사용자 의도 정보를 사용한 웹문서 분류

장영철

경민대학 영상공연학부
480-702, 경기도 의정부시 가능3동
Tel: +82- 31-828-7350, Fax: +82- 31-828-7950, E-mail:jdear@kyungmin.ac.kr

Abstract

복잡한 시맨틱을 포함한 웹문서를 정확히 범주화하고 이 과정을 자동화하기 위해서는 인간의 지식체계를 수용할 수 있는 표준화, 지능화, 자동화된 문서표현 및 분류기술이 필요하다. 이를 위해 키워드 빈도수, 문서내 키워드들의 관련성, 시소러스의 활용, 확률기법 적용 등에 사용자 의도(intention) 정보를 활용한 범주화와 조정 프로세스를 도입하였다.
웹문서 분류과정에서 시소러스 등을 사용하는 지식베이스 문서분류와 비감독 학습을 하는 사전 지식체계(a priori)가 없는 유사성 문서분류 방법에 의도정보를 사용할 수 있도록 기반체계를 설계하였고 다시 이 두 방법의 차이는 Hybrid조정프로세스에서 조정하였다.
본 연구에서 설계된 HDCI(Hybrid Document Classification with Intention) 모델은 위의 웹문서 분류과정과 이를 제어 및 보조하는 사용자 의도 분석과정으로 구성되어 있다. 의도분석과정에 키워드와 함께 제공된 사용자 의도는 도메인 지식(domain knowledge)을 이용하여 의도간 계층트리(intention hierarchy tree)를 구성하고 이는 문서분류시 제약(constraint) 또는 가이드의 역할로 사용자 의도 프로파일(profile) 또는 문서 특성 대표 키워드를 추출하게 된다. HDCI 는 문서간 유사성에 근거한 상향식(bottom-up) 의 확률적인 접근에서 통제 및 안내의 역할을 수행하고 지식베이스 (시소러스) 접근 방식에서 다양성에 한계가 있는 키워들 간 관계설정의 정확도를 높인다.

Keywords: document classification, clustering, intention, thesaurus, similarity, semantic weight, EBL

1. 서 론

웹문서 분류는 인간의 생각과 표현의 거리를 IT 기술로 좁히는 영역이다. 하지만 복잡한 인간의 지식체계와 부정확한 표현 환경하에서는 최적의 분류는 어려운 실정이다. 문서 분류 시 클래스의 기준을 미리 정하고 이를 찾는 접근(classification)과 초기에 제시된 기준 없이 문서간의 유사성에 (similarity)에 근거하여 분류하는 군집화(clustering) 접근이 시도되고 있다. 이를 구현하는 기술에는 관련된 키워드들의 유형에(빈도 등) 근거한 분류(TF*IDF, 베이지안), 인간의 표현 방식과 단어간의 관계가 들어있는 사전을 이용하는 방법(온톨로지, 시소로스) 등이 있다.[10] 또 인공지능적 처리기술을 사용하는 에이전트 방식은 사용자의 웹상 행동과 관심 문서에 대한 정보를 수집하고 이를 분석/학습하여 관심문서에 관련된 프로파일(keyword) 등을 제공한다.(Infofinder, Webwatcher 등)[3, 7]
인간이 사용하는 모든 표현을 저장하여 지식베이스로 활용하는 접근은 비용과 정확도의 한계가 존재하고 확률 기법을 사용하여 문서내의 유사성을 결정하는 방법도 단순히 관련 단어의 출현 빈도수, 사용자의 관심여부를 입력하는 등 충분히 체계적이지 못하다.
본 연구에서는 기존의 접근과 달리 문서분류를 위해 EBL(Explanation Based Learning)의 기법을[8] 응용해 먼저 사용자의 검색의도를 구조적으로 분석하고 문서내의 관련 단어선정 및 지식베이스 적용방법에 활용한다. 표현이 모호한 사용자 검색의도가 다양한 특징의 문서들 내에서 여러 표현(generalization, specialization 등)으로 적용될 수 있도록 기존의 문서분류 기법들을 수정 융합한 Hybrid 조정프로세스를 포함한 HDIC를 설계하였다.

