높은 연관 단어를 추출한다. 여기서, 유전인자는 마이닝 기법에 의해 추출된 연관 단어를 나타내며, 0과 1의 값을 갖는 비트로 표현된다. 염색체는 사용자가 선호하는 각 웹문서를 표현하며, 비트로 표현되는 유전인자의 집합으로 구성된다. 모집단은 사용자가 선호하는 웹문서로써 추출한 전체 문서를 나타내며 염색체로 구성된다. 모집단은 초기화, 적합율 계산, 재구성, 선택, 교배, 돌연변이, 평가의 과정을 통하여 진화한다. 진화된 모집단은 평가를 통하여 다음 세대로 진화를 계속할 것인가가 결정된다. 그림 5는 초기화, 적합율 계산, 재구성, 선택, 교배, 돌연변이, 적합율 평가 과정을 통하여 연관 단어 지식베이스를 최적화하는 유전자 알고리즘을 나타낸다. 함수 GA( )와 배열 Doc[class,doc,n], A_rule[class,wordruleset]은 그림 2에서 사용되는 의미와 동일하다. Gene[class,doc,rule]은 클래스의 웹문서를 유전인자로 표현한 배열이며, Fitness[class,doc]는 문서의 적합율을 저장한 배열이다. average_fitness는 클래스에 속한 전체 문서의 평균 적합율을 의미하며 fitness_ threshold는 적합율 임계값을 나타낸다. Pm, Pc는 각각 돌연변이율, 교배율을 의미한다. calculation_fitnesss( ), s_sel( ), *m_xovers*( ), *m_mut*( )은 각각 적합율 계산 함수, 염색체 선택 함수, 교배 함수, 돌연변이 함수를 의미한다.

```
Procedure GA(Doc[class,doc,n],A_rule1[class,wordruleset],
fitness_threshold,Pc,Pmos)
 /*class{(doc1:n11,n12,..),(doc2:n21,n22..)}명사추출*/
 Doc[class,docnum,n]<-MOR(htmlset)
 /* 모집단(population)을 초기화 */
 for (doc=1;doc<t;doc++) do begin {
   for (rule=1;rule<rulenum;rule++)
     If A_rule[class,rule] ∈ Doc[class,doc,n]
       Gene[class,doc,rule]<- 1 // 연관 단어가 웹문서에 있을
                                   경우
       else
       Gene[class,doc,rule]<- 0 //연관 단어가 웹문서에 없을
                                   경우
         endif}
 while (average_fitness < fitness_threshold) do begin {
 /* 각 웹문서의 적합율 계산 */
 calculation_fitness(Gene[class,doc,rule]);
  /* simple selection을 위한 재구성 */
    for (doc=1;doc<t;doc++) do begin {
  Fitness-s[class,doc]=Fitness[class,doc]/sum_fitness x 100
    }
    /* 선택, 교배, 돌연변이 */
    while (doc<=t) do begin {
     //simple selection
    s_sel(Gene[class,doc,rule],Fitness-s[class,doc]);//염색체
                                                    선택
      m_xovers(Gene[class,doc,rule], Pc);//교배
      m_mut(m_xovers(Gene[class,doc,rule], Pm));} //돌연변이
    /*평가: 평균적합율이 적합율임계값보다 작다면 진화 계속 */
    calculation_fitness(Gene[class,doc,rule]);
     if average_fitness >= fitness_threshold then exit;
     Fitness[class,doc] <=Fitness-s[class,doc]
    end
```

그림 5 최적화된 연관 단어 추출 알고리즘

초기화 단계에서는 수집된 문서를 염색체로 정의하고 유전인자로 표현한다. 이를 위해 클래스의 연관 단어(A_rule[class,rule])에 포함된 명사가 문서로부터 추출된 명사(Doc[class,doc,n])에 존재한다면 문서를 구성하기 위한 유전인자(Gene[class,doc,rule])는 1의 값을 갖는다. 존재하지 않는다면 유전인자는 0의 값을 갖는다.

