사용자 프로파일 구축을 위한 TDIDT기반 관심단어 추출기법

이선미 · 박영택
숭실대학교 컴퓨터학과
selly13@hanmail.net
park@computing.ssu.ac.kr

Attribute extract method based TDIDT for construction of user profile

Sun-Mi Lee, Young-Tack Park
Dept. of Computer Science, Soongsil Univ.

요 약

본 논문은 기존의 귀납적 결정 트리 방식에서의 문제점 개선을 통한 사용자 관심 프로파일 구축을 목적으로 한다. 특히, 사용자 관심 프로파일의 정확도 향상을 위한 속성 선택에 대한 연구에 초점을 맞추고 있다. 사용자의 관심, 비관심 문서를 대상으로 사용자 관심 키워드를 생성하고 이를 바탕으로 초기 문서들을 재표현한다. 재표현된 문서를 입력 집합으로 하여 기계학습을 진행한다. 본 논문의 의사 결정 트리 생성 알고리즘은 입력 집합을 클래스별로 가장 잘 나누는 속성을 선택하여 노드를 구성하는 면에서는 기존의 알고리즘과 같다. 그러나 기존의 의사 결정 트리 알고리즘에서는 hill-climbing방식을 사용함으로써 사용자의 관심을 나타내는 중요한 단어가 사용자 관심 프로파일에서 숨겨질 경우가 발생한다. 이를 최소화하기 위해 특징 추출을 통해 선택된 속성을 그대로 학습의 입력 데이터로 사용하는 것이 아니라 입력 데이터를 가장 잘 나누는 속성과 그 다음 속성을 대상으로 disjunctive 연산을 통해 새로운 속성을 생성하여 이것을 속성 집합에 포함시키고 이를 학습의 입력 데이터로 이용한다. 이와 같이 disjunctive operator를 이용하여 새로운 속성을 의사 결정 트리 형성 시 이용하면 사용자의 중요한 관심을 포함하는 의미 있는(semantic) 사용자 관심 프로파일 구축이 가능해지고, 사용자 관심 프로파일을 기반으로 사용자가 관심 있는 문서를 제공할 수 있는 개인화 서비스를 제공한다.

Keywords : 귀납적 기계 학습, 의사 결정 트리, 사용자 관심 프로파일

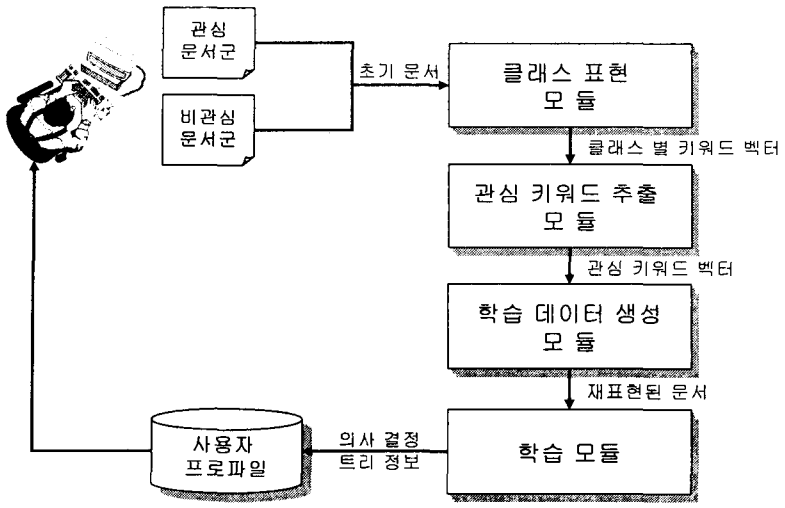
1. 서 론

인터넷을 통하여 제공되는 정보의 수가 기하급수적으로 증가함에 따라서 사용자 개개인이 원하는 정보를 정확하고 신속하게 검색하는데 많은 어려움을 가지고 있다. 이에 따라서, 인터넷에 산재해 있는 수많은 정보 중 사용자의 관심에 따라서 정보를 제공받을 수 있도록 하는 정보 여과 시스템에 대한 요구가 있어왔다.

