사용자 의도 트리를 사용한 동적 카테고리 재구성

김 효 래[†] · 장 영 철[†] · 이 창 훈[††]

요 약

기존에 단어의 빈도수를 근간으로 하는 문서 분류 시스템에서는 단일 키워드를 사용하기 때문에 사용자의 의도를 충분히 반영한 문서 분류가 어려웠다. 이러한 단점을 개선하기 위하여 우선 기존의 설명에 근거한 학습방법(explanation based learning)에서 한 예제만 있어도 지식베이스 정보와 함께 개념을 학습할 수 있다는 점에 착안하여 먼저 사용자 질의를 분석, 확장한 후 사용자 의도 트리를 생성한다. 이 의도 트리의 정보를 기존의 키워드 빈도 수에 근거한 문서분류 과정에 제약 및 보충 정보로 사용하여 사용자의 의도에 더욱더 근접한 웹 문서를 분류할 수 있다. 문서를 분류하는 측면에서 볼 때 구조화된 사용자 의도 정보는 단순한 키워드의 한계를 극복하여 문서 분류 과정에서 특정 키워드 빈도수의 임계값을 결정함으로서 잃게되는 문서 및 정보를 좀더 보유하고 재적용할 수 있게 된다. 질의에서 분석, 추출된 사용자 의도 트리는 기존의 통계 및 확률을 사용한 문서 분류기법들과 조합하여 사용자 의도정보를 제공함으로서 카테고리의 형성 방향과 범위를 결정하는데 높은 효율성을 보인다.

Dynamic recomposition of document category using user intention tree

Hyo Lae Kim[†] · Young Cheol Jang[†] · Chang Hoon Lee[††]

ABSTRACT

It is difficult that web documents are classified with exact user intention because existing document classification systems are based on word frequency number using single keyword. To improve this defect, first, we use keyword, a query, domain knowledge. Like explanation based learning, first, query is analyzed with knowledge based information and then structured user intention information is extracted. We use this intention tree in the course of existing word frequency number based document classification as user information and constraints. Thus, we can classify web documents with more exact user intention. In classifying document, structured user intention information is helpful to keep more documents and information which can be lost in the system using single keyword information. Our hybrid approach integrating user intention information with existing statistics and probability method is more efficient to decide direction and range of document category than existing word frequency approach.

키워드 : 의도정보(Intention Information), 문서분류(Document Classification)

1. 서 론

많은 정보량을 지닌 데이터들의 특성을 잃지 않으면서 정보량을 줄여 인간이 보기 편리하도록 데이터를 가공하거나, 여러 정보들이 내재된 큰 데이터 중에서 자신이 원하는 정보만을 골라 제공할 필요가 있다[1, 2]. 이러한 작업을 클러스터링(clustering)이라고 한다.

클러스터링 과정에서 유사한 특성을 가지는 데이터들은 함께 묶여 이들 데이터가 가지고 있는 공통적인 특징을 그 군집의 대표로 나타내어 전체에 산재되어 있는 데이터를 몇 가지의 특성 군으로 나누어 주게 된다. 기존에 대표적인 문서분류 및 검색 시스템들을 살펴보면 top-down 클러스터링 알고리즘인 Cobweb시스템, 사용자의 행동을 모니터링하고 사용자의 관심도를 학습하는 방식인 personal webwatcher[3], 사용자의 기존 검색을 분석하여 얻은 프로화일을 이용해서 사용자들에게 각 개인이 원하는 정보를 검색하고 분류할 수 있는 wise wire 시스템[4] 등이 있다.

이들 시스템중 인간의 점진적인 개념학습을 모델링하기 위해 개발된 Cobweb은 개념적 클러스터링 알고리즘으로 분류

† 정 회 원 : 경민대학 컴퓨터정보통신학부 교수
†† 종신회원 : 건국대학교 인터넷 멀티미디어학과 학장
논문접수 : 2001년 8월 9일, 심사완료 : 2001년 10월 15일

대상이 되는 개념을 계층적인 구조로 구성하고 하향식으로 분류를 수행한다. 또한, 점진적 학습을 하고 학습방식은 비 감독 학습방식으로 수행하는 등의 특징을 가지고 있다[5, 6]. 그러나 Cobweb은 단어의 빈도수를 이용한 통계 및 확률방식으로 문서를 분류하기 때문에 사용자의 관심과 의도가 제대로 반영될 수 없는 일이 발생할 수 있다. 따라서 본 연구에서는 이러한 문제를 해결하기 위하여 지식베이스나 개념 시소러스 등을 이용하여 사용자 질의를 분석, 확장한 후 사용자 의도 트리를 생성하여 기존의 키워드 빈도수에 근거한 문서 분류 기법인 cobweb 시스템에 제약 및 보충 정보로 사용하여 사용자의 의도에 더욱더 근접한 문서 분류 카테고리를 재구성할 수 있는 D-car(Dynamic category rebuilder) 시스템을 제안한다.

2. 관련 연구

2.1 Cobweb 알고리즘

Cobweb에서 사용하는 평가함수는 CU(Category Utility)를 사용하며 CU의 기본 개념은 주어진 분류에 대하여 어떤 개체의 속성이 기대되는 기대치의 합(X)에서 분류를 고려하지 않은 기대치(Y)를 뺀 값을 주어진 분류의 개수(K)로 나눈 값으로 표현되며[13], 이를 수식으로 표현한 것은 식 (1)과 같다.

$$\frac{\sum_{k=1}^{K} P(C_k) \sum_{i=1}^{I} \sum_{j=1}^{J} P(A_i = V_{ij} | C_k)^2 - \sum_{i=1}^{I} \sum_{j=1}^{J} P(A_i = V_{ij})^2}{K} \quad (1)$$

위에서 $P(C_k)$란 분류 K의 전체에 대한 비율이고, $P(A_i = V_{ij} | C_k)$는 주어진 분류에 대하여 개체의 속성이 특정 값을 가지는 확률을 나타내며, I는 속성의 개수, J는 학습개체의 개수를 나타낸다.

Cobweb의 특징으로는 분류 대상이 되는 개념을 계층적인 구조로 구성하고 하향식 분류를 수행하며 비 감독 학습방식이다. 또한 점진적인 학습(incremental learning)을 수행하고 새로운 학습대상에 대하여 힐 클라이밍(hill climbing)기법을 사용하여 효과적인 해를 구하는 방식으로 수행된다.

Cobweb에서 문서분류 트리를 구성하기 위해 사용되는 학습 연산자는 "Incorporate, Create-new-disjunct, Merge, Split"의 4가지가 있다. Incorporate 연산자는 새로운 입력 문서를 현재 노드의 자식 노드에 포함시키는 연산자이며 Create-new-disjunct 연산자는 새로운 입력 문서가 기존의 노드로 분류되기에는 너무 상이한 문서인 경우에 적용하는 연산자이다. Merge연산자는 현재 레벨의 자식 노드들 중에서 유사한 2개의 노드를 새로운 노드의 자식 노드로 변환하는 연산자이며 Split연산자는 현재 노드에 속하는 자식 노드의 성격이 너무 일반화된 경우에 세분화하기 위해서 적용하는 연산자이다[14].

2.2 학습 문서 전처리 알고리즘

학습 문서의 전처리 과정은 문서를 대표하는 단어를 색인어로 추출하거나 부여하는 과정이다. 색인은 특정한 정보가 필요한 사람에게 그 정보의 위치를 지시해주는 역할과 방대한 정보원으로부터 가장 유사한 내용의 정보만을 선별해 주는 역할을 한다. 이러한 색인어를 추출하기 위한 방법으로는 일단 웹 상에서 1차로 검색되어온 Snippet 문서에 대하여 Html Tag를 제거하는 알고리즘과 색인 단어의 수를 줄이기 위해 사용되는 불용어(stop list)제거 및 어근 추출(stemming) 알고리즘[9], 대상 단어의 집합에서 빈도수를 이용하여 주요한 키워드만을 추출하는 TF-IDF(Term Frequency-Inverse Document Frequency) 알고리즘이 있다. 또한 각 주요 단어의 문서 길이에 따른 영향력 불균형을 해결해주는 벡터길이 정규화(vector length normalization)알고리즘이 있다. 본 논문에서는 이러한 알고리즘들을 이용해서 학습 대상 전처리 과정을 수행하게 된다.

