단어의 연관성을 이용한 문서의 자동분류

신 진 섭† · 이 창 훈††

요 약

본 논문에서는 단어들 사이의 연관성을 이용하여 문서들을 사용자의 관심분야 만큼 자동으로 분류하는 다음과 같은 방법을 제안한다. 첫째, TF*IDF 알고리즘을 이용하여 각 문서를 대표할 수 있는 단어들을 찾아내고, 본 논문에서 제안한 연관성 계산을 위한 확률 모델을 이용하여 각 문서를 대표하는 각각의 단어들이 문서 전체집합에서 서로 어느 정도 연관성을 갖고 있는가를 계산한다. 둘째, 연관성이 가장 높은 두 단어를 중심으로 그 단어들에 밀접하게 연결되어 있는 단어들을 하나의 집합으로 구성하고, 그 집합을 이용하여 하나의 클래스와 프로파일을 생성한다. 연관성이 다음으로 높은 두 단어를 중심으로 위와 같은 과정을 임계 값 보다 낮은 값이 나올 때까지 계속적으로 반복함으로써, 사용자가 관심 있는 분야만큼의 프로파일을 생성한다. 또한, 본 논문에서는 생성된 각각의 프로파일이 각 문서들에 어느 정도의 영향력을 갖고 있는지를 평가하여 문서들을 분류하고, 기존의 자동 문서 분류 방법과의 비교를 통하여 본 논문에서 제시한 방법의 타당성을 입증한다.

Automatic Classification of Documents Using Word Correlation

Jin-Seob Shin† · Chang-Hoon Lee††

ABSTRACT

In this paper, we propose a new method for automatic classification of web documents using the degree of correlation between words. First, we select keywords from term frequency and inverse document frequency (TF*IDF) and compute the degree of relevance between the keywords in the whole documents, using the probability model proposed in this paper. Second, centering around two words having the most intimate relations, we extract the set of word that was closely connected with them and create a profile that characterizes each class. Finally, if we repeat the above process until lower than threshold value, we will make several profiles which are in keeping with users concern. And, we classified each document with the profiles and compared these with those of other automatic classification methods.

1. 서 론

인터넷을 통한 정보 교류(information exchange)가 보편화되면서 정보의 질적, 양적인 급격한 증가를 가져왔으나, 상대적으로 사용자가 원하는 가장 적절한 정보의 검색은 점점 더 어려워 지고 있다. 이런 문제를 해결하기 위해 사용자의 검색과 분류를 도와주는 에이전트에 관한 연구가 널리 이루어 지고 있다. 이러한 에이전트는 사용자 관심의 학습을 통해 사용자의 관심에 맞는 문서만을 선별하여 사용자에게 제공하는 기능을 수행한다. 이때 에이전트가 학습한 사용자의 관심사항에 대한 정보를 사용자 프로파일이라고 한다. 그러나 대부분의 사용자는 한가지 이상의 분야에 대해

※ 본 연구는 정보통신연구진흥원 대학기초연구지원사업의 연구비 지원에 의해 수행되었음.
† 정 회 원 : 대전보건대학 사무자동화과 교수
†† 종신회원 : 건국대학교 컴퓨터공학과 교수
논문접수 : 1999년 3월 16일, 심사완료 : 1999년 8월 16일

서 흥미를 지니고 있으며, 여러 분야에 대한 사용자의 관심을 하나의 프로파일에 관리하는 것은 매우 비효율적이다. [1, 4, 6]

기존의 자동 문서 분류 모델에서는 사용자의 여러 가지 관심에도 불구하고 하나의 분류 틀에 적용시키기 때문에, 사용자의 관심을 적절하게 고려하지 못하였으며, 또한 이로 인하여 동적으로 변화하는 사용자의 관심에 대처하기가 어려웠다.[3, 7] 한편, 각각의 문서들을 단지 하나의 분류 틀에만 고정시키므로 인하여 문서의 다중 특성을 고려하지도 못했다. 이를 해결하기 위한 방안으로 본 논문에서는 에이전트가 자동으로 문서를 분류할 때 사용자의 관심 있는 분야만큼의 독립된 프로파일을 생성할 수 있는 모델을 제안한다. 또한 각각의 프로파일의 특성에 따라 문서들을 분류하며 문서의 특성을 고려하여 여러 프로파일에 적용되는 문서 분류도 가능한 새로운 방법을 제시한다.