최근 이러한 요구는 개인화(personalization)라는 방법을 통해 사용자에게 맞춰진 정보를 제공하는 형태로 연구가 진행되고 있으며, 인터넷을 이용하는 사용자의 프로파일(profile)을 작성하고 이를 기반으로 학습한 결과를 가지고 정보를 여과하여 제공하고 있다[1]. 이때, 사용자 관심 프로파일은 사용자의 선호도와 필요를 표현하는 것으로, 사용자 관심 프로파일의 정확도에 따라 사용자에게 제공되는 정보의 질이 좌우되는 것이다. 따라서 정확한 사용자 관심 프로파일 구축에 대한 연구가 요구되고 있다.

본 논문에서 사용자의 관심은 감독학습(supervised learning) 방법을 이용하기 위해 관심 문서 영역과 비관심문서 영역으로 구성하고, 두 문서군 간의 관계를 통하여 관심영역을 나타내는 키워드 벡터를 기반으로 표현한다[2]. 특징추출 과정을 통하여 관심 키워드를 추출하고 수정된 귀납적 결정 트리 방식을 이용하여 사용자별 관심 프로파

* 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.



[그림1] 시스템 구성도

일을 구축하는 TDIDT(Top-Down Induction of Decision Tree) 기반 관심 단어 추출 기법에 대하여 서술하고자 한다.

2장에서는 관련연구에 대하여 알아보고, 3장에서는 프로파일 구축 시스템을 구성하고 있는 각 모듈에 대하여 소개한다. 4장에서는 관심 문서군과 비관심 문서군을 바탕으로 특징추출과정을 통해 얻어진 관심 키워드 추출과 학습을 위한 초기 문서 재표현 과정에 대하여 소개한다. 5장에서는 재표현된 문서 집합을 입력 데이터로 하여 추출된 관심 키워드를 대상으로 수행되는 수정된 속성선택 알고리즘을 이용하는 귀납적 의사 결정 트리 방식과 그 결과로 구축된 사용자 관심 프로파일을 기존의 귀납적 의사 결정 트리 방식의 그것과 비교한다. 6장에서는 5장의 내용을 실질적인 실험을 통해 사용자 관심 프로파일의 정확도를 비교·실험 하고 마지막으로 7장에서 결론 및 향우 과제에 대하여 서술하여 맺는다.

2. 관련연구

사용자가 웹을 검색하는데 사용자 관심 프로파일을 이용하여 개인화된 정보를 제공하는 기술은 다방면에서 연구되고 있다. 그중 본 논문에서 구현하고자 하는 사용자 관심 프로파일 구축 시스템과 유사한 시스템에 대하여 알아보기로 한다.

앤더슨 컨설팅 연구실에서 만든 지능형 에이전트 시스템인 InfoFinder[3]는 사용자가 검색하는 문서와 사용자의 관심을 관측함으로서 사용자의 프로파일을 만든다. 이것은 본 논문과 같이 사용자의 직접적인 관심여부에 대한 정보를 입력받는 감독 학습 방식을 이용한다. 검색 문서들을 이용하여 ID3 귀납적 기계학습을 수행하여 추출된 관심 키워드는 의사 결정 트리의 형태로 만들어지며, 이러한 의사 결정 트리를 규칙(rule)의 형태로 표현하여 사용자의 관심 프로파일을 형성한다. InfoFinder의 단어 추출과정은 단순히 단어의 출현 빈도수에 의한 것이었으나, 본 논문에서는 관심영역과 비관심 영역의 단어를 대상으로 한 가중치 설정 기법을 통한 특징 추출작업을 적용하여 사용자 관심을 대표하는 단어의 정확도를 높였다.

다음은 본 시스템의 근간이 되는 C4.5[4]에 대한 설명이다. C4.5 알고리즘은 TDIDT를 위한 대표적인 알고리즘으로서, 여기서 TDIDT란 루트로부터 클래스에 의해 주어진 입력 예제를 가장 잘 나누어주는 속성을 선택하여 의사 결정 트리를 만듦으로 감독학습을 진행하는 방식을 말한다. 1993년 Quinlan에 의해 제안된 C4.5는 엔트로피, 즉 복잡도를 기반으로 하여 트리의 노드를 구성하는 알고리즘이다. 본 논문에서는 가장 좋은 속성 하나만을 고려하여 노드를 구성하는 hill-climbing 방식에서 벗어나 그 다음으로 좋은 속성, 즉 사용자의 관심을 나타내는 중요한 속성을 의사 결정 트리에 적용하기 위해 두 번째 속성까지를 고려해주는 새로운 속성을 속성집합에 추가하여 학습을 진행한다. 이로서 영역 분류에 커다란 비중을 가지는 속성으로 인하여 의사 결정 트리에 나타나지 않을 수 있는 사용자 관심 속성을 추출할 수 있게 한다.