3. 의도 정보 분석

3.1 의도(intention)

기능적, 지역적, 시간적으로 분산되고 모듈화되어 가는 현 시스템들은 효율적으로 관리하기 위하여 모듈(에이전트) 및 프로그램에 자율권과 상호 협력하는 사회성을 부여하여 문제 해결능력을 향상시키는 접근 방법이 널리 시행되고 있다. 이때 시스템의 행동을 직관적으로 빨리 예측하기 위하여 의도를 사용한 시스템을 사용하게 된다. "학생은 비가 올 것을 믿기(believe) 때문에 우산을 가지고 갔다.", "학생은 대학 진학을 소망하기(desire) 때문에 열심히 공부했다." 위 두 문장에서 학생의 행동을 예측하기 위해 학생의 태도를 "believe", "desire"로 특성화했다. 이를 "의도 시스템"이라고 부른다[7].

의도는 소망, 믿음 등의 관계 또는 시스템이 취하는 행동들의 관계로 정의된다. 즉, 시스템(에이전트)은 능력(자원)과 환경에 대한 믿음을 가지고 어떤 일을 할지를 결정하게 되는데 이를 "의도"라고 볼 수 있다[12].

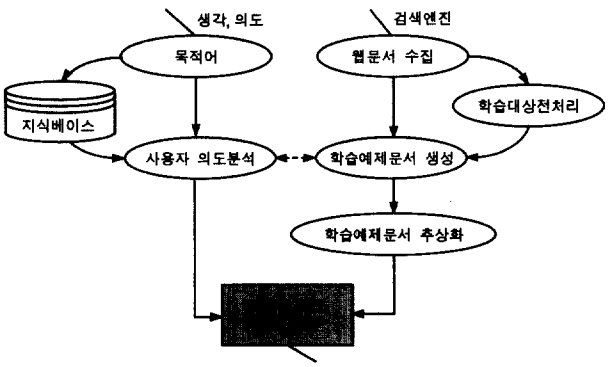
앞으로 할 일을 미리 정적으로 결정하는 측면에서 보면 계획(plan)의 성격을 나타내고 현 상황에서 의도에 맞춰 가능한 여러 행동중에서 상황에 적절한 행동을 취하려는 입장에서는 전략(strategy)의 성격을 나타낸다[8].

이 중 의도의 전략적인 해석은 시스템이 상황 변화에 능동

적으로 대처하고 문제의 제어방법을 함축성있게 표현하도록 해준다. 본 연구에서도 이러한 의도 개념을 이용하여 사용자의 의도 정보를 추출하여 문서를 분류하는데 있어서 사용자의 의도가 반영될 수 있도록 하였다.

3.2 의도 트리의 구성

본 절에서는 사용자의 의도를 포함한 생각이라든가, 목적을 이용하여 지식베이스나 개념 시소러스를 통해서 목적 키워드를 근거로 하여 개념을 확장한다. 본 연구에서 사용되는 지식베이스는 실험용으로서 건강과 의학분야로 도메인을 설정하여 작은 규모의 지식베이스를 구축, 실험하였으며 실험에서 적용된 몇 개의 카테고리는 <표 1>과 같다.



(그림 1) 사용자 의도트리 생성과정

〈표 1〉 실험 지식베이스의 구조

```
Health and Medicine
  {disease (cancer (stomach cancer
                    (prevention (leisure (sport or exercise,
                                          tour or trip ···)) ;
                                (prohibition (drink, smoke ···)) ;
                                (drug (vitamin, dietition or
                                       dietision ···)) ;
                    (treatment (radiobiology)) ;
                                (herb (needle ···)) ;
                                (drug (···)) ;
           (liver cancer) ;
           (prevention) ;
           (treatment) ;
     (diabetes)} ;
  {diet (method (exercise (running, cycle, ···)) ;
           (alimentotherapy or diet cure (fruit (apple, graph, ···)) ;
                                         (cereals (rice, ···)) ;
                                         (vegetables (cucumber,
                                                      potato,
                                                      carrot, ···)) ;
           (chinese medicine (matrimony vine, milk vetch ···)) ;
           (nature (black pea, water, ···)) ;
     (diet consultation)} ;
```

웹 환경에서 수집된 예제문서들은 학습 대상 전처리 과정인 HTML 태그 제거, 불용어 제거, 스태밍처리, TF-IDF 알고리즘 등을 거쳐 학습 예제 문서가 생성된다. 이때 생성된 학습 예제문서에서 각각의 키워드로부터 지식베이스를 통해 사용자 의도를 추출해내고 문서내에서의 주(major)의도와 이와 관련된 세부 의도들을 추상화한다. 추상화된 사용자 의도들은 주(major)의도와 세부 의도의 보정 작업으로 사용자 의도 트리가 생성되며 이 의도 트리는 4장에서 설명되는 동적 카테고리 재구성시에 사용자 의도로서 반영된다.

사용자의 검색의도는 상위레벨의 추상화된 단어로 구성되어 있어 이들은 다시 관련된 세부의도들로 나뉘어질 수 있다. 이같은 전개과정은 현재 웹 문서내의 키워드들과의 관계를 찾기 용이하게 해준다.

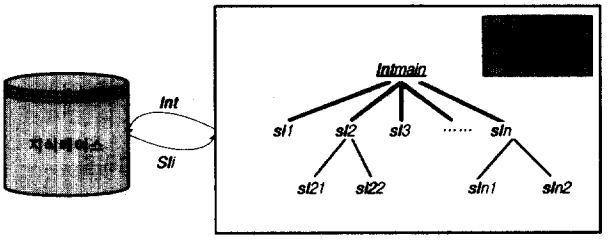
예를 들어 검색시에 (병 : 예방), (접시 : 장식), (망치 : 목공) 등의 (키워드 : 의도)조합이 있을 수 있다. 위 예들은 다시 (병 : 치료), (접시 : 요리), (망치 : 목공)등의 다른 의도로도 검색이 가능하다.

즉, 의도를 단어의 의미로 해석하는 방향(dimension) 또는 문맥(context)을 잡아가는 방향으로 보게 된다. 이는 점점 세부의도들로 전개됨에 따라 이를 표현하여 얻을 수 있는 단어들과 연결된다.

아래 그림의 ***sIi***는 [LB, UB], $\{C_1 \cup C_2 \cup \cdots \cup C_n\}$ 등의 범위, 제약조건, 순서 등의 일반화된 개념을 나타내게 된다. 이 세부의도들이 다음절에서 추상화된 키워드들과 비교된다. 위에서 ***sI1***, ***sIn***은 하나의 개념을 나타내며 하위 의도들을 포함한다.

전개과정은 영역별로 정의된 사용자 의도의 부분기능, 부분능력, 관련수단, 세부계획 등을 관련 시소러스에서 찾아 나열한다.

이같은 전개가 이루어지면 다음과 같은 의도들간의 포함관계, 순서, 전이규칙이 정의되어야 한다.



(그림 2) 지식베이스를 통한 의도 전이도

[포함관계]

<***Int***> : : + <leafword> | <non-leafword> | {<***Int***>··· <***Int***>}

[순서]

<*Inti*> | → [*Intl, Intm, Intn*]

의도 *Intl, Intm, Intn*이 차례로 수행되어야 *Inti*가 얻어진다는 의미이다.