2. 관련연구

2.1 EBL

EBL(Explanation Based Learning)은 한 예제만 있어도 도메인 지식(domain knowledge)을 이용하여 개념학습을 할 수 있는 학습방법이다. EBL 은 사용자가 제시한 의도가 다양한 세만틱을 수용할 수 있도록 구조화시키고 변화시킨 의도트리를 생성하게 되고 의도관련 여러 표현을 포함하게 된다. 특히 같은 키워드도 웹문서의 표현 및 관계에 따라 다양한 해석이 가능한 환경에 적합도록 운용성(operationalize) 변화 특성이 있다. 의도트리 내 주의도, 부의도들과 관계 값들은 특정 임계값(threshold)으로 사용함으로써 키워드 선정의

제약, 지식베이스의 적용 한계를 극복하는 문서분류가 가능하게 된다.
EBL 기법이 사용되어 생성된 의도트리는 일반화(generalization), 세분화(specialization)의 전이과정을 수행하여 다양한 형태의 관련된 의도들을 생성한다.

2.2 유사성에 근거한 문서분류

일반적인 문서분류 방법은 1) 각 문서를 대표하는 키워드(keywords)들의 추출과정 즉, 색인추출과정, 2) 문서분류과정으로 나뉘어진다. 색인추출과정에서는 불용어 제거, 어근추출, TF*IDF, 벡터길이 정규화 알고리즘 등이 사용된다. 문서분류과정에서는 단어의 연관성계산, 연관 테이블 구축, 프로파일 생성(사용자 관심과 연관된 단어), 문서분류가 이루어진다.[4]

문서의 유사성을 찾는 학습과정은 컨텍스트(context)에 영향을 받으며 어떤 목표(goal), 목적(purpose), 의도(intention) 등을 고려하여 생성된다. 효율적으로 형태적, 의미적(semantic) 유사성을 찾기 위해서는 대상 키워드들의 단순한 거리만이 아닌 사용자 의도 정보 등을 적극 활용해야 한다.[5]
문서분류시 고려사항은 1)문서내 속성, 관계, 이들간의 가중치 2) 표현형식 3) 분류기준(의도 고려) 4) 알고리즘의 시간과 공간의 한계 등이며 유사도(similarity),연관관계(association), 적합도(fitness) 등을 이용하여 한계를 극복한다.

2.3 지식베이스에 근거한 문서분류

정보의 복잡도가 증가함에 따라 단어들의 관계를 문서내의 의미 관점에서 분석해야 하고 이를 위해 동의어, 반의어, 상하위 포함관계 등을 기록한 온돌로지 체계와 시소러스 등을 이용하고 있다. [9] 이러한 접근은 그 동안 어려웠던 멀티미디어 데이터 내 시간 공간 표현문제, 상황인식, 개인화된 정보분류 등 인간의 사고와 비슷한 분류가 가능하게 되었다. 하지만 방대한 자료의 입력, 관리비용 등으로 인해 일반화되지 못하고 특정영역에 국한되어 사용되고 있다.
적은 양의 영역지식 또는 시소러스 등을 사용하여 사용자의 의도를 정확히 분석하고 이를 운용 정보로 사용함으로써 기존의 유사성에 근거한 문서분류, 지식베이스에 근거한 문서분류에서 능동적이고 효율적인 문서분류가 이루어진다.

3. 의도 사용을 위한 기반 구조

3.1 의도 트리(intention tree)

웹스터 사전에서 의도는 "명확하고 고의적인 형식화", 목적(purpose)은 의도보다 더 많이 정해서 결정한 것으로 설명된다. 의도는 형식안에서 상황에 따라 변화가 가능한 의미를 내포하고 있다.
의도는 사건에 대한 믿음(belief) 과 욕구(desire)의 요소들로 표현될 수 있으며[8] 사용자의 검색의도도 이 같은 관련성에 근거한 표현, 상하위 포함관계 등으로 표현할 수 있다. "비가 올것을 믿으니 우산을 가지고 간다", "대학에 들어가기를 욕망하기에 열심히 공부한다." 는 의도가 믿음, 욕구의 형태로 표현된 예이다.
문서분류시 사용자 의도를 표현하는 키워드는 지식베이스, 시소러스를 이용하여 확장 분석되어 의도트리를 구성한다.
다음은 의료/건강 분야 지식베이스의 예를 보여준다.