1.1 연구 방향

인터넷의 수많은 정보를 사용자에게 효율적으로 제공하는 역할을 담당하는 에이전트를 웹 에이전트(web agent)라고 한다. 웹 에이전트의 기능은 크게 두 가지로 나눌 수 있다. 하나는 사용자가 원하는 정보를 찾아주는 정보 검색(information retrieval)에 관한 기능이고, 또 다른 하나는 찾아온 정보들 가운데서 사용자에게 가장 적절한 정보만을 선택 제공하여 정보의 질을 높여주는 정보 여과(information filtering)에 관한 기능이다. 정보 검색이나 정보 여과의 공통적인 목적은 사용자의 부담을 최소화하면서 사용자에게 필요 없는 정보는 제거하고 필요한 정보만을 찾아서 제공하는데 있다. 그러나 현재의 웹 에이전트로서는 사용자의 관심에 대한 정보를 좀더 체계적으로 분류 및 관리하는 것이 어렵다. [1, 6, 9]

본 논문에서는 사용자들의 다양한 요구로 인하여 다양한 문서들이 수집되었을 때 이들 문서들 중에서 사용자가 관심 있는 분야의 수만큼 자동적으로 사용자의 프로파일을 생성하고 이에 따라 문서들을 분류하는 문서 자동 분류를 수행한다. 또한 위와 같은 과정을 통하여 필요 없는 정보를 제거하고 사용자가 요구하는 정확한 정보를 검색하여 올 수 있도록 한다.

위와 같은 결과를 도출하기 위하여, 본 논문에서는 첫째, 불용어 제거 알고리즘, 어근 추출 알고리즘과 TF*IDF를 이용하여 주요단어를 추출한다. 또한, 문서의 길이에 따른 단어의 영향력 불균형을 해소하기위해서 벡터 길이 정규화 알고리즘을 이용한다. 둘째, 주요 단어들간의 문서들에서 차지하는 상호 연관성을 조사하기 위하여 확률을 이용한다. 셋째, 상호 연관성이 가장 높은 단어들을 중심으로 연관성을 재계산한 테이블을 만들고 이를 통하여 사용자 프로파일을 만들었다. 이 과정을 반복하면 사용자가 관심 있는 여러 분야의 프로파일이 생성된다. 마지막으로, 이 프로파일을 기준으로 각각의 문서들을 분류하여 사용자에게 정렬된 순서대로 보여줄 수 있도록 한다.

2. 문서 자동 분류

문서의 자동 분류는 문서 안에서 중심이 되는 단어들을 찾아내는 색인 추출(indexing) 과정과 추출된 색인을 이용하여 문서간의 연관성에 따라 문서들을 분류하는 클러스터링(clustering) 과정으로 구성된다. [3, 4, 7, 11]

2.1 색인 추출

문서에서 색인을 추출하기 위한 방법으로는 일단 색인 대상 단어의 수를 줄이기 위해 사용되는 어근 추출 알고리즘, 불용어 제거 알고리즘, 동의어(thesaurus) 사전을 이용하는 방법과 대상 단어 집합에서 주요한 키워드만을 추출하는 TF*IDF 알고리즘이 있다.[2, 3, 5, 8]

2.1.1 어근 추출 알고리즘

접두사 또는 어미 등 어근에 붙어있는 부분을 제거하는 알고리즘으로서 단어를 어근으로 분리하므로 같은 단어이지만 접미사나 어미의 변화에 의해 다른 단어로 인식되는 것을 막을 수 있다.

2.1.2 불용어 제거 알고리즘

문서에서 많이 나오는 일반적인 단어는 키워드로 사용하는데 부적절 하므로 주요 키워드 후보에서 제거하는 알고리즘이다. 일반적으로 영어의 접속사(and, or, but), 동사(is, are)와 관사(a, the) 등이 있으며, 문서마다 그 문서의 특징에 따라서 불용어들은 약간의 차이점을 갖는다.