3. 시스템 구성

본 논문에서 구현한 사용자 관심 프로파일 구축 시스템은 문서 전처리(preprocessing)를 통해 관심과 비관심 클래스를 표현하는 클래스 표현 모듈과 각 클래스를 대상으로 특징추출 과정을 통해 사용자 관심 키워드를 추출하는 관심 키워드 추출 모듈, 관심 키워드를 바탕으로 초기 문서를 재표현하여 학습에 필요한 입력데이터를 생성해주는 학습 데이터 생성 모듈, 마지막으로 기계학습 과정을 거쳐 사용자 관심 프로파일을 구축하는 학습 모듈로 구성되어 있다.

사용자는 먼저 자신의 관심 정보를 특정 관심 영역으로 생성한 후 관심 문서와 비관심·문서를 지정한

다. 문서 전처리 과정을 통해 사용자가 정의한 관심 영역과 비관심 영역을 구성하는 키워드를 추출하여 각 클래스를 키워드 벡터의 형식으로 구성한다. 관심 키워드 추출 모듈을 통해 score measure를 사용하여 각 관심 클래스를 표현하는 전체 키워드를 대상으로 가중치를 줌으로써 특징추출 과정이 진행되며, 이 과정을 거쳐 최종적으로 사용자의 관심을 표현하는 관심영역의 대표 벡터를 추출할 수 있다. 이것은 초기 문서를 학습에 필요한 입력 데이터로 표현할 수 있는 근간이 되며, 의사 결정 트리를 구성하는 요소인 속성 집합이 된다.

관심영역의 대표 벡터를 구성하고 있는 다량의 키워드들은 학습 작업을 통해 관심 영역을 표현하는 각 키워드 간의 연관 관계를 갖고 있는 사용자 프로파일을 구성하며, 이것은 사용자별 개인화 서비스를 위한 기초 자료가 된다.

중요 키워드간의 관계를 나타내기 위해 본 논문에서는 속성 선택 시 새로운 속성 추출을 통한 의사 결정 트리의 노드를 구성하는 방식을 적용한다. 클래스별로 초기문서들을 잘 나누는, 즉 최상의 information gain값을 가지는 속성을 선택하여 의사 결정 트리의 노드(루트)를 구성하는 것이 아니라 경미한 차이로 인하여 의사 결정 트리에서 사용자의 중요 관심 속성이 숨겨지는 것을 막기 위해 상위 두 번째의 속성을 포함하는 새로운 속성을 속성집합에 적용하는 것이다. 추가된 속성집합을 대상으로 속성에 대한 평가가 이루어지며 최상의 information gain값을 갖는 속성을 선택하여 의사 결정 트리의 노드를 구성하는 작업을 진행한다. 이를 통해 구성된 의사 결정 트리를 기반으로 사용자 관심 프로파일은 중요 키워드들을 결정 트리 구조로서 표현하게 된다. 의사 결정 트리의 표현으로 인해 단순히 다량으로 나열된 키워드들로 구성되었던 초기 중요 키워드들은 의사 결정 트리를 구성하는 최상위 노드를 중심으로 일정한 규칙 형태로 사용자의 관심영역을 표현하게 된다.

4. 문서 표현

기계 학습을 이용하여 사용자의 관심도를 구체화하는 프로파일을 구축하기 위해서는 학습의 입력데이터가 되는 문서의 표현이 정확해야 한다. 본 논문에서는 관심 영역을 나타내는 관심 키워드를 추출해 내는 특징 추출 기법과 복합 명사 처리를 위한 2-그램 단위의 키워드 추출 기법, 임계값(threshold)을 기반으로 한 문서 재표현 기법 등을 이용하여 문서를 표현한다.

4.1 특징 추출

학습의 입력데이터인 사용자 관심 문서와 비관심 문서들은 학습을 수행할 수 있는 형태로 표현되어야 한다. 본 연구에서는 이를 위해 키워드 정보검색과 텍스트기반 학습에서 주로 사용되는 벡터형식[5], 즉 추출된 키워드와 키워드의 출현 빈도수 등의 값으로 구성된 키워드 벡터로 각 문서를 표현하고 이 벡터를 바탕으로 관심(positive)와 비관심(negative) 클래스를 표현한다.