[전이규칙]

<Belief> : : = <*Inti*> | → <*Intj*>

*Intj*가 수행되어 *Inti*로 바뀐다는 신뢰도를 표현

지정임계신뢰도 수치 이상이면 전이

Func TransFORM(*Int, k*)

{While *(Belief* > ***k****인 전이규칙 존재)*

{***IntTree*** = ***NewInt***

TransFORM(*Int, k*)

}

}

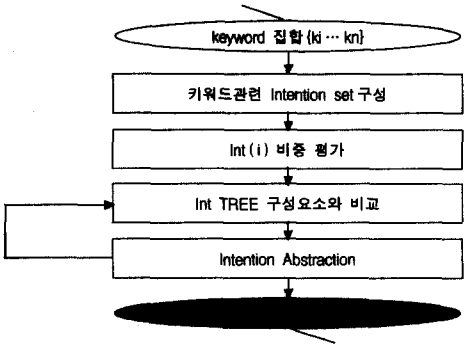
Int *: intention*

k *: 임계 신뢰도*

IntTREE *: 현재까지 전개된 의도트리*

3.3 학습 예제 문서 추상화

현재 웹 문서가 사용자 의도의 어느 세부의도와 어느 정도 부합하는가를 결정하는 과정이다. 이를 위해 학습대상 웹 문서의 전처리 과정에서 생성된 학습 예제 문서를 사용한다. 이들을 추상화시켜 관련된 의도들을 찾아내는데 웹 문서의 학습대상 전처리 후 학습 예제문서 집합이 $\{K_1, K_2, \cdots, K_n\}$이라 할 때 관련 의도를 찾아가는 과정은 아래와 같은 단계로 진행된다.



(그림 3) 주의도(major intention) 결정 알고리즘

3.3.1 키워드관련 Intention Set구성

지식베이스에서 다음과 같은 키워드 K_i관련 의도들을 찾는다.

$$K_i = \{ Inti_1, Inti_2, \cdots, Inti_n \}$$

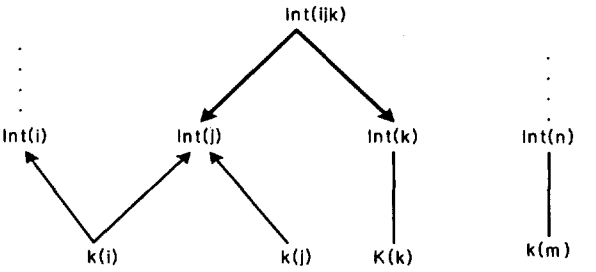
$$K_j = \{ Intj_1, Intj_2, \cdots, Intj_n \} \cdots$$

3.3.2 Int(i) 비중평가

생성된 의도에 대한 현 문서 내 소속 키워드 집합을 구하고 키워드 충족율 $\{ R_k = \frac{\text{문서내구성키워드수의총합}}{KB\text{내구성키워드수의총합}} \times 100 \}$를 구한다.

3.3.3 Int TREE구성 요소와 비교

의도 전이 트리의 구성에서 세부의도와 일치하는 의도로 탐색한다.



(그림 4) 사용자 의도 추상화

3.3.4 Intention Abstraction

임계치 수준까지의 키워드 구성율이 나오도록 세부의도를 추상화한다. 즉, 여러 키워드를 소유한 일반화된 의도를 생성하여 다시 IntTREE 내용과 비교한다.

3.3.5 주의도(major Intention) 생성

IntTREE의 내용과 일치하고 R_k 값이 가장 높은 의도가 현 문서의 주의도가 된다.

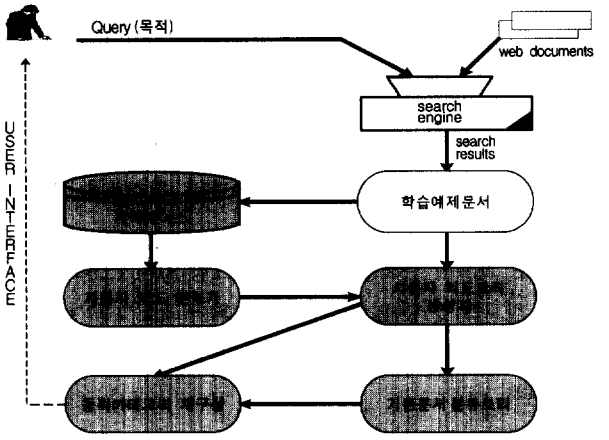
4. 동적 카테고리 재구성

4.1 D-car 시스템의 구조

동적 카테고리 구성방법의 주요 개념은 다음과 같다. 초기 사용자 의도 트리내의 문서들 간의 관계가 관련 문서들 사이에서 그룹정보로 작용하여 빈도수를 이용한 문서분류 방법에서 이동범위 또는 문서분류 가능한 범위로의 제약조건이 된다. 즉, 문서 그룹별로 사용자의 세부의도가 반영된 문서분류가 이루어진다. 이와 같이 사용자 의도 트리를 사용하여 문서분류 카테고리를 재구성하는 시스템인 D-car(Dynamic category rebuilder)의 구조는 (그림 5)와 같다.

(그림 5)에서 사용자의 검색 키워드로 일반 범용 검색 엔진을 통해 웹 문서를 수집하고 수집된 웹 문서는 학습 대상 전처리 과정인 HTML 태그 제거, 불용어 제거, 스태밍처리, TF-IDF 알고리즘 등을 거쳐 학습 예제 문서가 생성된다. 또한 사용자의 목적어는 지식베이스를 통해 사용자의 의도를 분석하고 개념을 확장해서 학습 예제 추상화 과정을 거쳐 사용자 의도 트리가 생성된다. 학습 예제 문서는 다시 기존의 단

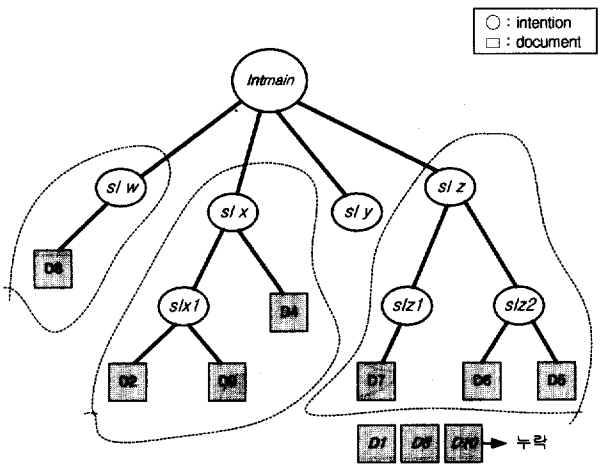
어 빈도수에 의한 분류 기법인 cobweb을 통해 문서 분류 카테고리가 생성된다. 생성된 카테고리는 앞에서 생성된 사용자 의도 트리와의 병합 과정을 거쳐 사용자의 의도가 반영된 카테고리로 재구성된다.



(그림 5) D-car 시스템 구성도

4.2 의도 트리를 사용한 문서분류

{D1, D2, D3, D4, …, D10}의 문서에서 학습 대상 전처리 과정을 수행 한 후 생성된 학습 예제 문서에서 대표 키워드의 집합을 선정하고 이들 중 지식베이스에서 사용자 쿼리 의도가 전개된 세부 의도들과의 부합된 문서들의 집합을 선정하면 아래 그림과 같다.



(그림 6) 의도들과 부합된 문서분류

위 트리의 구성은 다음과 같다.

(1) 전체 학습대상중 ***IntTREE*** 내에서 의도를 가진 문서만 선정

(2) ***Intmain***으로부터 세부의도(***sIi***)를 전개하고 관련문서 분류

문서들간의 관계분석(의도그룹 기준)

위 트리에서 ***sIw, sIx, sIy, sIz*** 등의 의도그룹을 찾을 수 있다.