2.1.3 동의어 사전

동의어 사전이란 같은 의미를 갖고 있는 단어이지만 단어의 철자가 다른 경우 또는 상위 개념의 단어와 하

위 개념의 단어 사이에서 발생할 수 있는 문제점을 해결하기 위해서 제안된 방법이다. 예를 들어 "bike", "bicycle"은 같은 의미를 갖고 있으나 서로 철자가 틀리므로 다른 단어로 인식할 수 있다. 또한 탈것 이라는 단어의 개념과 "자전거", "자동차", "기차" 등의 단어들 사이에는 포함 관계가 존재하나 이를 인식하기가 쉽지 않다. 이러한 문제를 해결할 수 있다는 장점에도 불구하고 동의어 작성이라는 방대한 작업과 사장되는 단어와 새로이 만들어지는 단어들, 그리고 의미가 변경되는 단어들을 관리하는 비용이 막대한 단점으로 인하여 특정하게 제한된 분야인 경우를 제외하고는 사용하기가 어렵다.

2.1.4 TF*IDF 알고리즘

많은 문서들 중에서 그 문서들을 대표할 수 있는 특징을 추출하기 위해서 단어의 빈도수(term frequency)를 많이 이용한다. 그러나 단어의 빈도수가 높은 것이 그 문서를 정확히 대표하는 단어가 된다고 확신할 수는 없다. 실제로 많은 문서에서 그 문서를 대표하는 단어는 빈도가 그리 높게 발생하지 않고 있다. 이러한 단어 빈도수의 문제점을 해결하기 위하여 여러 문서에서 많은 빈도를 나타내는 용어는 일반적인 용어로서 문서의 대표성과는 관련성이 떨어진다고 볼 수 있다. 그러므로 TF*IDF는 역 문서 빈도수(inverse document frequency)를 단어의 빈도수와 같이 적용 함으로서 그 문서를 대표하는 단어들을 효율적으로 찾을 수 있는 알고리즘이다.

문서의 빈도 df_i 는 N개의 문서들 중에서 단어 t_j가 존재한 문서의 개수를 의미하며, 단어의 빈도 tf_{ij} 는 문서 d_i에서 단어 t_j가 나타난 수를 의미한다. 이때 $\log(N/df_i)$는 역 문서 빈도수를 의미하며, 역 문서 빈도수와 단어 빈도수를 곱한 값을 문서 d_i에서 단어 t_j의 중요도 또는 영향력(weight) 이라고 말하며 이를 w_{ij} 부른다.

$$w_{ij} = tf_{ij}\log(N/df_i) \tag{1}$$

2.1.5 백터 길이 정규화 (vector length normalization) 알고리즘

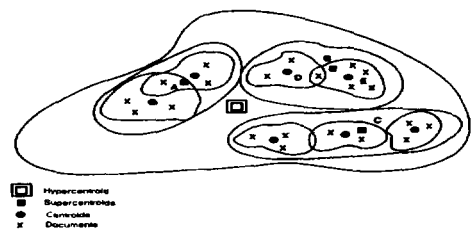
TF*IDF 에 의해 생성되는 단어와 문서에 대한 영향력의 모음은 하나의 백터의 형태로 간주된다. 이 경우 큰 문서의 프로파일은 전체 사용자 프로파일 생성시 높은 영향력을 지닐 수 있게 된다. 이러한 문서의 길이에 의한 영향력의 불균형을 해결하기 위해 각 벡터의 길이를 동일하게 하는 벡터 길이 정규화 과정이 필요하다. 이는 단어의 가중치를 가중치의 총합으로 나누어 줌으로서 수행된다. 아래 식 (2)에서 n은 전체 단어의 개수이다.