이에 앞서 본 논문에서는 복합 명사 처리를 위해 2-그램이라 일컫는 단어 집합 개념을 적용하였다. 각각 의미를 갖는 단일 명사뿐만 아니라 이웃하는 단어들의 조합을 통해 복합 명사를 추출해낼 수 있다. 예를 들어 "중국 정부"의 두 음절로 구성된 키워드의 경우, 단일명사인 "중국", "정부" 뿐만 아니라 복합명사인 "중국 정부"를 추출하여 보다 구체적인 의미를 갖는 키워드를 추출하였다.

이렇게 추출된 세부 단어들은 정보 검색과 텍스트 기반 학습 시스템에서 주로 사용되는 문서 처리 기법인 TF(term. frequency), DF(document frequency) 기법을 적용한다[6]. 미리 정의된 명사 사전에 의해서 이미 한번 걸러진 단어들은 TF, DF값과 함께 이를 바탕으로 계산된 P(*word*|*pos*)와 P(*word*|*neg*)라는 확률 값을 속성으로 가진다.

그러나, 실질적으로 학습 작업에 중요한 비중을 갖는 키워드는 실험적으로 밝혀진 바와 같이 관심영역내의 문서를 구성하고 있는 전체 키워드 중 2-5%에 속하는 소수에 불과하다[7]. 따라서 관심 클래스 학습을 수행하기 전에 노이즈가 적은 입력 예제 집합을 만들기 위해서 특징추출 과정을 거쳐 양질의 키워드를 추출하는 과정이 필요하다. 이를 위해 관심 문서를 구성하는 단어들의 출현 확률 값과 비관심 문서를 구성하는 단어들의 출현 확률 값을 이용한 가중치 설정 기법을 사용한다. 확률 값을 이용한 가중치 설정 기법으로는 정보 검색 분야에서 응용되고 있는 OddsRatio[8]기법이 있다. 본 논문에서는 이것을 기초로 한 Exp*P*기법을 사용하여 가중치를 줌으로써 중요 관심 속성을 추출한다. Exp*P*의 수식은 다음과 같다.

$$expP = e^{P(word|pos)-P(word|neg)}$$

- $P(word|pos)$: 키워드 *word*가 *positive*문서에 출현할 확률
- $P(word|neg)$: 키워드 *word*가 *negative*문서에 출현할 확률

[수식1] Exp*P* 공식

Exp*P*는 해당 키워드가 positive문서에 포함될 확률과 negative문서에 포함될 확률의 차이가 클수록 값이 커지는 특성을 가진다. 즉, 관심 문서에는 많이 포함되고 반대로 비관심 문서에는 적게 포함될수록 사용자의 관심 속성일 확률이 높음을 의미한다.

위의 계산 값을 관심 클래스를 구성하는 모든 단어에 적용하여 가중치를 주는 방법을 사용하였다.

이와 같은 가중치 방식은 사용자의 관심 속성의 중요도를 다르게 설정하므로 써 귀납적 학습 시스템이 보다 효과적인 사용자 프로파일을 구축할 수 있도록 한다. 뿐만 아니라 관심 문서에 포함된 속성만은 고려하는 것이 아니라 비관심 문서에 포함되어 있는 속성을 특징 추출 시 고려함으로서 사용자의 관심에는 가깝되 비관심과는 거리가 먼 특징들을 추출해 낼

수 있다.

4.2 문서 재표현

특징 추출을 통해 생성된 관심 속성 집합을 통해 초기 입력 문서들을 재표현한다. 관심 속성 집합에 포함되는 관심 키워드도 비관심 문서에 포함될 수 있는 가능성을 고려하고 이로 인한 노이즈를 최소화하기 위해 1이라는 임계값을 두어 불린(boolean)의 형태로 문서를 재표한다.

다시 말해서, 각 문서에 관심 속성 집합에 포함되는 키워드가 출현하지 않았거나 한번 출현했을 경우에는 0으로 평가, 두 번 이상 출현했을 경우에는 1로 평가, 불린 형태의 키워드 벡터로 문서를 재표현한다.