(그룹의 형성기준은 다양하게 변화 가능)

각 그룹내에서 문서들간의 특징을 추출해 낸다.

예) ***Intention*** 그룹 1) 그룹의 깊이 3

선정그룹{{D8}, {D2, D9}, {D4}, {D7}, {D6, D5}}

2) 그룹의 넓이 3

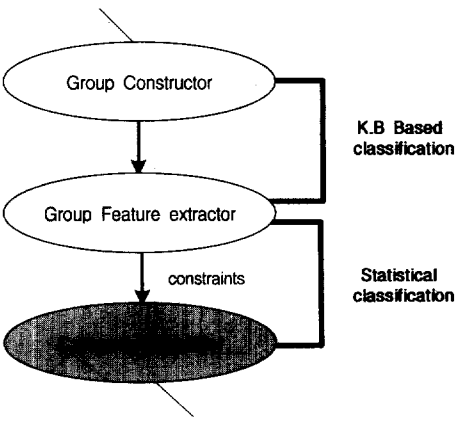
(***sIy***는 관련문서 없음)

이같은 그룹내 특징은 사용자 의도를 반영한 문서 분류이며 키워드 빈도수 등을 고려하지 않고 지식 베이스내에서 단어들간의 관련성을 개괄적으로 반영한 것이다. 이때 그룹의 크기를 조절하며 의도 트리의 성능변화를 실험 할 수 있다.

위에서 예)의 기준 1, 2는 빈도수, 통계 방법에 근거한 문서 분류 카테고리의 재구성 기준으로 사용되게 된다.

4.3 동적 카테고리 재구성

지식베이스에 근거한 사용자 의도 트리 정보를 가지고 분류된 문서들은 각 의도들에 대하여 그룹을 형성하게 된다. 이때 문서들의 그룹내 위치는 그룹의도의 의미을 만드는 한 특성일 뿐만 아니라 그룹내 문서들 상호간 관련성의 근거가 된다. 이러한 사용자 의도와 문서 키워드 관련성에 관한 정보를 기존의 키워드 빈도수를 이용한 문서 분류 방법과 병합해서 카테고리가 재구성되는 과정을 아래 그림에서 설명한다.



(그림 7) 카테고리 재구성 과정

4.3.1 Group constructor

4.2절의 (그림 6)에서 세부 의도들에 따른 문서들의 그룹을 형성한다. 문서들간의 관계구조에서 의미를 부여하기 위

해서는 그룹의 사이즈를 크게 구성하고 문서내에서 키워드 출현에 의미를 두기 위해서는 그룹 사이즈를 적게 구성한다.

4.3.2 Group feature extractor

그룹내에서 관련정보를 뽑아내는 단계로 다음과 같은 정보를 사용한다.

① 그룹내 문서를 포함하는 개념범위(개념 노드수 : 트리 넓이)

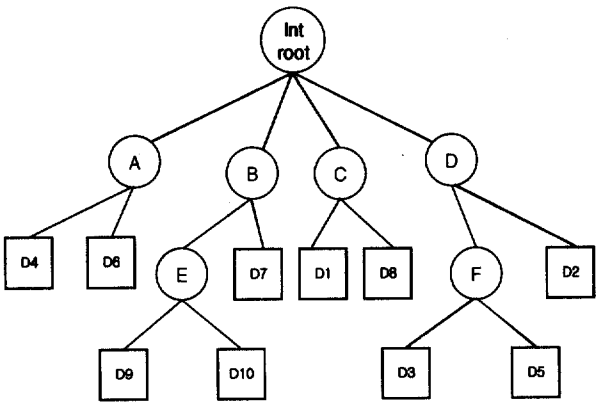
② 그룹의 깊이 (세부개념표현)

그 외에 다양한 그룹 특징을 문서간 관계를 제시할 수 있다.

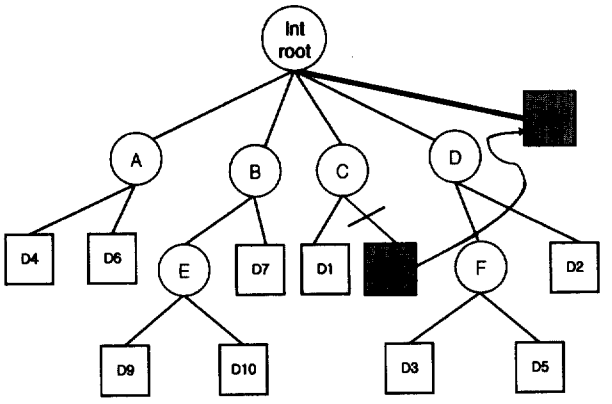
4.3.3 Category rebuilder

빈도수 확률에 근거한 카테고리 체계에서 Group feature extractor 단계에서 생성된 제약조건에 위배되는 문서의 위치를 변경, 조정한다.

아래의 카테고리를 (그림 6)의 의도 트리에서 생성된 기준인, 그룹 원소간 깊이 3, 넓이 3 이내의 유지 규칙을 적용하면 (그림 9)와 같은 문서분류 카테고리가 재구성된다.



(그림 8) 빈도수확률에 근거한 카테고리 구성



(그림 9) 문서분류 카테고리 재구성

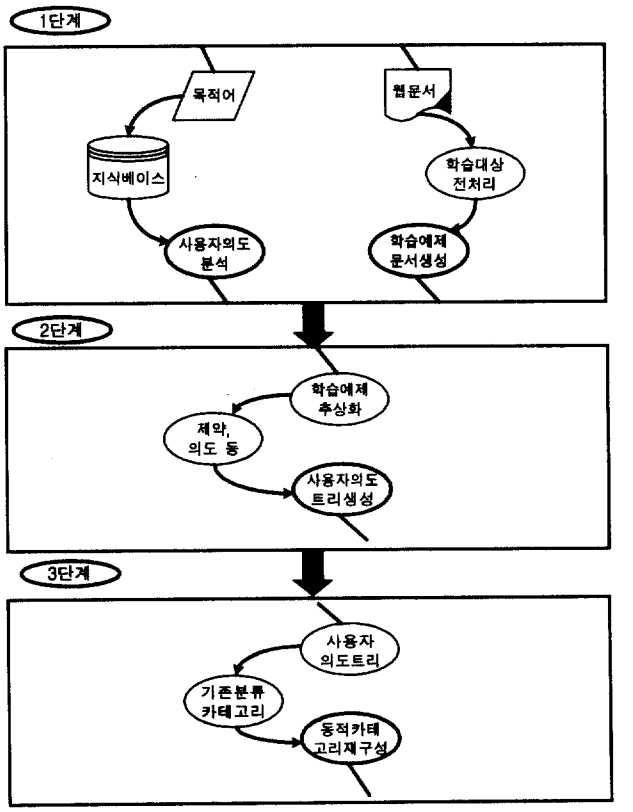
카테고리 재구성 알고리즘은 다음과 같다.

- 깊이조정 : 문서를 너무 세분화하였기 때문에 발생
  - 위배노드 문서 일반화(generalize)
  - 한 단계 상위개념으로 추상화
- 넓이조정 : 문서를 여러 의도로 해석할 때 발생하는 불필요한 해석방향을 삭제
  - 최소 원소그룹 선정 또는 공통 부모 축을 벗어난 소수그룹 문서 선정
  - 이동할 원소의 의도 트리내 관련 그룹원소 리스트 작성
  - 여러 관련 그룹원소 존재 → 공통 부모노드에 링크
  - 단일 관련 그룹원소 존재 → 단일 원소소속 노드에 링크

빈도수 확률에 근거한 카테고리 (그림 8)은 (그림 6)과의 기준에 비교해 볼 때 넓이규정이 위배되어 최소그룹 "C 그룹"을 선정하여 *Introot*로 "D8 문서"가 이동되어 카테고리가 재구성된다.