유전자 알고리즘에서는 개체의 성능을 다른 개체와 비교하기 위하여, 혹은 개체가 얼마나 유전자 알고리즘이 적용되고 있는 가상의 환경에 잘 적응하고 있는지를 나타내는 척도로서 적합율을 모든 개체에 부여한다. 본 논문에서 선택한 적합율의 기준은 선택한 문서간에 유사도가 높다면 적합율이 높은 것으로 유사도가 낮다면 적합율이 낮은 것으로 판정하였다. 문서간의 유사도[15]는 유전자 알고리즘에서 사용되는 적합율은 비트간의 일치수를 사용하여 유사도를 구해야 하므로 식(3)의 Jaccard 방법[5]을 이용한다. 식(3)에서 $\#(docn \cup docm)$은 문서n(docn)을 표현한 염색체에서 1의 값을 갖는 유전인자와 문서m(docm)을 표현한 염색체에서 1의 값을 갖는 유전인자의 합집합의 개수를 의미하며, $\#(docn \cap docm)$은 문서n(docn)을 표현한 염색체에서 1의 값을 갖는 유전인자와 문서m(docm)을 표현한 염색체에서 1의 값을 갖는 유전인자의 교집합의 개수를 의미한다.

$$Fitness(docn,docm)=\#(docn \cap docm)/\#(docn \cup docm) \quad (3)$$

식(3)을 이용하여 클래스에 포함된 각 염색체의 적합율을 구할 수 있으며, 또한 클래스에 포함된 전체 문서의 평균 적합율을 구할 수 있다. 그림 6은 적합율을 구하는 함수 calculation_fitness( )이다.

```
calculation_fitness(Gene[class,doc,rule])
for (doc=1;doc<t;doc++)
   for (doc1=1;doc1<t;doc1++) do begin {
      P1=Gene[class,doc,rule]; P2=Gene[class,doc1,rule]
      /* Jaccard의 score에 의한 계산 */
   Fitness[class,doc]=Fitness[class,doc]
                  + ( Σ_{i=1}^{rulenum} P1P2 / ( Σ_{i=1}^{rulenum} P1^2 + Σ_{i=1}^{rulenum} P2^2 − Σ_{i=1}^{rulenum} P1P2 ) )
   Fitness[class,doc]=Fitness[class,doc]/t
   sum_fitness=sum_fitness+Fitness[class,doc]
   }
    average_fitness=sum_fitness/t;
```

그림 6 각 문서의 적합을 계산

재구성 단계에서는 적합을 계산을 위한 목적 함수의 값을 다른 값으로 재구성한다. 이러한 작업을 적합율 조정(fitness scaling)이라고 한다. 본 논문에서는 적합율의 조정을 위해 식(4)를 이용한다. 식(4)는 클래스에 속한 전체 문서의 적합율의 합에 대한 각 문서의 적합율(Fitness[class,doc])의 비를 계산한다. 이에 따라 적합율을 재조정되어 Fitness-s[class,doc]에 저장된다.

$$Fitness-s[class,doc]=\frac{Fitness[class,doc]}{\sum_{doc=1}^{t} Fitness[class,doc]} \quad (4)$$

선택 단계에서는 재구성된 적합율을 바탕으로 교배할 문서를 선택한다. 교배 단계에서는 선택된 문서를 대상으로 교배를 한다. 교배 방법에는 1점 교차방법, 2점 교차방법, n점 교차방법, 균일 교차방법 등이 있으나 본 논문에서는 유전인자의 수가 많으므로 1점 교차방법을 사용한다. 1점 교차 방법은 임의의 한 지점을 선택하여 그 이후의 유전인자를 다른 문서의 유전인자와 맞바꾸는 형식으로 동작한다. 그림 7은 교배의 예이다. 그림 7에서 Doc[1], Doc[2]는 부모를 의미하며, Doc1[1], Doc1[2]는 자손(child)을 의미한다. Doc[1], Doc[2]의 염색체가 교배를 하기 위해 선택된 임의의 수는 [1..18] 중 6이므로 6번째 이후의 유전인자를 서로 맞바꾸는 형식으로 교배가 이루어져서 Doc1[1], Doc1[2]의 자손이 탄생한다.

```
[1..18]에서 발생한 임의의 수가 6일 경우
교배 전 :Doc[1]=[1 0 0 1 0 1 1 1 1 1 0 0 0 1 1 1 1]
         Doc[2]=[1 1 0 0 0 0 0 1 1 1 1 0 0 0 1 0 0]
교배 후: Doc1[1]=[1 0 0 1 0 0 0 1 1 1 1 0 0 0 1 0 0]
         Doc1[2]=[1 1 0 0 0 1 1 1 1 1 0 0 0 1 1 1 1]
```

그림 7 염색체의 교배

교배를 하기 위해서는 교배율을 지정해야한다. 통상적으로 교배율(Pc)은 0.7-0.9를 사용하며, 본 논문에서는 0.9를 사용하였다. 교배율을 적용하는 방식은 선택된 부모 염색체에 [0..1]사이의 임의의 수(r)를 적용하여 r<=Pc이라면 선택된 부모 염색체는 교배되며, r>Pc이라면 선택된 부모 염색체는 교배되지 않고 부모 염색체가 바로 자손이 된다.