	단어 재표현 값(Boolean)								관심도
	북한	탈북자	대사관	진입	난민	체포	안내인		
문서1	0	1	1	0	1	0	0	1	positive
문서2	0	1	1	1	1	0	0	1	positive
문서3	1	1	1	0	0	0	1	1	positive
문서4	1	1	0	0	0	0	0	1	negative
문서5	0	1	0	0	0	1	1	0	negative
문서6	0	0	1	1	1	1	0	0	negative
	0	0	0	0	0	0	1	1	negative
문서n	0	1	0	0	0	0	0	1	negative

[그림2] 임계값 기반 문서 재표현

위의 [그림2]와 같은 형태로 문서를 표현한다. 불린 형태의 키워드 벡터로 표현된 문서들은 귀납적 기계 학습의 입력 집합으로 이용된다.

5. 사용자 관심도 학습

특징추출 과정을 통해 사용자 관심 영역을 대표하는 다량의 키워드군이 설정되었다. 이 경우 사용자가 지정한 관심 문서를 정확히 구분 지을 수 있는 중요 키워드는 다량으로 나열된 키워드가 아니라 새로운 문서를 분류하는데 적합한 정확한 키워드여야 할 것이다. 본 논문에서는 귀납적 기계 학습을 적용하고 이의 결과 값인 의사 결정 트리 형태로 중요 키워드를 추출하고자 한다. 또한 사용자의 관심을 좀 더 구체적으로 반영하는 사용자 관심 프로파일을 구축하기 위한 목적으로 새로운 속성 추가를 통한 관심 단어 추출 기법을 이용하여 학습에 이용하고자 한다.

5.1 귀납적 기계 학습

귀납적 기계 학습은 주어진 예제들의 유사성을 이용하여 이들이 암시적으로 표현하고 있는 가설을 추출하는 것이다. 이처럼 귀납적 기계 학습은 예제들 간의 유사성(Similarity)을 기반으로 예제가 암시적으로 나타내는 개념을 추출하는 방법을 사용하므로 흔히 유사성 기반의 학습(Similarity-based learning)이라고 불린다. 본 논문의 기반이 되고 있는 C4.5는 ID3와 더불어 분류모델을 생성하기 위해 만들어진 귀납적 기계 학습 시스템이다. 의사 결정 트리를 이용하여 분류 모델을 생성하는 C4.5의 학습 예제 집합은 속성(attribute)과 속성의 값으로 이루어진 쌍의 집합으로 이루어진다. 이때, C4.5는 한 클래스 값의 예제는 모두 포함하면서 그 외의 다른 클래스 값을 가진 예제는 하나도 포함하지 않는 가장 간단한 형태의 의사 결정 트리를 생성하는 것을 목적으로 한다[9]. 의사 결정 트리의 노드를 구성하기 위해 사용되는 것이 정보이론(Information theory)에 따른 복잡도(Entropy) 개념을 이용한 information gain값이다. 정보이론에서는 한 메시지에 포함된 정보는 그 메시지가 나올 수 있는 확률에 따라 달라지며, 정보량은 밑수가 2인 형태의 logarithm에 마이너스 값의 비트로서 계산한다. 다음은 본 논문에서 사용하는 전체집합 T에 대한 복잡도, 즉 전체 복잡도(entropy)를 구하는 수식이다. 이때, 복잡도 함수는 전체 불명확함(uncertainty) 총량의 평균값을 구한다[10].

$$info(T) = -(P_{pos} \times log_2(P_{pos}) + P_{neg} \times log_2(P_{neg}))$$

- P_{pos} : 문서가 *positive* 클래스에 속할 확률
- P_{neg} : 문서가 *negative* 클래스에 속할 확률

[수식2] 복잡도 공식

전체 문서는 positive문서인지 negative문서인지에 따라서 나뉘며 특징추출 과정을 통하여 구성된 속성 집합을 대상으로 각 속성으로 나누었을 때의 복잡도를 구하고 전체 복잡도 info(T)와 복잡도의 차이를 구한다. 이것이 앞서 언급했던 information gain인데, information gain은 어떤 속성을 선택하여 입력 예제를 분류하였을 경우 얼마나 복잡도를 줄일 수 있는가를 나타내는 도구가 된다. 속성집합에 포함된 모든 속성에 대하여 information gain값을 구하고 가장 큰 값을 가지는 속성을 선택하여 의사 결정 트리의 노드를 구성한다. 그러나, information gain값을 이용하여 의사 결정 트리를 생성하는 것은 전략상 내부적인 결함을 가지고 있다. 즉, 속성을 선택할 때, 분류되어지는 값의 개수가 일정하지 않음으로 야기되는 문제를 해결하지 못한다는 것이다. 따라서 이를 해결하기 위해 Gain Ratio를 제안하고 있다. Gain Ratio는 값의 개수를 정규화(normalization) 할 수 있도록 Split Info값을 이용한다. Split Info 값은 다음과 같다.