5. 실험 및 평가

D-car 시스템의 성능을 실험하기 위하여 문서 분류대상 도메인을 건강과 의학분야로 설정하였고 입력예제로는 HTML 형식으로 이루어진 웹 페이지로 실험하였다. 단계별 실험과정은 (그림 10)과 같다.



(그림 10) 단계별 실험 과정

1단계에서는 사용자의 검색어로 범용 검색 엔진을 이용하여 웹 문서를 수집하고 학습 대상 전처리 과정인 HTML 제거, 불용어 제거, 스태밍 처리, TF-IDF 알고리즘 등을 수행한 후 학습 예제 문서를 생성하고 사용자의 목적어는 지식베이스를 통하여 사용자의 의도를 분석하고 개념을 확장하는 실험을 한다.

2단계에서는 사용자의 의도 분석과 학습 예제 문서를 이용하여 학습 예제를 추상화하고 트리의 방향과 형태를 구성할 수 있는 제약 조건 등을 이용하여 사용자 의도 트리를 생성하는 과정을 실험한다.

3단계에서는 사용자 의도 트리를 이용하여 빈도수 확률에 근거한 문서 분류 카테고리를 재구성하는 실험을 한다.

또한 사용자의 만족도 조사를 위해서 임의로 사용자 2명을 선정하여 각자 관심있는 분야를 설정토록 하여 웹 상에서 문서를 검색, 수집하여 D-car 시스템에 입력예제로 사용하였다. 각 사용자들의 관심분야로는 <표 2>와 같으며 이를 입력 카테고리로 하여 문서를 분류하였다.

〈표 2〉 입력 데이터 set

실험	구 분	관심분야	목 적	웹문서수집	검색엔진
1차	사용자(A)	암	예방	암, 예방(10)	야 후
2차	사용자(B)	다이어트	방법	다이어트, 방법(9)	네이버

5.1 D-car 시스템 실험

5.1.1 사용자 의도 트리 생성

1단계 실험에서는 지식베이스를 통해서 사용자의 의도개념을 전개해가는 과정과 학습 예제 문서를 생성하는 실험을 하였는데 (그림 11)은 사용자의 목적이 지식베이스를 통해서 개념을 전개해 나가는 과정과 사용자의 검색어를 사용하여 웹 환경에서 웹 문서를 수집하고 학습대상 전처리 과정인 HTML 제거, 불용어 제거, 스태밍처리, TF-IDF 알고리즘 등을 수행한 후 결과를 보여준다.

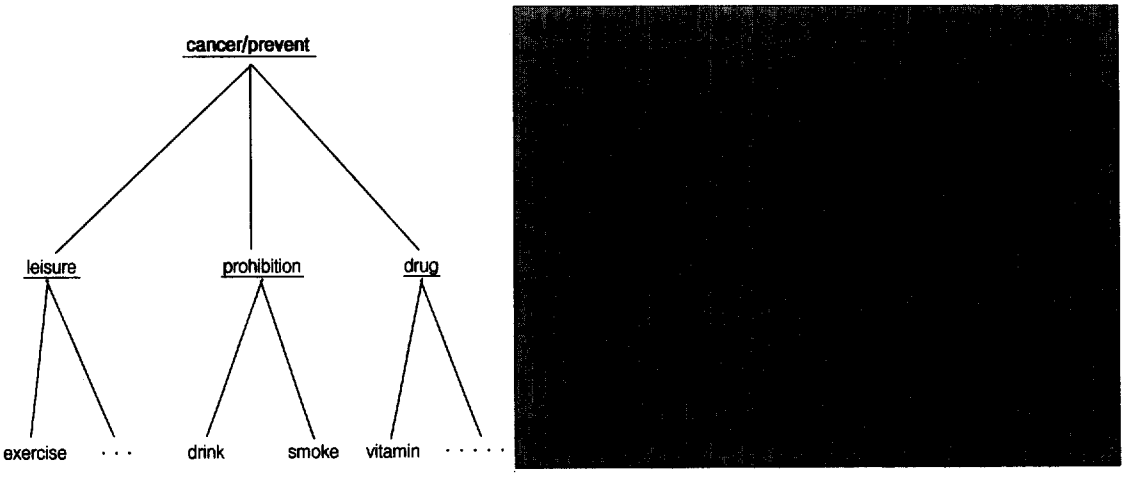
2단계 실험에서는 앞에서 생성된 사용자 의도 전개 트리를 이용하여 학습 예제 문서에서 사용자의 의도를 추출하는 단계로서 추출과정은 아래와 같다.

추출된 상위 개념 <표 3>은 사용자 의도 예제 생성시 그 문서의 대표 키워드가 된다.

D1 {prevent, treatment, cancer, diagnose, gastric, isbn, ···} *none*
D2 {diet, drug, prevent, smoke, screen, cancer, ***diet***, ···}
← **prohibition**
D3 {cerreal, diet, esophagu, fiber, ***health***, protect, h, ···}
← **leasure**
D4 {associ, cancer, chanc, decreas, ***tour***, factor, gastric, ···}
← **leisure**
D5 {agent, cohort, ***drug***, diet, ***exercise***, ***medicine***, smoke, ···}
← **drug**
D6 {drank, drink, tea, research, ***cholesterol***, coffee, tea, ···}
← **drug**
D7 {chanc, decreas, gastric, ***smoke***, stomake, ···} ← **prohibition**
D8 {diet, evid, prevent, risk, supplement, *guot*, *metaplasa*, ···} *none*
D9 {fear, doctor, ***pain***, ***stress***, ***worry***, persist stomach, ···}
← **prohibition**
D10{food, reduc, ***smoke***, treatment, ***tobacco***, ***prevent***, ···}
← **prohibition**

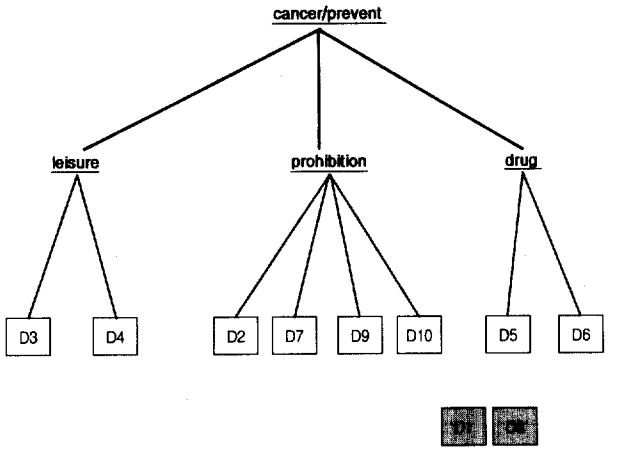
〈표 3〉 regression table

words	intention
{excercise, tour, health, ···}	{leisure}
{medicine, colesterol, tea, ···}	{drug & nutrition}
{stress, worri, smoke, tobacco, diet.}	{prohibition}
{running, vegetable, ···}	{diet}



(그림 11) 의도 전개트리와 학습예제 생성 : 1차예제

3단계 실험에서는 위에서 생성된 사용자의 의도 전개 트리를 이용하여 보정 작업을 거친 후 (그림 12)와 같이 사용자 의도 트리가 생성된다. 위 표에서 보여지는 바와 같이 "leisure", "prohibition", "drug"등은 의도가 전개된 "예방" 트리내에 존재하기 때문에 메이저의도로 선정되고 D2 문서의 경우 "diet"라는 개념은 "예방" 트리내에 존재하지 않으므로 다음의 서브의도인 "smoke"라는 단어로 바로 상위개념인 "prohibition"이 메이저 의도로 보정, 변경되는 것을 보여준다. 그리고 D1, D8의 경우는 의도전개 트리내에서 개념추출이 어려운 문서이므로 사용자의 의도 트리에는 구성되지 않으나 빈도수 확률에 의해 문서가 분류되는 과정을 볼 수 있다.