돌연변이 단계에서는 하나의 비트를 주어진 확률에 따라서 다른 값으로 바꾸는 단계이다. 돌연변이율은 염색체의 유전인자가 다른 값으로 바뀌는 확률을 의미한다. 돌연변이율이 100%라면 모든 유전인자의 값이 다른 값으로 변경되며, 0이라면 돌연변이가 일어나지 않는다. 돌연변이율은 0.01-0.05사이의 값을 많이 사용하며 본 논문에서는 돌연변이율을 0.01로 정한다. 그림 8은 그림 7에 나타난 교배 후의 염색체에서 돌연변이가 일어나는 예를 나타낸다.

```
*임의의 수(r)가 Pm(0.01)보다 작을 경우만 돌열변이가 일어남
예1>r=0.001일 경우의 돌연변이
Doc1[1]에서 임의의 유전인자(3)를 선택하여 다른 값으로 바꿈.
돌연변이후=>Doc1[1]=[1 0 1 1 0 0 0 1 1 1 1 0 0 0 1 1 0 0]
예2>r=0.1일 경우 Doc1[1]은 변하지 않음
```

그림 8 염색체의 돌연변이

평가 단계에서는 진화를 계속할 것인가 종료할 것인가를 결정하는 단계이다. 평가하는 기준은 평균 적합율이 적합율 임계값과 같거나 크다면 진화를 종료하고 작다면 재구성 단계부터 다시 진화를 반복한다. 본 논문에서는 적합율 임계값을 1로 하여 계산된 평균 적합율이 1보다 작다면 진화를 계속 진행한다.

4. 연관 단어 추출 및 정제 과정

본 논문의 실험을 위해 컴퓨터 분야의 웹문서를 8개의 클래스로 분류한다. 실험 대상의 웹문서를 8개의 클래스로 분류한 기준은 야후, 한미르, 알타비스타 등 기존의 정보검색 엔진에서 이용하는 분류 통계를 따른 것이다.

표 1은 그림 2에 속한 Block1의 MOR( )을 실행하였을 경우에 대한 한 예이다.

표 1 클래스별로 선정된 URL과 형태소 분석 결과 출력된 명사의 예

클래스	훈련 URL(트랜잭션필드)	형태소 분석 결과 추출된 명사들
게임	http://www.thegame.co.kr	게임,경고,인가,사용자,이벤트,참가...
그래픽	http://www.animatus.com	멀티미디어,출판사,컴퓨터,인테리어,활용.
뉴스와 미디어	http://www.chollian.net	인터넷,날씨,방송,신문,환경,오염,...
반도체	http://www.daejn-semi.co.kr	구축,설계,창업,기술,산업,메모리,...
보안	http://165.246.33.33	해킹,접근,발표,정보,활동,인공지능...
인터넷	http://www.gagaweb.co.kr	네트워크,컴퓨터,정보,교환,프로토콜....
전자출판	http://www.hiebook.com	서점,결제시스템,출판,기획,제작,내용...
하드웨어	http://www.savitmicro.com	메인보드,하드웨어,하드디스크,모니터...

그림 2의 Block2에 나타난 APRIORI( )는 표 1의 결과를 대상으로 Apriori 알고리즘을 적용하여 연관 단어를 생성한다. 이러한 연관 단어들 중에서 반복되는 선행 단어(antecedent)를 갖는 연관 단어를 삭제하면 표 2의 지식베이스를 구축할 수 있다.

표 2 각 클래스에서 추출된 연관 단어의 예

클래스	선행단어(Antecedent)	후행단어(Consequent)	평균 신뢰도	평균 지지도	연관 단어 수
게임	게임&구성&선수&경기&스포츠&참가	선발	91.30%	20.1039%	118
그래픽	빙법&중심&제작&사용	평가	88.10%	21.4286%	109
뉴스와미디어	뉴스&제공&홍보&속보	안내	99.9%	20.2838%	120
반도체	시스템&사업&활용&기법	컴퓨터	96.20%	20.3839%	122
보안	세계&네티즌&인물&조직&수법	해커	95.30%	21.7683%	125
인터넷	컨텐츠&사이트&관리&쇼핑몰	웹	94.90%	19.3833%	128
전자출판	입력&편집&출력&컬러&종류	출판	91.30%	18.2128%	129
하드웨어	보드&주변기기&슬롯&랜터엄	기기	90.20%	21.2532%	130

표 3은 게임 클래스를 대상으로 APRIORI( )에 의해 연관 단어를 추출한 예로, 18개의 연관 단어를 보인다.