$$splitinfo(X) = -\sum_{i=0}^{n} \frac{|T_i|}{|T|} \times log_2[\frac{|T_i|}{|T|}]$$

[수식3] Split Info 공식

Gain Ratio의 값은 기존의 information gain값을 앞서 소개한 Split Info값으로 나눈 것을 말한다. 만약 값의 수가 같다면 데이터의 수가 많은 것의 Split Info값이 작아져 결과적으로 Gain Ratio값을 정규화해준다. Gain Ratio의 수식은 다음과 같다.

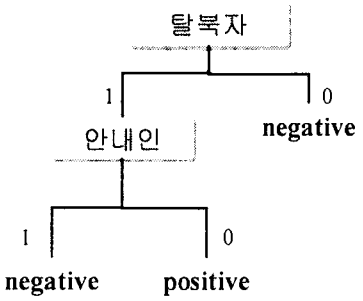
$$gainratio(X) = \frac{gain(X)}{splitinfo(X)}$$

[수식4] Gain Ratio 공식

C4.5는 Gain Ratio값이 가장 큰 속성을 하나 선택하여 루트를 구성하고 선택 속성을 제외한 속성 집합을 대상으로 하여 재귀적으로 속성평가와 속성선택 과정을 진행함으로서 의사 결정 트리를 구성한다. 이와 같은 작업은 모든 속성이 의사 결정 트리에 모두 포함되거나, 단말 노드(leaf node)를 구성하는 속성에 의해서 분류된 모든 문서가 positive나 negative 어느 한쪽 클래스에 포함될 때까지 반복한다.

이와 같이 C4.5는 의사 결정 트리를 구성하고, 이를 이용하여 사용자 프로파일을 구성한다. 이 과정에서 사용자의 관심을 결정짓는 가장 영향력 있는 속성은 의사 결정 트리의 루트 노드에 위치한 키워드로서, 해당 노드에 연결된 다른 중요 속성과 함께 관심 영역에 대한 키워드 규칙을 생성한다.

다음과 같이 가장 클래스별로 문서를 잘 나누는 속성인 "탈북자"는 의사 결정 트리의 노드를 구성하지만, 이미 "탈북자"로 나뉜 하위 집합을 대상으로 속성을 평가하게 되므로 두 번째로 중요한 속성인 "대사관"이라는 속성은 의사 결정 트리에 고려되지 않아 결과적으로는 사용자 관심 프로파일에서 숨겨졌음을 볼 수 있다. 또한 루트를 구성하고 있는 속성에 의해 전체적으로 편중된 의사 결정 트리의 형태를 볼 수 있을 것이다.



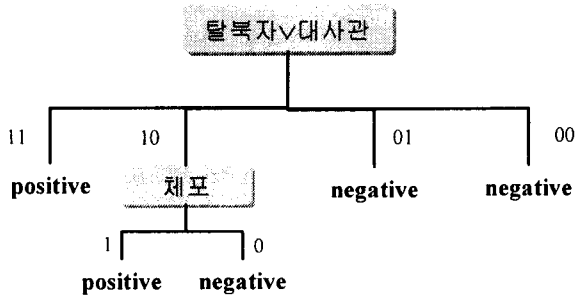
[그림3] C4.5의 의사 결정 트리

5.2 수정된 속성 선택 알고리즘

귀납적 기계 학습 과정을 통해 관심 영역을 구체적으로 표현하는 키워드 구조를 결정할 수 있다. 이러한 키워드 구조는 첫 번째 선택된 키워드의 중요도가 매우 높기 때문에 루트에 따라서 하위 노드를 구성하는 키워드들이 좌우될 가능성이 매우 높다. 그러므로 최상위 노드의 키워드에 의해서 사용자의 관심을 표현하는 중요한 키워드임에도 불구하고 약간의 차이로 인하여 중요한 키워드가 의사 결정 트리에서 배제될 수도 있는 것이다.