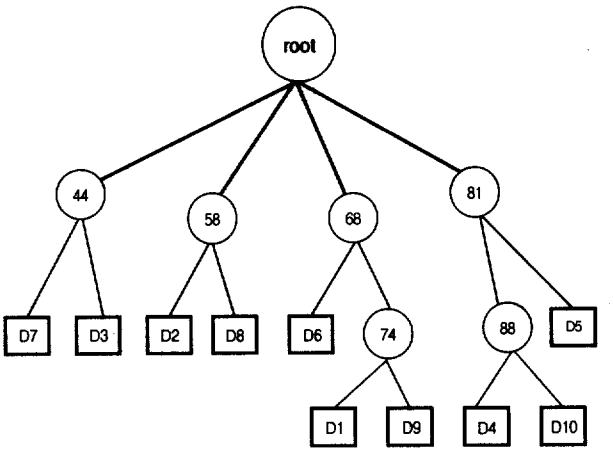
(그림 12) 사용자 의도트리 : 1차예제

사용자 의도 트리를 이용하여 기존에 분류되어진 카테고리와 병합하여 새로운 카테고리가 자동으로 재구성되는 것을 보여주어야 한다. 카테고리를 재구성하기 위해서 우선 cobweb에 의해 생성된 문서 분류 트리를 살펴보고 의도 트리가 병합될 때 변경, 조정되어야 하는 부분과 처리과정을 기술한다. 아래 그림은 cobweb에 의해 생성된 문서 분류 트리이다.

(그림 13)의 문서분류 트리를 사용자 의도 트리와 병합하여 카테고리를 재구성하기 위해서는 보정 및 변경 작업이 필요하다. 우선 (그림 12)의 사용자 의도 트리를 그룹으로 설정하고 그룹기준을 "깊이 3", "넓이 3"으로 설정한다. 이 기준은 고정값이 아니고 상황에 따라서 조정하여 적용한다.

빈도수확률에 의해 만들어진 분류 트리는 사용자 의도 트리에서 설정한 기준에 위배됨으로 변경, 조정이 필요하다.

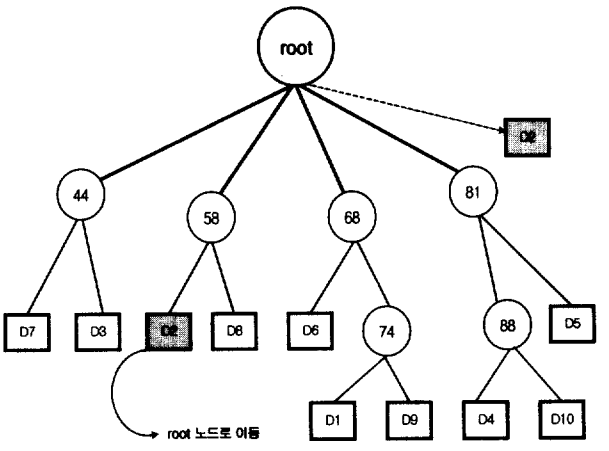
D = {{D3, D4} {D2, D7, D9, D10} {D5, D6}}을 빈도수확률의 문서분류 트리에 반영하면 변경, 조정 필요부분은 아래와 같다.



(그림 13) cobweb에 의한 문서분류트리 : 1차예제

① (그림 11)의 트리에서 D의 원소를 다 포함하는 공통부모 root의 자식노드는 4개이다.(넓이 제약기준 위배 : 재구성 필요(불필요한 측면으로 분류제거))
② {{D7, D3} {D2, D8} {D6{D1, D9}} {D5{D4, D10}}} 중 가장 적은 그룹 D2 선택
③ 문서 D2의 의도 트리내 그룹 탐색
④ D2그룹의 원소 D7, D9, D10 선정
⑤ D7, D9, D10 문서의 공통 부모노드 선정(root)
⑥ root 아래로 D2 이동
⑦ ④번에서 의도그룹 원소수 2일 때 원소관련 그룹내 바로 삽입시킨다.

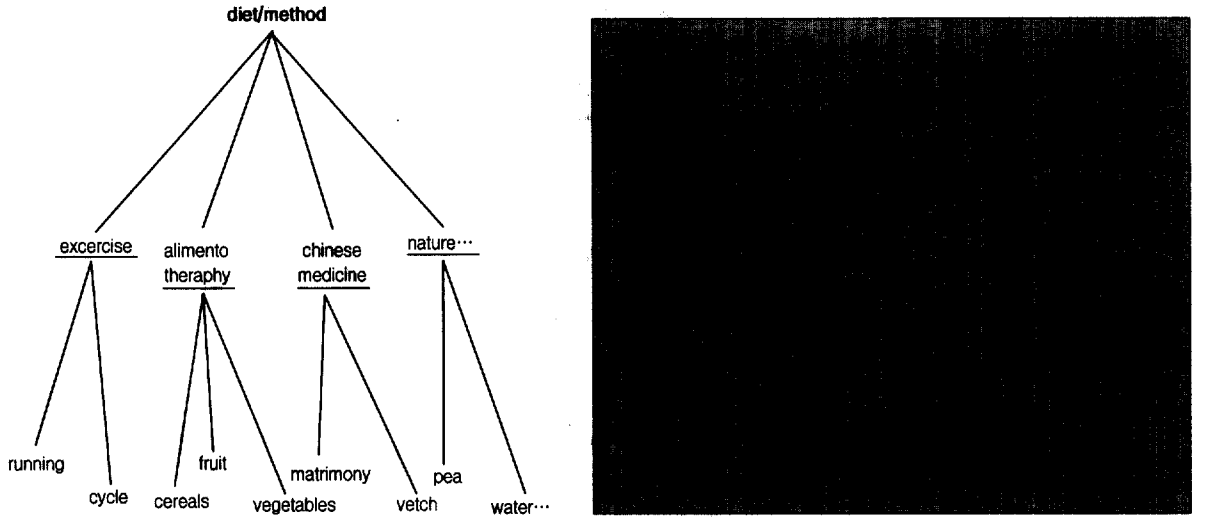
아래 그림은 사용자 의도 트리에 의해 변경, 조정된 분류 트리이다.



(그림 14) 문서분류 카테고리 재구성 : 1차예제

5.1.2 동적 카테고리 재구성

2차 실험에서는 사용자 의도 반영 외에 1차 실험에서 "암" 관련문서를 입력예제로 생성된 문서 분류 카테고리에 추가로



(그림 15) 의도전개 트리와 학습예제문서 : 2차예제

"다이어트" 관련문서들을 입력시켜 분류한다. 이때 두 부류의 입력 예제들은 각각의 그룹을 형성하게 된다. 하지만 단어 빈도수에 근거한 cobweb같은 시스템은 (그림 17)에서 보여지는 것처럼 e2의 다이어트 문서가 암 관련 문서 그룹으로 잘못 분류되는 과정을 보게된다. 이를 사용자 의도 트리 정보를 사용하여 카테고리의 보정 과정을 거쳐 재구성하는 과정을 실험하였다.