표 3 게임 클래스에 나타난 연관 단어

(1)게임&구성&선수&경기&스포츠&참가=>선발	(9)데이터&암호&통신망=>가입
(2)국내&최신&기술&설치=>개발	(10)게임&이용&문제=>규칙
(3)게임&참가&인기&사용자&접속=>이벤트	(11)그림&인기&서비스=>음악
(4)운영&선발&경기&순위&규칙=>평가	(12)그림&데이터&서비스=>엔진
(5)게임&순위&이름=>스포츠	(13)데이터&프로그램=>음악
(6)운영&스포츠&위원회&선수=>선발	(14)그림&데이터&프로그램=>사진
(7)게임&구성&선발&순위=>경기	(15)게임&설명&제공=>공략
(8)게임&일정&선수&참가&운영=>스포츠	(16)게임&이용&기술=>개발
	(17)삭제&게임&개인전=>경고
	(18)게임&제공&일러스트=>설명

그림 2의 APRIORI( )를 사용하여 연관 단어를 추출하는 경우 상당수 부적절한 연관 단어가 포함된다. 이는 웹문서 수집기에 의해 추출된 웹문서 중에 각 클래스와 상관 없는 문서가 상당수 포함되는 것을 의미한다. 또한 적합한 웹문서에서 추출한 연관 단어일지라도 질의 확장에 사용될 가능성이 희박한 연관 단어가 존재하게 된다. 따라서 그림 5의 유전자 알고리즘을 적용하여 연관 단어 지식베이스를 정제해야 한다.

표 3의 연관 단어에서 부적절한 연관 단어를 제거하기 위해서 사용자가 선정한 10개의 웹문서를 대상으로 Block3의 MOR( )를 실행하여 표 4와 같이 명사를 추출한다.

그림 5의 초기화 단계에서는 문서를 유전인자의 집합으로 표현한다. 표 3의 게임 클래스에는 18개의 연관 단어가 존재하므로 모든 염색체는 18비트로 표현된다. 예로서, 첫 번째 비트는 표 3의 (1)의 연관 단어를 나타내며, 두 번째 비트는 (2)의 연관 단어를 표현하게 되어 염색체는 총 18개의 비트로 구성된다. 염색체의 각 비트는 표 3의 연관 단어가 표 4의 명사에 포함된다면 1(on)이 되고, 그렇지 않다면 0(off)임을 나타낸다. 이러한 초기화 과정을 거침으로써 각 문서는 표 5와 같이 유전인자로 표현된다.

표 4 모집단으로 선택한 웹문서의 명사 추출

웹문서	형태소 분석 결과에 의해 추출된 명사들
1	구성,페이지,창세기,규칙,이미지,배경,공략,갤러리,링크,설명,아이템,...
2	참가,결혼,선수,커플,갤러리,공략,배경,설명,수록,시리즈,시스템,에피소드,..
3	개인전,개편,게임,경고,경기,규칙,그룹,기준,담변,대전,대표,대립,문제,방법,..
4	기술,규칙,선수,공략,인기,접속,사용자,파일,개발,참가,운영,순위,평가,경기,..
5	리스트,게임,버튼,대전,이벤트,컴퓨터,하드웨어,정보,제공,구매,사용,대결,..
6	게임,이용,규칙,설명,제공,인기,사용자,선수,참가,운영,일정,문제,제공,공략,..
7	스포츠,야구,용병,위치,경기,구성,기록,적용,실전,자동,제품,대회,진행,참가,..
8	게임,국내,참가,인기,접속,작업,기술,설치,운영,순위,이름,선수,선발,경기,..
9	멀티미디어,사운드,국내,제공,제작,소프트웨어,음악,데이타,프로그램,그림,
10	개인전,스포츠,단체전,프로,아마,신청,대회,참가,우승,준우승,토너먼트,..

표 5 웹문서의 염색체 표현

웹문서	염색체(1세대)
문서$_1$	1011000000000100000
문서$_2$	0010000000000001001
문서$_3$	1010111101000011111
문서$_4$	1111111101000011111
문서$_5$	1010101101000011111
문서$_6$	1111111101000011111
문서$_7$	1000111100000000000
문서$_8$	1111111101000011111
문서$_9$	1111101101011111111
문서$_{10}$	1010111010000000010

표 5와 같이 문서를 염색체로 표현한 후에 그림 6의 문서 적합율 계산 알고리즘을 이용한 각 문서의 적합율과 식(4)에 이용하여 계산된 재구성된 적합율은 표 6에 나타난다.