따라서 본 논문에서는 중요 관심 속성이 숨겨지는 것을 최소화하고 보다 사용자의 관심을 많이 포함할 수 있는 사용자 관심 프로파일을 구축하기 위해 다음과 같은 속성 선택 방식을 제안한다. 특징추출을 통해 구성된 속성집합을 그대로 귀납적 기계 학습의 입력 데이터로 사용하는 것이 아니라 disjunctive operator를 이용하여 최고의 속성인 "탈북자"와 다음으로 큰 information gain값을 갖는 "대사관"을 포함하는 새로운 속성을 속성집합에 추가하여 학습에 이용한다. 새로운 속성인 ("탈북자"∨"대사관")은 두개의 속성을 모두 포함하는 경우, 혹은 둘 중 하나의 속성만을 포함하는 경우, 마지막으로 두개의 속성 모두를 포함하지 않는 경우, 이렇게 네 개의 값(value)을 가지며 이것으로 문서를 분류하게 된다. 새로운 속성의 information gain값을 구하고 원래 최고속성의 그것과 비교한다. 이때, 두 가지의 경우가 발생할 수 있다.

첫 번째, 최고 속성인 ("탈북자")의 information gain값이 클 경우가 그것이다. 이때는 기존의 C4.5 알고리즘을 수행하는데, ("탈북자")를 선택하여 의사 결정 트리의 노드를 구성하고 다음 단계에서 선택 속성을 제외한 속성 집합을 대상으로 속성 선택을 한다.



[그림4] 추가 속성 선택을 통한 의사 결정 트리

두 번째 경우로, disjunctive operator를 이용하여 생성한 새로운 속성인 ("탈북자"∨"대사관")의 information gain값이 더 클 경우를 들 수 있다. 이때는 연산을 통해 생성된 속성을 속성집합에 추가하고, 이것을 구성하는 ("탈북자")과 ("대사관") 속성을 속성집합으로부터 제거한다. ("탈북자"∨"대사관")을 선택하여 의사 결정 트리의 노드를 구성하고 다음 단계에서 선택 속성을 제외한 속성 집합을 대상으로 속성 선택을 한다.

6. 실험

본 논문에서 구현한 사용자 관심 프로파일 구축 시스템의 각 모듈은 java언어로 구현되었다. 또한 각 입력 문서집합은 실제 문서 검색 시에 사용자의 실제 관심문서에 비해 비관심 문서가 더 많은 것을 고려하여 positive문서 21개와 negative문서 50개를 한 집합으로 하여 구성하여 이를 대상으로 관심도 학습을 진행하였다. 다양한 분야 사용자 관심도를 실험에 반영하기 위해 여러 분야에 걸쳐 위와 같은 구성의 문서

집합 50개를 대상으로 실험을 진행하였다.

6.1 사용자 프로파일 구축 결과

다음은 기존의 C4.5알고리즘을 이용하여 구성한 사용자 관심 프로파일과 개선된 관심 단어 추출기법을 이용한 TDIDT기반 알고리즘을 이용하여 구성한 사용자 관심 프로파일을 비교한 것이다.

```
Rule 1:
        탈북자 = 0
         -> class Negative
Rule 2:
        탈북자 = 1
        안내인 = 0
        -> class Positive
```

[그림5] 사용자 관심 프로파일(C4.5)

```
Rule 1:
        탈북자 ∨ 대사관 = 11
         -> class Positive
Rule 2:
        탈북자 ∨ 대사관 = 10
        체포 = 1
         -> class Positive

Rule 3:
        탈북자 ∨ 대사관 = 10
        체포 = 1
         -> class Positive
Rule 4:
        탈북자 ∨ 대사관 = 01
         -> class Negative
Rule 5:
        탈북자 ∨ 대사관 = 00
         -> class Negative
```

[그림6] 사용자 관심 프로파일(개선된 C4.5)

여기서 disjunctive operator를 이용하여 생성한 속성은 11, 10, 01, 00, 네 개의 속성 값을 가지며, 각각은 이 속성을 구성하고 있는 두개의 키워드를 모두 포함하고 있는 문서와 두개의 키워드 중 하나만을 포함하고 있는 문서, 두개의 키워드 중 어느 것도 포함하지 않는 문서로 전체 문서를 분류하는 것을 의미한다. 네 가지의 경우에 대하여 각각의 규칙을 추출하고 이를 바탕으로 사용자 관심 프로파일을 구축한다.