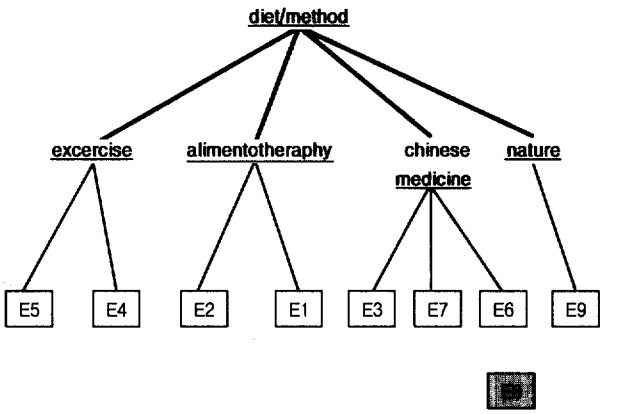
<표 2>에서 사용자(B)의 데이터 SET이 입력되어 1단계 과정에서 사용자의 의도 전개트리와 학습대상 전처리 과정인 HTML태그 제거, 불용어 제거, 스태밍처리, TF-IDF 알고리즘 등을 수행하고 생성된 학습예제 문서는 (그림 15)과 같다.

앞에서 생성된 사용자 의도 전개 트리를 이용하여 학습예제 문서에서 사용자의 의도가 추출된 상위개념은 1차 실험에서와 같이 사용자 의도 예제 생성시 그 문서의 대표 키워드가 된다. 사용자 의도 예제 문서는 다시 의도 전개 트리를 이용하여 보정 작업을 거친후 (그림 16)과 같이 사용자 의도 트리가 생성된다.

위에서 생성된 사용자 의도 트리는 (그림 17)에서 보여주는 cobweb에 의해 생성된 문서 분류 카테고리와의 병합으로 새로운 카테고리가 재구성된다. E8 문서의 경우는 의도 전개 트리내에서 개념 추출이 어려운 문서이므로 사용자의 의도 트리에는 구성되지 않으나 빈도수 확률 기법인 cobweb에 의해 문서가 분류되는 과정을 볼 수 있다.

아래 분류 트리는 "다이어트, 방법" 의도 트리의 제약조건을 "깊이 4", "넓이 4"로 설정하였다.

이를 보정, 변경하여 트리를 재구성하기 위해서는 문서 추가시 흐트러진 트리형태 (그림 17)를 사용자 의도 트리와 병합해서 문서 분류 카테고리를 재구성해주는 과정이 필요하다.



(그림 16) 사용자 의도 트리 : 2차예제

두 의도의 보정, 변경으로 재구성되는 과정은 다음과 같다.

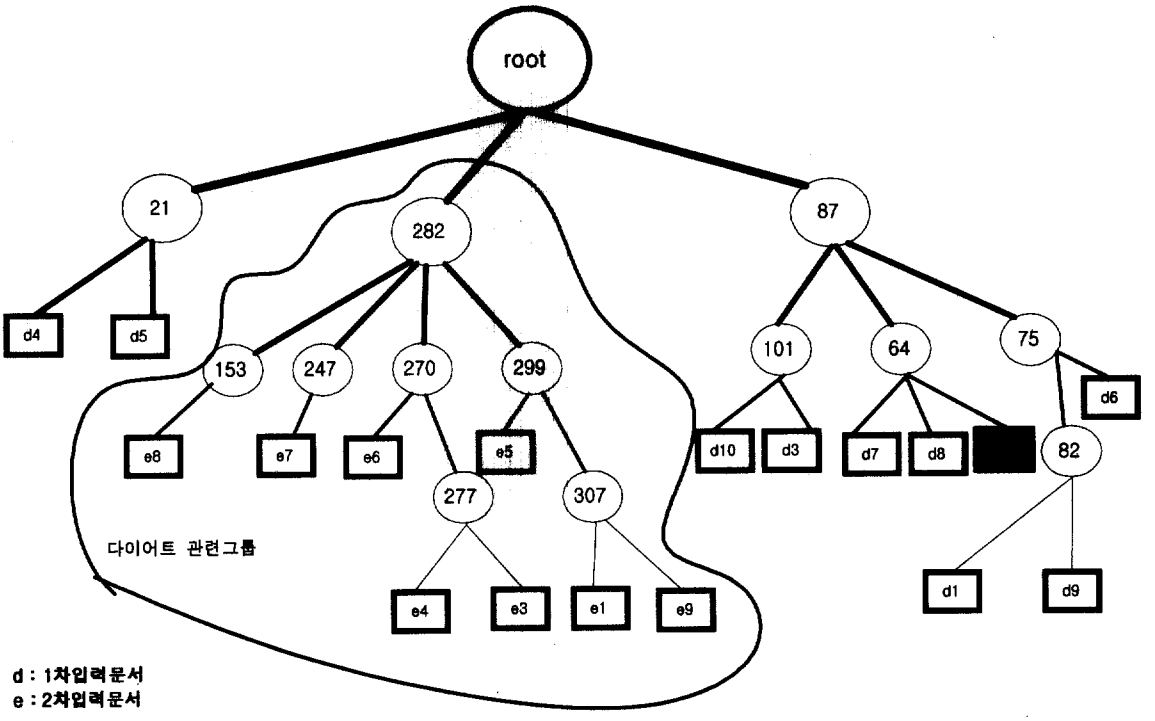
① 중심축 구하기

"다이어트, 방법" 의도 관련 문서를 그룹중에서 최다수 문서를 보유한 그룹을 "축"으로 선정한다.

② 소그룹을 중심축 그룹에 병합

- 소그룹내 원소와 의도 트리내에서 같은 그룹인 원소들을 찾는다.
- 의도 트리내 같은 그룹이었던 원소들의 공통부모 노드를 찾는다.
- 공통 부모노드에 소그룹 원소를 연결시킨다.

③ 구성원이 둘이었던 의도 트리내 그룹은 소그룹 원소를 바로 의도



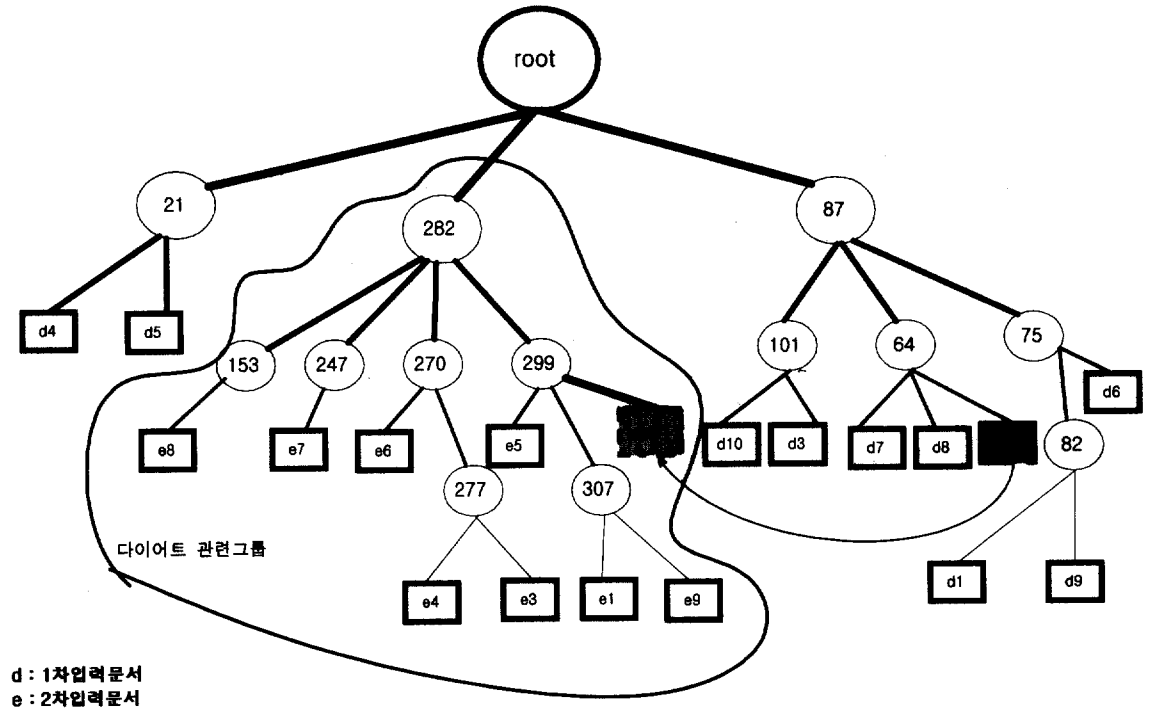
(그림 17) cobweb에 의한 문서분류트리

트리와 같은 그룹 원소의 그룹에 연결시킨다.