표 6 적합율의 조정

웹문서	적합율	재구성적합율
문서$_1$	0.277170	5.354191
문서$_2$	0.275670	5.325215
문서$_3$	0.649913	12.554600
문서$_4$	0.612508	11.832030
문서$_5$	0.616830	11.915520
문서$_6$	0.665659	12.858770
문서$_7$	0.391846	7.569429
문서$_8$	0.665659	12.858770
문서$_9$	0.557083	10.761370
문서$_{10}$	0.464354	8.970091
평균	0.517669	10

선택 단계에서는 재구성된 적합율을 바탕으로 교배할 문서를 선택한다. 이러한 경우, 재구성된 적합율은 각 문서가 선택될 확률이다. 그러므로, 표 6의 문서$_3$, 문서$_4$, 문서$_5$, 문서$_6$, 문서$_8$은 부모 염색체로 선택될 확률이 높다.

교배 단계에서는 선택된 문서를 대상으로 교배율에 따라 교배를 한다. 돌연변이 단계에서는 돌연변이율 0.01에 따라 하나의 비트를 주어진 확률에 따라서 다른 값으로 바꾼다. 교배된 염색체에 [0..1] 사이의 임의의 수(r)를 적용하여 r=<0.01이라면 교배된 염색체는 돌연변이가 일어나며, r>0.01이라면 교배된 염색체는 돌연변이가 일어나지 않는다.

표 7은 표 5의 부모 염색체가 선택, 교배, 돌연변이 단계를 거쳐 탄생된 1세대의 자손인 2세대의 염색체이다.

표 7 2세대 염색체 및 적합율

웹문서	염색체(2세대)	적합율
문서$_1$	1010111111000011111	0.700600
문서$_2$	1111111101000001111	0.716735
문서$_3$	1010001101000011111	0.629429
문서$_4$	1010110101000011111	0.693230
문서$_5$	1101111101000000000	0.525678
문서$_6$	1000111101000011111	0.700791
문서$_7$	1101111101011111111	0.625147
문서$_8$	1111101001000011111	0.646749
문서$_9$	1111111101100010010	0.625490
문서$_{10}$	1010110100010000111	0.567764
평균		0.643161

표 8 세대별 염색체 및 평균 적합율

진화	염색체	적합율	평균 적합율
1세대	문서$_1$:1011000000000100000 문서$_2$:0010000000000001001	0.277170 0.275670	0.517669
2세대	문서$_1$:1010111111000011111 문서$_2$:1111111101000001111	0.700600 0.716735	0.643161
3세대	문서$_1$:1010111101000011111 문서$_2$:1111111101000011111	0.774610 0.774650	0.684610
4세대	문서$_1$:1010111111000011111 문서$_2$:1111111100000011111	0.790659 0.736813	0.8215736
5세대	문서$_1$:1000111101000011111 문서$_2$:1010111111000011011	0.834790 0.768407	0.8215740
6세대	문서$_1$:1011111111000011111 문서$_2$:1110111111000011011	0.779121 0.718132	0.8597770
7세대	문서$_1$:1010111101000011111 문서$_2$:1010111101000011111	0.991667 0.991667	0.9850000
8세대	문서$_1$:1010111101000011111 문서$_2$:1010111101000011111 문서$_3$:1010111101000011111 문서$_4$:1010111101000011111 문서$_5$:1010111101000011111 문서$_6$:1010111101000011111 문서$_7$:1010111101000011111 문서$_8$:1010111101000011111 문서$_9$:1010111101000011111 문서$_{10}$:1010111101000011111	1 1 1 1 1 1 1 1 1 1	1.0000000

평가 단계에서는 진화를 계속할 것인가 종료할 것인가가 결정된다. 표 8의 3열에 계산된 2세대 염색체의 평균 적합율 0.643161이므로 유전자 알고리즘의 종료 조건에 부합된다. 따라서 다음 세대로의 진화가 계속된다.

진화는 위와 같은 과정을 반복하며 평균 적합율이 1이 될 때까지 계속된다. 표 8은 1세대에서 마지막 세대인 8세대까지의 염색체와 각 세대의 평균 적합율을 보인다. 1세대에서 7세대까지는 대표로 문서$_1$,문서$_2$만을 보인다.

표 8과 같이 염색체가 진화됨에 따라 각 세대의 평균 적합율이 변화되는 과정을 그래프로 나타내면 그림 9와 같다.