앞서 본 의사 결정 트리의 경우와 마찬가지로 기존의 C4.5알고리즘을 이용하여 사용자 관심 프로파일을 구축한 결과 루트를 구성하고 있는 속성에 집중되어 사용자 관심 프로파일에 사용자의 관심을 제대로 반영하지 못했음을 알 수 있다. 이에 비하여 최고의 속성 하나만을 보는 것이 아니라 의사 결정 트리의 노드를 구성하기에 앞서 두 번째 속성까지 포함하는 새로운 속성을 의사 결정 트리를 구성하는 속성집합에 포함시킴으로서 다음과 같이 기존의 C4.5에서 숨겨질 수 있었던 중요 속성을 사용자 관심 프로파일에 포함시킬 수 있었다.

[그림5],[그림6]은 기존의 C4.5알고리즘과 본 논문에서 제안한 속성 선택 방식을 이용한 개선된 C4.5알고리즘의 결과물로 구성된 의사 결정 트리를 규칙의 형태로 보이는 것이다.

7. 결론 및 향후연구

본 논문은 보다 정확한 사용자의 관심 프로파일을 구축하기 위해 웹 문서 내에서 사용자의 관심을 나타내는 중요한 키워드를 추출하는 특징추출 기법과 사용자 요구 정보를 분석하는 사용자 프로파일 구축 기술, 의사 결정 트리 구성 시 속성선택의 과정에서 hill-climbing방식으로 최고의 information gain값을 가지는 속성 하나만을 선택함으로 인해 숨겨졌던 사용자 관심 속성을 의사 결정 트리에 반영하기 위해 새로운 속성 선택 알고리즘을 제안하였다.

먼저, 사용자가 자신의 관심 영역과 비관심 영역으로 생성한 임의의 문서군을 대상으로 확률을 기반으로 한 특성 추출기법을 통해 관심 영역을 대표하는 다량의 키워드 집합을 추출한다. 이들은 수정된 속성 선택 알고리즘에 의해 의사 결정 트리의 구조를 구성하며 이렇게 구축된 의사 결정 트리 구조는 키워드간의 구체적인 규칙을 생성하여 사용자 관심 프로파일로 작성된다.

disjunctive operator를 이용하여 생성한 새로운 속성을 의사 결정 트리 형성 시 이용하면 사용자의 중요한 관심을 포함하는 의미 있는(semantic) 사용자 관심 프로파일 구축이 가능해지고, 사용자 관심 프로파일을 기반으로 사용자가 관심 있는 문서를 제공할 수 있는 개인화 서비스를 제공할 수 있을 것이다.

본 논문에서는 사용자 프로파일을 구축하는데 감독학습방식을 적용하였다. 이는 학습 결과가 사용자의 관심 영역을 보다 정확하게 표현할 수 있다는 장점이 있는 반면 사용자에게 자신의 관심 영역을 항상 표시하게 하는 부담을 준다는 단점을 가지고 있다. 향후에는 이러한 불편 없이도 사용자의 관심도를 추출할 수 있도록 사용자 행위 모니터링을 통한 비감독학습(unsupervised learning) 방식을 적용할 수 있도록 확장이 이루어져야 할 것이다.

참고문헌

[1] L.Dent, J.Boticario, J.McDermott, T.Mitchell, D.Zabowski, "A Personal Learning Apprentice", 1994.

[2] Michael Pazzani, Daniel Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites", 1997.

[3] Bruce Krulwich, Chad Burkey, "The InfoFinder Agent: Learning User Interest through Heuristic Phrase Extraction,", 1995.

[4] J R. Quinlan, "C4.5 Programs for Machine Learning", 1993.

[5] Dunja Mladenic, "Text-Learning and Related Intelligent Agents: A Survey", 1999.

[6] Thorsten Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", 1996.

[7] Dunja Mladenic, "Feature subset selection in text-learning", 1998.

[8] van Rijsbergen, C,J,. Harper, D.J., Porter, M.F., "The Selection of Good Search Terms.", 1981.

[9] J R. Quinlan, "Induction of Decision Tree", 1986.

[10] Rechard W.Hamming, "Coding and Information Theory", 1986.