$$w' = \frac{w}{\sum_{i=1}^{n} wi} \tag{2}$$

2.2 문서 클러스터링

문서를 분류하기 위해서는 문서들의 특징을 전혀 모르는 상황에서 각각의 문서의 특징을 추출할 수 있어야 하며, 그들 사이의 공통된 패턴을 발견하고, 목적과 일치되는 판별기준을 수식화 할 수 있어야 한다. 이 때, 공통된 패턴에 따라 모여진 집합을 클러스터(cluster)라고 하고 주어진 패턴들을 수식에 의해 무리 지워 나가는 과정을 클러스터링(clustering)이라고 한다.[3, 9, 11]

아래의 (그림 1)은 클러스터된 전형적인 파일 구조를 보여주고 있다. 파일은 A, B, C의 3개의 큰 클러스터로 나누어 지고 각 클러스터는 더 작은 클러스터로 나누어지며 가장 작은 클러스터는 문서들을 포함하고 있다. 클러스터된 파일에서 모든 클러스터는 클러스터 중심(centroid)이라 불리는 특별한 단어 백터로 표현될 수 있다. 그리고, 중심은 한 클러스터 안의 모든 문서에 동일하게 적용된다. 클러스터를 생성하는 알고리즘에는 크게 계층적 알고리즘과 휴리스틱 알고리즘이 있다.



(그림 1) 클러스터된 파일 구조의 전형

첫째, 계층적 클러스터 생성 알고리즘으로 클러스터링을 하기 위해서는 두 개의 문서 사이의 모든 유사도의 쌍이 필요하다. 모든 유사도 중에서 가장 큰 유사도를 가진 두 개의 문서들을 하나의 클러스터로 묶고 다음으로 큰 유사도를 지닌 문서들을 클러스터로 묶어

나가서 결국은 하나의 커다란 클러스터로 구성되게 된다. 계층적 클러스터링은 큰 클러스터에 접근해서 주제 범위가 넓은 탐색이 가능하고 또한, 작은 클러스터에 접근하여 특정한 주제에 대한 탐색이 가능하다.

둘째, 휴리스틱 생성 알고리즘은 처음에는 하나의 문서를 클러스터로 놓고, 다른 문서를 선택해 이것과 비교하여 유사하면 그 클러스터에 위치시키고 그렇지 않으면 새로운 클러스터로 놓는다. 계층적 생성 알고리즘과는 달리 유사도의 쌍이 필요 없고 비교적 적은 비용으로 빨리 클러스터링을 할 수 있다. 기존의 클러스터와 새로운 문서와의 유사도를 결정하기 위해서 중심으로 각 클러스터를 표현하는 것이 편리하다. 새로운 문서가 클러스터에 덧붙여질 때, 대응되는 중심은 적절하게 갱신되어야 한다.

3. 단어들 사이의 연관성 계산

인터넷에서 내려 받은 많은 문서들 중에서 유사한 문서들을 그룹화 시키기 위해서는 유사 정도를 판단할 수 있는 기준이 필요하다. 이러한 판단 기준으로 HTML문서의 형식과 내용이 있을 수 있다. 자동분류의 성능을 높이기 위해선 사람처럼 양자 모두를 고려해야 하겠지만, 본 논문에서는 우선 후자인 문서의 내용에 초점을 맞추어서 연구한다.

문서의 내용을 분류 기준으로 적용하기 위해서는 자연어처리에서 사용되는 구문 및 의미 분석 과정을 거쳐야만 된다. 그러나 이는 현재까지도 연구 진행 중인 분야이며 아직까지 만족할만한 성능에 이르지 못하고 있다. 또한 실제 구현이 되어도 방대한 처리량에 따른 속도 문제가 남아있다. 그러므로 본 논문에서는 문서에 사용된 키워드를 중심으로 이를 분류할 수 있는 방법을 제시하고자 한다. [1, 9]

키워드를 이용하여 문서를 분류하기 위해서는, 우선 문서에서 주요한 역할을 차지하는 키워드를 찾아내야 한다. 처음에는 문서 전체의 모든 단어를 중심으로 실험을 하였으나 문서를 분류하는 시간이 너무 오래 걸리는 문제점으로 인하여 기존의 불용어 제거 알고리즘, 어근 추출 알고리즘 그리고 TF*IDF 알고리즘과 벡터 길이 정규화 알고리즘을 이용하여 각 문서에서 주요한 역할을 하는 단어들을 우선 색출한 후에 각 단어들간의 전체 문서 집합에서 연관성 정도를 조사한다.