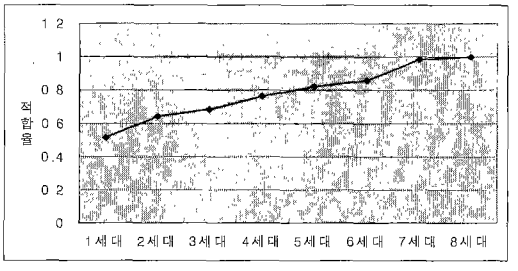


그림 9 각 세대의 평균 적합율

그림 9에서와 같이 진화가 진행되는 동안 평균 적합율은 일정하게 증가되어 평균 적합율이 1이 되는 세대인 8세대에서 진화는 종료되었다. 8세대의 염색체에서 1의 유전인자를 나타내는 연관 단어는 채택되고 0의 유전인자를 나타내는 연관 단어는 제거된다. 이에 따라 표 3에 나타난 게임 클래스의 연관 단어 중에서 0의 유전인자를 나타내는 연관 단어를 제거하면 연관 단어 지식베이스의 게임 클래스는 최적화된다.

5. 성능 평가

본 논문에서는 Apriori 알고리즘을 사용하여 연관 단어 지식베이스를 구축하였으며 유전자 알고리즘을 사용하여 구축된 지식베이스를 최적화하였다. 이를 평가하기 위해 Apriori 알고리즘은 연관 단어를 추출하는 기존의 방법인 상호정보량[14]과 Rocchio 알고리즘[8]과 비교하였으며 유전자 알고리즘은 TF·IDF를 이용한 텍스트 여과 방법[13]을 응용한 TF·IDF를 이용한 단어 정제 방법과 비교하였다. TF·IDF를 이용한 단어 정제 방법은 먼저 사용자가 선호하는 문서를 형태소 분석한 후 명사를 추출한다. 다음으로 추출된 명사를 대상으로 식(5)를 사용하여 TF·IDF값을 계산한다. 계산 결과, 연관 단어 지식베이스의 연관 단어에 속한 명사의 TF·IDF값이 1보다 작은 경우의 단어는 연관 단어 지식베이스에서 제거한다.

$tfidf_{ik}$= 문서$_k$에서의 단어$_i$의 발생 빈도*log_2(전체문서수)/(단어$_i$가 나타난 문서의 수) (5)

평가를 수행하기 위해 각 클래스에 대해 분류 결과를 대상으로 표 9와 같은 분할표[6]를 작성한다.

분류의 측정은 식(6)의 F_β 측정식을 이용한다.

표 9 2 x 2 분할표

분류A \ 분류B		분할 B (알고리즘2로 생성된 분류)	
		YES	NO
분할 A (알고리즘1이 만든 분류)	YES	a	b
	NO	c	d

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P+R} \quad (6)$$

$$\left(P=\frac{a}{a+b}100\%\ ,\ R=\frac{a}{a+c}100\%\ \right)$$

식(6)에서 P는 정확도, R은 재현율을 의미하는데, F_β의 값이 클수록 분류가 우수함을 의미한다. 여기서, β는 F_β 측정식에서 정확도에 대한 재현율의 상대적인 가중치를 나타내는 수치로 1.0일 경우 정확도와 재현율의 가중치가 같다. 본 실험에서는 β의 값을 1.0로 설정하여 클래스별로 분류 결과를 분석하였다.

그림 10, 그림 11, 그림 12은 연관 단어 지식베이스를 구축하기 위한 Apriori 알고리즘, Rocchio 알고리즘, 상호정보량의 3가지 방법의 정확도와 재현율을 식(6)에 대입하여 분석한 결과를 나타낸다.

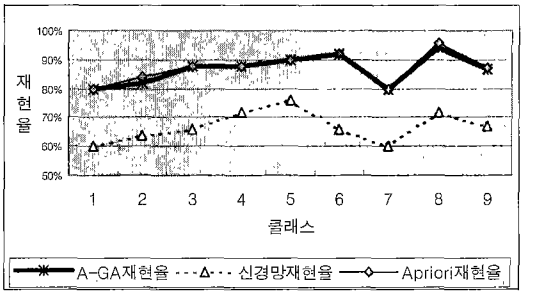


그림 10 Apriori, Rocchio, 상호정보량의 단어 분류 재현율

그림 10에서 Apriori 알고리즘의 재현율은 86.80%로 Rocchio 알고리즘을 이용한 방법보다 19.65%의 높은 재현율을 나타내고, 상호정보량을 사용한 방법보다는 0.53% 저하된 재현율을 나타낸다. 여기서, 0.53%는 100개의 웹문서만을 대상으로 사용함으로써 발생한 오차이다.

그림 11에서 Apriori 알고리즘의 정확률은 91.50%로 Rocchio 알고리즘을 사용한 방법보다는 평균 0.04%, 상호정보량을 사용한 방법보다는 19.34%의 향상된 정확도를 나타낸다.