표1 - 지식베이스 구조 예

```
Health and Medicine
 disease(cancer(stomach cancer
        ( prevention(leisure(sport or exercise,
                        tour or trip ...));
            (prohibition(drink, smoke ..))
            (drug(vitamin, diet pills ..));
        (treatment(radiobiology));
            (herb(acupuncture ...));
            (drug(...));
       (liver cancer);
         (prevention);
         (treatment);
       (diabetes);
            ...
  diet(method
      (exercise(running, cycle . . .));
      (alimentotheraphy or diet cure
                  (fruit( apple, grape ...))
                  (vegetables(cucumber.
                        potato, carrot ...));
       (diet consultation);
```

이 지식베이스 내 의도들은 EBL의 룰 적용과 regression 과정을 거치면 상위의도와 하위의도의 계층구조로 표현된 트리를 생성한다.
이 같은 주의도, 부의도(sub-intention)의 계층화된 구조는 키워드들간의 관계 파악이 쉽고 정교한 분류 영역 설정이 가능하다.
사용자는 검색시 다음과 같은 (키워드 : 의도) 조합형태를 입력으로 사용된다.

예)
(disease : prevention), (dish : decoration),

(saw, carpentry)

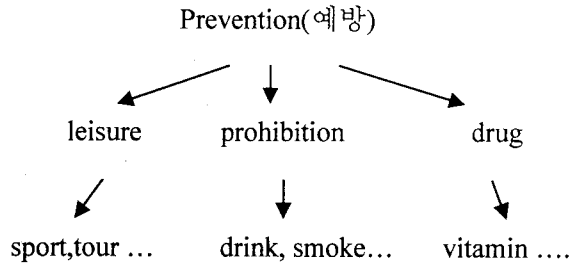

그림 1 - 의도 트리

3.2 의도사용 유사성 문서 분류 구조

비감독 학습형태의 유사성 문서분류 구조는 검색시 다음의 (keyword : intention) 의 형태로 입력이 주어지고 intention 은 <Int i> | → [subInt-1, subint-2, […, subInt-k]] 의 형태로 변형된다. 유사성에 근거하여 문서들 내에서 이들 keyword, Int, subInt-k 들과의 관련성을 분석할 수 있는 구조는 다음과 같다.
전체 문서에 출현한 단어의 집합 중에 n번째 단어와 m번째 단어의 연관성은 다음과 같이 계산한다.

$$Rnm = -ln \qquad (1)$$

이는 우연히 두 단어가 문서에서 중복되어 나타날 확률이며 확률이 낮을수록 두 단어는 서로 연관성이 높은 관계를 의미한다. *"한 단어가 10개 문서 중 6개 문서에서 사용되었고 또 다른 단어는 10개의 문서 중 3개의 문서에서 사용되었다. 이 때 3개 문서에서 위 두 단어가 동시에 사용되었다고 가정하자."* 아무 연관도 없는 단어들 사이에 이런 일이 발생할 확률은 다음과 같다.

$$(\frac{6}{10}\times\frac{3}{10})^3\times(1-\frac{6}{10}\times\frac{3}{10})^7\times\frac{10!}{7!\times3!} \qquad (2)$$

이를 일반화 시킨 결과는 다음과 같다.

$$P_{wc}^{Dc}\times(1-P_{wc})^{D-Dc}\times {}_DC_{Dc} \qquad (3)$$

Pwc 는 하나의 문서가 두 단어를 동시에 포함할 확률로서 식 (3) 과 같다. 이때 Dc는 두 단어가 동시에 사용된 문서의 수이며 D는 전체 문서의 숫자이다.

$$P_{wc} = P_{wn} \times P_{wm} \qquad (4)$$

Pwn 은 전체 문서 중 n번째 단어가 포함된 문서의 비율이며 Pwm는 m번째 단어가 포함된 비율이다. 이를 기반으로 n번째 단어와 m번째 단어간의 연관성을 추정하면 식 (5)와 같다.

$$R_{nm}=-\ln(P_{wc}^{Dc}\times(1-P_{wc})^{D-Dc}\times {}_DC_{Dc}) \qquad (5)$$

표2 - 키워드와 의도들 간의 관계

	keyword	intention	sub-int
keyword			
intention	3.1		
sub-int	1.7	4.1	

표1 에서 문서내의 출현 빈도 확률에 의하면 keyword는 주 intention과 높은 관계성을 가지고 있으며 의도트리의 여러 subIntention으로 확장해 의도의 다양한 변형된 형태로 문서를 분류할 수 있다.