④ 이같은 과정을 의도 트리 생성 그룹수만큼 반복한다.

다음 그림은 의도 트리와 병합해서 카테고리가 재구성된 후의 문서 분류 트리이다.

위 그림에서 보듯이 cobweb에서는 1차에 실험했던 암 관련 카테고리에 다이어트 관련 문서인 e2 문서가 포함되어 있었으나 사용자 의도 트리를 이용하여 카테고리 재구성을 수행한 결과, e2 문서가 다이어트 문서 그룹으로 재구성되는 것을 볼 수 있다.



(그림 18) 문서 분류트리의 재구성 : 2차예제

6. 결 론

본 연구는 웹 상에서 사용자의 검색 의도에 적합한 문서를 찾는데 목적을 두고 설계되고 실험되었다. 접근 방법은 확률 및 통계를 사용한 문서 특징 추출 방법과 사용자 의도 정보 지식을 사용한 문서특징 추출 방법을 혼합한 통합적인 웹 문서 분류 방법을 제안하였다.

이 접근 방법은 확률 및 통계적 접근 방법에서 어려웠던 문서들 사이에 의미를 부여하는 부분에서 지식베이스의 지식을 사용하여 사용자의 의도가 반영된 문서 분류가 이루어지도록 하였다.

이 방법의 성과는 크게 다음과 같다. 첫째, "의도"에 관련된 지식베이스 정보를 사용함으로서 사용자 의도를 표현하는 넓은 의미의 키워드에서 관련된 세부의도들의 키워드들을 추출하여 문서들의 분류가 정확하고 융통성 있게 이루어지는 효과를 가져왔다.

둘째, 의도에 관련된 키워드 그룹, 즉, 세부의도들 그룹간의 관계를 구조적으로 분석하여 이 정보가 기존의 확률 및 통계를 사용한 문서분류 카테고리를 재구성하여 사용자의 검색 의도에 따르는 문서분류 능력을 향상시켰다.

셋째, 단어의 빈도수를 근간으로 하는 cobweb 시스템은 입력 문서의 순서에 종속적이므로 동일 문서가 입력 순서에 따라 다른 부류의 노드들에 포함될 수 있으며 모든 문서들이 하나의 루트노드 아래에 존재하는 단말 노드로 분류됨으로 특정 문서의 카테고리 확인이 어려웠다. 또한, 단일 키워드로 사용자의 관심과 의도를 반영하기가 어려웠으나 이들의 보정 및 그룹화로 의도 반영이 용이하게 되었다.

앞으로 웹 문서의 특징추출 및 사용자 의도 반영을 위한 분야에서 D-car 시스템은 다음과 같은 확장을 시도하고 있다. 기존에 개발되어 운영중인 지식베이스나 개념 시소러스 등에서 사용자 의도를 표현할 수 있는 방법 등을 연구하고 그룹내 키워드 구조를 이용한 사용자 의도분석 방법, 키워드에서 의도를 파악하기 위한 추론방법, 그리고 사용자의 누적된 프로 파일에서 의도정보를 추출하는 방법 등이다. 특히 계층적으로 구성된 의도 트리내에서 지역별, 그룹별 의도 파악 기법 연구는 분산 및 객체기술과 연계되어 진행되어야 할 것으로 본다.

참 고 문 헌

[1] Jayanarayan Bhuyan, "Cluster-Based Adaptive Information Retrieval," Ph.D. dissertation, Univ of Southwestern Louisiana, 1990.

[2] W. Bruce Croft, "Clustering large files of documents using the single link method," JASIS 28, 6, pp.341-344, Nov. 1977.

[3] Webcatcher, http://plum.tuc.noao.edu/webcatcher/webcatcher.Html

[4] "wisewire," http://www.wisewire.com/http://home.wisewire.com/press/netscape.Html.

[5] Fisher, D. H., & Langley, P., "Methods of conceptual clustering and their relation to numerical taxonomy," in W. Gale, AI and statistics, reading MA : Addison Wesley, 1986.

[6] Doug. Fisher, "interactive optimization and simplification of Hierarchical clustering," AI access foundation and Morgan Kaufmann publishers, 1996.

[7] M. Wooldridge, N. R. jennings "Agent Theories, Architectures and language," Intelligent Agent, Springer Verlag, pp.1-39, 1994.

[8] M. P. Georgeff, A. S. Rao "The semantics of intention for rational Agents," IJACI-95, pp.710-804, 1995.

[9] Harman, D. "How Effective is Suffixing?," Journal of the American Society for Information science, 1991.

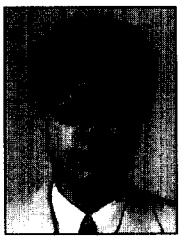
[10] Buckely C., Salton G. "Improving Retrieval Performance by Relevance Feedback," Journal of the American Society for Information science, 1990.

[11] Joachims T. "A Probabilistic Analysis of the Rocchio Algorithm with TF-IDF for Text Categorization," March 1996.

[12] E. Werner "A Unified View of Information, Intention and Ability," Decetralized AI2, pp.109-125, 1990.

[13] Gluck, M., & Corter, J., "Information Uncertainty and the Utility of Categories," Proceeding of the Seventh Annual Conference of the Cognitive Science Society, pp.283-287. 1985.

[14] Fisher, D. H.,"Knowledge Acquisition via Incremental Conceptual Clustering," Machine Learning, 1987.



김 효 래

e-mail : hrsaeg@kyungmin.ac.kr

1986년 서울산업대학교 전자계산학과 졸업 (공학사)

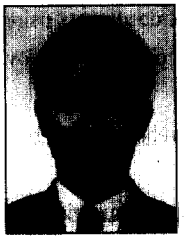
1991년 건국대학교 대학원 컴퓨터공학과 (공학석사)

1996년 건국대학교 대학원 컴퓨터공학과 (박사과정수료)

1985년~1997년 현대투신증권 정보기술부 근무

1998년~현재 경민대학 컴퓨터정보통신학부 교수

관심분야 : Agent, 데이터마이닝, 인공지능



장 영 철

e-mail : jdear@kyungmin.ac.kr

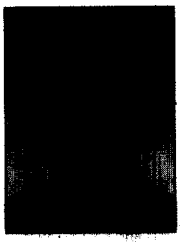
1987년 한양대학교 수학과 졸업(이학사)

1989년 건국대학교 대학원 컴퓨터공학과 (공학석사)

1998년 건국대학교 대학원 컴퓨터공학과 (공학박사)

1996년~현재 경민대학 컴퓨터정보통신학부 교수

관심분야 : Agent, 인공지능, 데이타마이닝



이 창 훈

e-mail : chlee@konkuk.ac.kr

1975년 연세대학교 수학과(이학사)

1977년 한국과학기술원 전산학과(이학석사)

1993년 한국과학기술원 전산학과(이학박사)

1980년~현재 건국대학교 컴퓨터공학과 교수

1996년~2000 건국대학교 서울캠퍼스 정보통신원 원장

2000년~현재 건국대학교 정보통신 대학원 원장

2001년~현재 건국대학교 인터넷 멀티미디어학과 학장

관심분야 : IDS, Agent, Data Mining, 인공지능