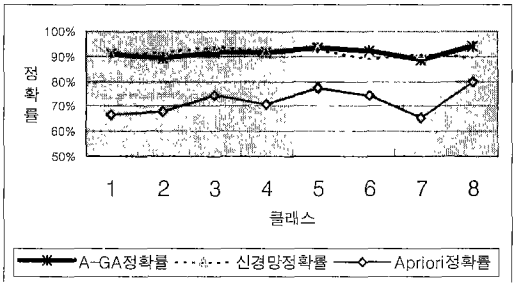


그림 11 Apriori, Rocchio, 상호정보량의 단어 분류 정확도

그림 12와 같이 β=1.0일 경우, F_β에 의한 단어 분류 성능은 Apriori 알고리즘은 89%로, 상호정보량을 이용한 방법보다는 10.04%, Rocchio 알고리즘에 의한 방법보다는 11.80% 높은 것으로 나타났다.

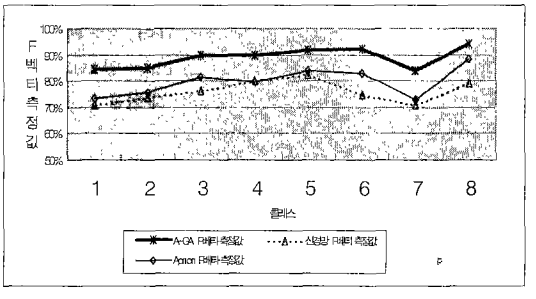


그림 12 Apriori, Rocchio, 상호정보량의 F_β 측정값

그림 13, 그림 14는 Apriori 알고리즘에 의해 구축된 연관 단어 지식베이스를 정제하기 위한 유전자 알고리즘과 TF · IDF를 이용한 단어 여과 방법의 정확도와 재현율을 식(6)에 대입하여 분석한 결과를 나타낸다. 유전자 알고리즘을 이용한 방법과 TF · IDF를 사용한 단어 여과 방법의 재현율은 각각 87.23%, 87.25%로 거의 흡사한 결과를 나타내므로 정확도의 관계만을 보인다.

그림 13은 식(6)을 바탕으로 한 정확도의 관계를 나타낸다. 그림 13에서 유전자 알고리즘의 정확도는 94.27%로 TF · IDF를 사용한 단어 여과 방법보다 14.12% 높게 나타났다.

그림 14과 같이 β=1.0일 경우, F_β에 의한 연관 단어 지식베이스의 정제 성능은 유전자 알고리즘에 의한 방법이 90.53%로 TF · IDF를 사용한 단어 여과 방법에 의한 방법보다는 7.02% 높은 것으로 나타났다.

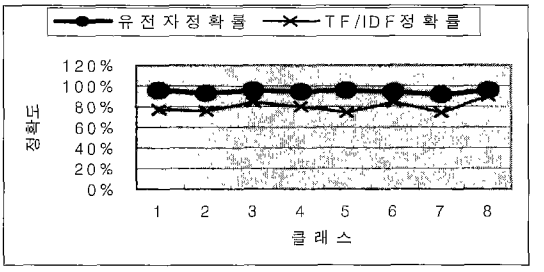


그림 13 유전자 알고리즘과 TF · IDF에 의한 단어 정제의 정확도

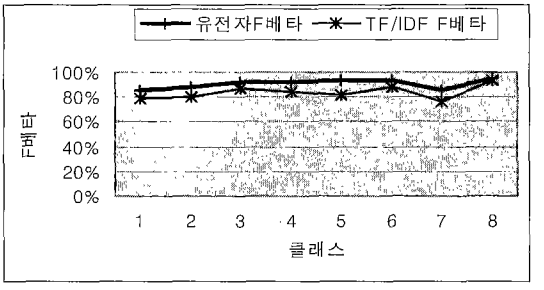


그림 14 유전자 알고리즘과 TF · IDF에 의한 단어 정제의 F_β 측정값

6. 결 론

본 논문에서는 지식 기반 정보검색 시스템에서의 질의 확장에 이용되는 지식베이스를 구축하기 위하여, 컴퓨터 분야의 웹문서를 수집하여 클래스별로 명사를 추출하고 Apriori 알고리즘을 이용하여 연관 단어를 추출하였다. 이를 대상으로 유전자 알고리즘을 적용하여, 사용자가 선호하지 않는 연관 단어를 제외시킴으로써 클래스별로 최적화된 연관 단어 지식베이스를 구축하였다. 본 논문에서 설계한 Apriori 알고리즘을 이용하여 클래스별 웹문서를

대상으로 연관 단어를 추출할 경우, F_β 측정 방법에 의한 단어 분류의 성능에서 상호정보량에 의한 방법보다는 평균 5.46%, Rocchio 알고리즘에 의한 방법보다는 평균 7.26% 높은 측정 결과를 얻을 수 있었다. 또한 Apriori 알고리즘에 의해 구축된 연관 단어 지식베이스를 정제하기 위한 유전자 알고리즘은 TF·IDF를 이용한 단어 여과 방법보다 7.02% 높은 분류 성능을 나타낸다.