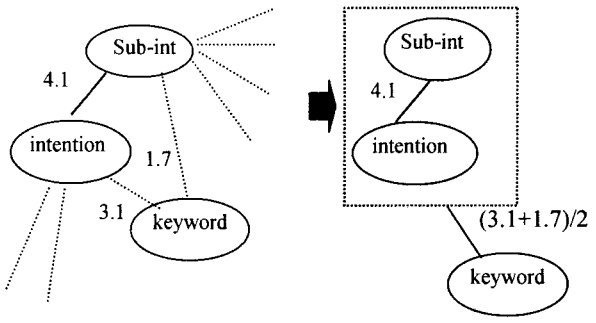

그림 2 - 병합을 통한 그룹의도와의 관계

3.3 의도 사용 지식베이스 문서분류 구조

지식베이스, 시소러스를 이용하는 세만틱을 고려한 문서 분류에서 의도 정보를 사용하기 위해서는 의도트리내의 구조화된 정보가 적용될 수 있도록 지식베이스 내 룰(rule), 표현 등이 다양한 관계,

가중치 등을 표현할 수 있도록 구조화되어야 한다. [1, 6] 문서 내에서 상/하위관계, 동의관계, 동위관계 등이 고려된 키워드간 세만틱 관계가 형성되어야 한다. 전체 프로세스는 다음과 같다.

1) 감독학습으로 시소러스를 사용하여 문서의 의미를 나타내는 여러 키워드 그룹을 선정
2) 각 그룹내 세만틱을 고려한 가중치 계산 (의도트리내 정보 반영)
3) 최적 그룹내 대표 주제어 선정.

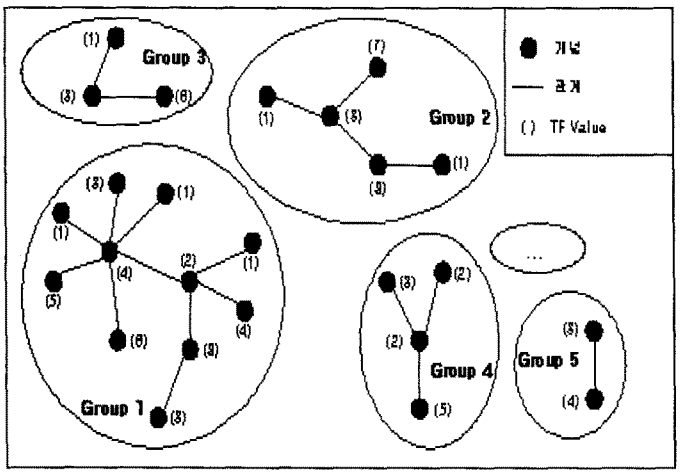

그림 3 – 키워드(프로파일) 그룹핑

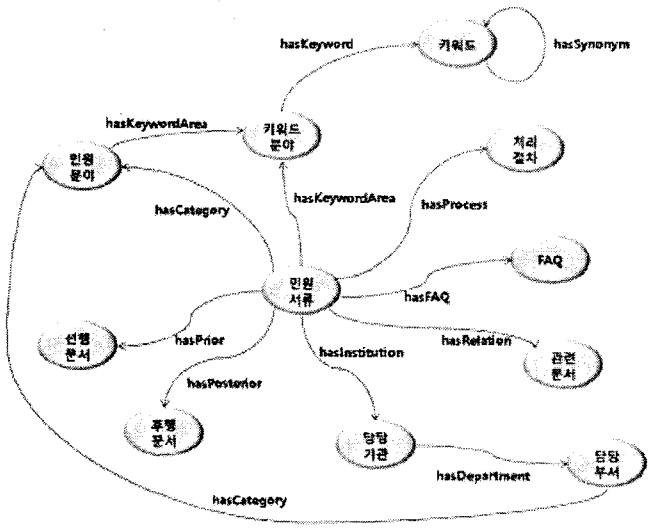

그림 4 - 의미를 고려한 온톨로지 설계 사례

다음의 연산자를 사용하여 의도 정보를 사용한 세만틱 키워드 그룹들의 조정이 이루어 진다.

Composition
세만틱 가중치(semantic weight)의 증가를 위해 키워드 또는 키워드 그룹을 병합한다.

Dconposition
세만틱 가중치(semantic weight)가 임계치 이하일 때 키워드 또는 키워드 그룹을 분리한다.

Generalized-Subject
빈도수는 적으나 의도정보나 문서내 관련된 다수의 세만틱 키워드들을 대표하여 상위레벨의 키워드가 주제어 또는 프로파일로 선정된다.

Specialized-Subject
의도 정보의 정해진 목적(임계치)에 부합하는 세분화된 키워드가 주제어로 선정된다

4. HDCI

4.1 HDCI와 Hybrid 조정프로세스

인간의 생각과 표현의 차이는 크며 이를 줄이기 위해서는 사용자의 검색의도를 충분히 지능적으로 분석하는 것이 필요하다.
HDCI(Hybrid Document Classification with Intention)에서 문서내의 일반적인 단어들의 관계를 넘어서 사용자의 의도를 정확하게 분석한 후 이 정보를 활용하여 기존의 문서분류 방법과 융합한 범주화가 이루어진다.
HDCI는 크게 사용자 ***의도정보 분석과정***, ***웹문서 분류과정***으로 구성되어 있다. 웹문서 분류과정은 다시 ***지식베이스 문서분류***, ***유사성 문서분류***, ***Hybrid 조정프로세스***의 3단계로 구성된다.
의도 정보는 지식베이스 문서분류, 유사성 문서분류, Hybrid 조정프로세스에서 사용된다. 앞장에서 설계된 기반구조 위에서 의도트리(intention tree) 내의 관련 의도, 부의도(subIntention), 이들간의 관계를 이용하여 지식베이스 문서분류는 인간이 표현한 복잡한 의미(semantic, keyword)들과 이들간의 관계를 사용자 중심으로 가중치를 부과하여 (topdown 제어) 대표 주제어를 선정할 수 있다. 또 유사성 문서분류에서는 상향식 일반화(bottomup generalization)를 통하여 관련 의도그룹과 키워드 간 관계의 경중을 고려하여 변화에 능동적인 프로파일(keyword)이 선정된다.

지식베이스 특성상의 분류기준과 확률에 의한 유사성을 이용한 분류기준과의 차이는 Hybrid 조정프로세스에서 조정된다. 과도한 지식사용으로 인한 편향된 분류는 문서내 출현된 단어관계와 빈도수를 근거로 조정되고 세만틱이 잘못 해석된

빈도수 만을 고려한 편향성은 시소러스 기반구조와 의도트리 정보로 조정된다.
범위(scope), 제약조건(constraint), 가중치(weight), 순서(order) 등으로 제시된 의도정보는 문서 특성을 Hybrid조정프로세스에서 다음의 연산자들을 사용하여 조정한다.

Relaxation
의도트리를 기준으로 편향된 기준을 완화한다.

Logrolling
문류방법의 특성상 강점이 있는 기준과 의도트리 정보를 기준으로 중요 기준들은 상호 선택하고 부족한 정보의 결정, 과도한 inductive .leap 이루어진 분류는 배제한다.

Extending
한 분류 기준을 확장하여 분류범위를 확장한다.

Restriction
의도트리에 근거하여 포괄적인 기준 및 영역을 제한한다.

CompositBridge
두 분류 기준을 통합 합성하여 새 기준을 생성한다. 여러 조건의 단계적 연결(bridge) 효과를 가져온다.