향후, 복합 명사를 대상으로 Apriori-Genetic 알고리즘을 이용하여 클래스별 연관 단어 지식베이스를 구축한다면 정보검색 시스템의 성능을 더욱 높일 수 있을 것으로 기대된다.

참 고 문 헌

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.

[2] R. Agrawal and T. Imielinski and A. Swami, "Mining association rules between sets of items in large databases," Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, May 1993.

[3] P. Brown and P. Della and R. Mercer, "Class-based n-gram models of natural language," Computational Linguistics, 18(4), pp.467-479, 1992.

[4] C. Clifton and R. Steinheiser, "Data Mining on Text," Proceedings of the Twenty-Second Annual International Computer Software & Applications Conference, 1998.

[5] M. Gondon, "Probabilistic and genetic algorithms for document retrieval," Communication of the ACM,31, pp.1208-1218, 1988.

[6] V. Hatzivassiloglou and K. McKeown, "Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning." Proceedings of the 31st Annual Meeting of the ACL, pp.172-182, 1993.

[7] K. Hyun-Jin and P. Jay-Duke and J. Myung-Gil and P. Dong-In. "Clustering Korean Nouns Based On Syntactic Relations and Corpus Data," Proceedings of the LASTED International Conference Artificial Intelligence and Soft Computing, 1998.

[8] T. Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," Proc. 14th International Conference on Machine Learning, 1997.

[9] H. IU and R. Setiono and H. Liu, "Effective Data Mining Using Neural Networks," Proceeding of the IEEE Trans. Knowledge and data engineering, V.8 N.6, pp.962-969, 1996.

[10] G. Miller, "Wordnet:An on-line lexical database," International Journal of Lexicography. 3(4), pp. 235-244, 1990.

[11] K. Miyashita and K. Sycara, "Improving System *Performance in CaseBased Iterative Optimization* through Knowledge Filtering," Proceedings of the International Joint Conference on Artificial Intelligence, 1995.

[12] T. Michael, *Maching Learning*, McGraw-Hill, pp. 249-273, 1997.

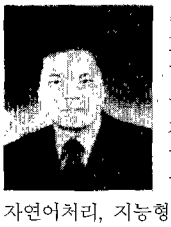
[13] D. W. Oard and G. Marchionini, "A Conceptual Framework for Text Filtering," Tehcnical Report CAR-TR-830, Human Computer Interaction Laboratory, University of Maryland at College Park, 1996.

[14] C. Plaunt and B.A.Norgard, "An association based method for automatic indexing with a controlled vocabulary," *Journal of the American Society for* Information Science, 49, pp.888-902. 1998.

[15] 한승희, 이재윤, "문헌 클러스터링을 위한 유사계수간의 연관성 측정", 제6회 한국정보관리학회 논문집, pp. 25-28, 1999.

고 수 정

1990년 인하대학교 전자계산학과 졸업(학사). 1997년 인하대학교 교육대학원 전자계산교육(교육학석사). 2000년 인하대학교 대학원 전자계산공학과 박사과정 수료. 1990년 ~ 1991년 수협중앙회 전자계산소 근무. 1992년 ~ 2000년 문성여자상업고등학교 교사. 2001년 ~ 현재 유니버셜 소프트 정보통신(주) 연구원. 관심분야는 데이타마이닝, 정보검색, 기계학습



최 준 혁

1990년 2월 경기대학교 전자계산학과(이학사). 1995년 2월 인하대학교 전자계산공학과(석사). 2000년 8월 인하대학교 전자계산공학과(박사). 1997년 ~ 2001년 현재 김포대학 컴퓨터계열 (소프트웨어개발 전공) 조교수. 관심분야는 정보검색, 자연어처리, 지능형 알고리듬

이 정 현

1977년 인하대하교 전자공학과 졸업. 1980년 인하대학교 대학원 전자공학과(공학석사). 1988년 인하대학교 대학원 전자공학과. (공학박사). 1979년 ~ 1981년 한국전자기술연구소 시스템 연구원. 1984년 ~ 1989년 경기대학교 전자계산학과 교수. 1989년 ~ 현재 인하대학교 전자계산공학과 교수. 관심분야는 자연언어처리, HCI, 정보검색, 음성인식, 음성합성, 계산기 구조