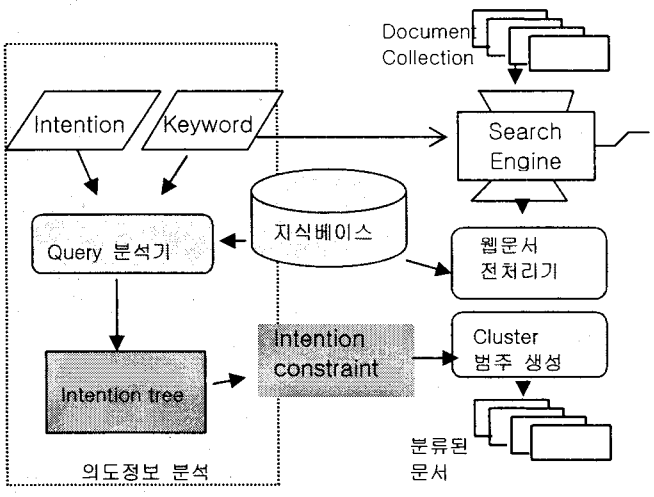

그림 5 – HDCI 구성도

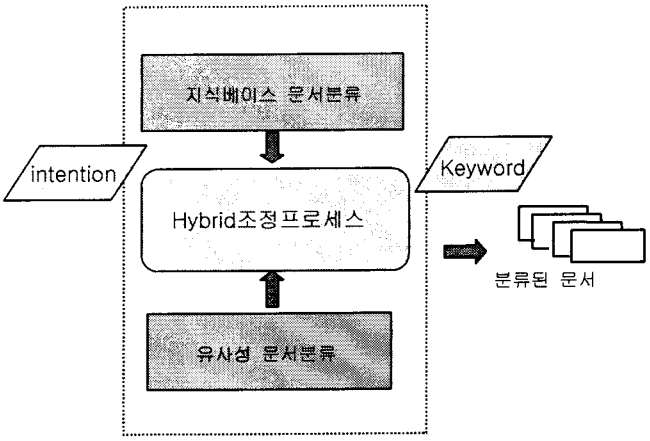

그림 6 - Hybrid 조정 프로세스

4.2 HDCI 내 의도 구조

의도트리 구조내의 sI_i는 범위, 제약조건, 순서(order), 가중치 등을 나타내는 일반화된 개념이다. [LB, UB], { c_1 U c_2 U . . . U c_n } 같은 예이다. 다음에서 다음에서 *Intention-ma i n*은 여러 *subIntention*으로 구조화된다.

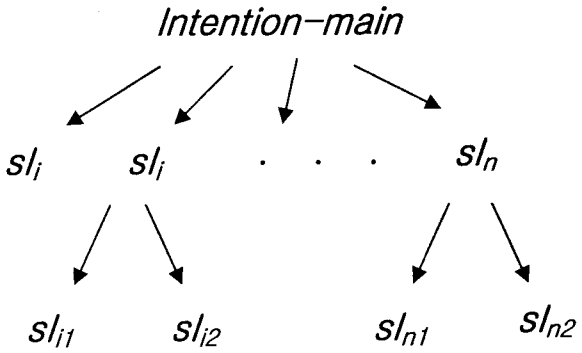

그림 7 - HDCI 의도트리

[Relationship]
< ***Int*** > : : + <leafword> | <non-leafword>
| {< ***Int*** > . . . < ***Int*** >}

[Order]
< Int_i > | → [Int_l, Int_m, Int_n]

Int_i 는 Int_l, Int_m, Int_n 가 처리된 후 얻는다.

[Transformation Rule]
< ***Belief*** > : : = < Int_i > | → < Int_j >

Int_i 는 Int_j 의 변형된 형태이며 특정 신뢰도가 임계치를 상위할 때 얻어진다.

```
Func TransForm ( Int, k )
{ While ( transformation rule : Belief > k )
        { IntTREE = NewInt
                TransFORM( Int, k )
        }
}
```

Int : *Intention*
k : 신뢰도를 위한 정해진 임계치
IntTREE: *지금까지 전개된 intention tree*

5. 결 론

웹문서 분류는 인간의 생각과 표현의 거리를 IT 기술로 좁히는 과정이다. 복잡한 인간의 지식체계, 다량의 문서 및 표현 형식으로 인해 지능화, 자동화가 시급히 이루어져야 할 영역이다.
본 연구에서는 사용자의 검색의도 정보를 효과적으로 표현하고 구조화하여 기존의 주요 웹문서 분류방법과 융합하는 모델을 제시하였다.

사용자의 검색 키워드와 의도를 한 쌍의 입력으로 받아 지식베이스의 도움으로 검색의도를 주의도와 관련 부의도(subIntention)들의 트리형태로 구성하여 문서분류의 성능을 향상시키게 된다. 트리형태로 된 의도 그룹은 지식베이스, 유사성을 근간으로 삼는 기존의 문서분류 방법의 지식 편향성, 세만틱의 깊은 의미 반영시 미숙함을 보완하게 된다.

제안된 HDCI에는 크게 *사용자 의도 분석과정*, ***웹문서 분류과정*** 으로 설계되어 있다. 웹문서 분류과정은 *유사성 문서분류*, *지식베이스 문서분류*, ***Hybrid조정프로세스***의 3 단계로 구성되어 있다.
사용자 의도분석과정의 결과인 의도트리(intention tree) 는 웹문서 분류과정에서 범위, 제약, 순서, 가중치 등에 영향을 주어 기존 분류시스템의 프로세스 및 결정알고리즘이 개선되고 조정되어 사용자의 의도에 적합한 웹문서를 분류하게 된다.
기존의 웹문서 분류단계에서 의도 그룹군의 정보를 수용하기 위해서 구조 및 체계가 설계되었다.
유사성 문서분류 단계에서는 의도 그룹군의 주의도와 부의도들의 연관성 조사 체계, 의도 그룹군의 합병 및 분할에 따른 키워드와 의도그룹별 연관성을 계산하는 지식 구조와 체계가 설계되었다.

지식베이스 문서분류 단계에서는 기존 지식베이스의 체계의 세만틱 정확성의 한계를 극복하기 위해 의도트리내의 요소들과 문서내의 세만틱 가중치(semantic weight)를 표현하고 계산하는 체계를 설계하였다.

Hybrid 문서분류 단계에서는 하향식의 지식베이스 문서분류와 상향식의 유사성 문서분류의 차이가 최종적으로 통합 보정되었다. 의도트리의 분석에 근거하여 ***Relaxation, Logrolling, Extending, Restriction, CompositBridge*** 의 연산자를 사용하여 조정이 이루어진다.
HDCI는 의도정보의 분석 및 구조화에 따라 상향 및 하향 문서분류 방식이 동조화할 수 있도록 기반 표현 및 처리과정을 제시한 특성이 있으며 여러 문서분류 방법을 통합할 수 있는 프레임웍을 제공하고 있다. 특히 인간 지식표현의 복잡성과 세만틱의 다양성을수용하는 EBL 학습기법을 활용한 의도정보 활용은 HDCI의 장점이다.

참고문헌

[1] Berners-Lee, T., Hendler, J., Lassila, O.(2001) "The Semantic Web" . Scientific American
[2] Buckley, C., Salton, G., Allan, J.(1994.) "The Effect of Adding Relevance Information in a Relevance Feedback Environment", In Proc. 17th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 292-298
[3] E. M. Voorhees, (1986) "The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval", Doctoral Dissertation, Cornell University, Ithaca, NY.
[4] J.S. Shin, J.H. Kwak and C.H. Lee,(1999) "Automatic Classification of Web Documents with Word Accordance of Degree using Probability Model", Proceedings of ICOIN 13, Jan., pp. 6A-3.1-6A-3.4
[5] Kofod-Petersen, A., Cassens, J.,(2006) "Using Activity Theory to Model Context Awareness", Modeling and Retrieval of Context: MRC 2005, volume 3946 of LNCS, pp.1-17, Edinburgh, Springer
[6] Kong, H.J., Hwang, M.G., Hwang, H.S., Shim, J.H., Kim, P.K.,(2006) "Topic Selection of Web Document Using Specific Domain Ontology", MICAI2006, LNAI4293, pp.1047-1056.
[7] Krulwich, B., Burkey, C., (1997) "The InfoFinder Agent: Learning User Interests through Heuristic Phrase Extraction", IEEE Expert/Intelligent Systems & Their Applications, Vol. 12, No. 5.
[8]M. Wooldridge, N. R. Jennings,(1994) "Agent Theories, Architectures and Language" Intelligent Agents, Springer Verlag, pp.1-39
[9] 고광섭, (2007) "의미기반 기술을 사용한 전자정부 정보시스템 활용성 향상방안에 관한 연구", 박사논문, 건국대학교
[10] 최옥경, 한상용,(2006) "자동화된 통합 프레임워크를 위한 시맨틱웹 기반의 정보 검색 시스템" , 정보처리학회논문지C, 제13-C권, 제1호