본 논문에서는 우선적으로 2개의 그룹으로 나뉘어진 문서를 분류할 수 있는 모델을 제시하고 또한 이를 통해 2개 이상의 그룹 또는 하나의 관련성 있는 문서 집합과 그 외 관련성 없는 잡다한 문서로 이루어진 문서 집합에서 관련성 있는 그룹의 추출 등에 적용할 수 있는 확장 가능성에 대해 논하고자 한다.

두개의 성격을 가진 문서의 그룹이 있다. 하나는 인공지능에서 "신경 망"을 다룬 문서들이고 다른 하나는 유전자 "알고리즘"을 다룬 문서들이다. 이 경우 우리는 쉽게 신경 망 및 유전자 알고리즘에 관련된 몇몇 단어를 이용해 이 문서를 분류할 수 있다는 것을 알 것이다. 그러므로 이러한 단어를 찾음으로써 문서의 분류가 가능하다. 사람은 문서들의 내용을 이해하고 어떤 기준으로 나뉘어졌는지를 이해하고 있기 때문에 쉽게 이러한 단어를 유추할 수 있다. 그러나 컴퓨터는 이러한 내용에 대한 이해가 어렵다. 그러므로 본 연구는 각 문서의 그룹을 특징지을 수 있는 키워드를 찾는데 초점을 맞추었다.

문서의 그룹이 정의 되어 있다면 쉽게 그 그룹을 특징지을 수 있을 정도로 많이 나오는 키워드를 찾을 수 있다. 그러나 본 연구에서 제안하는 자동 분류기법에 있어서는 역으로 문서 그룹의 정의 자체가 목적이고 키워드 추출은 수단이 된다. 이러한 키워드를 찾기 위해서 본 연구에선 키워드의 밀집성을 이용한 방법을 제시한다. 키워드의 밀집성이란 하나의 그룹을 대표하는 단어는 하나가 아니며 이들은 그들이 대표하는 문서에 사용될 확률이 다른 문서에 사용될 확률보다 높다라고 정의한다. 그러므로 같은 주제에 많이 사용되는 두 키워드는 같은 문서에 사용될 가능성이 서로 연관이 없는 두 단어에 비해 높다. 이는 확률의 문제로서 우리는 두 단어가 같은 주제를 대표할 가능성을 두 단어사이의 연관성으로 규정하였다. 본 연구에선 전체 문서에 출현한 단어의 집합 중에 n번째 단어와 m번째 단어의 연관성을 다음과 같은 방법으로 계산한다.

$$Rnm = -ln(\text{우연히 두 단어가 문서에서 중복되어 나타날 확률})$$

이는 우연히 두 단어가 중복해서 나올 확률이 낮을수록 두 단어는 서로 연관성이 높은 확률을 갖는다는 것을 의미한다. 만일 어떤 단어가 10개의 문서 중 6개의 문서에 사용되었고 다른 단어는 10개의 문서 중에서 3개의 문서에 사용되었다고 가정하자. 이 때 3개의 문서에서 위의 두 단어가 동시에 사용되었다고 가정하

자. 아무 연관도 없는 단어들 사이에 이런 일이 발생할 확률은 다음과 같다.

(6/10 * 3/10)**3 * (1-6/10*3/10)**7 * (10! / (7!*3!))

이를 일반화 시키면 다음과 같다.

$$P_{wc}^{Dc} \times (1-P_{wc})^{D-Dc} \times dC_{Dc} \quad (3)$$

P_{wc}는 하나의 문서가 두 단어를 동시에 포함할 확률로서 식(4)와 같다. 이때 Dc는 두 단어가 동시에 사용된 문서의 수이며 D는 전체 문서의 숫자이다.

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

P_{wn}은 전체 문서 중 n번째 단어가 포함된 문서의 비율이며 P_{wm}는 m번째 단어가 포함된 비율이다. 이를 기반으로 n번째 단어와 m번째 단어간의 연관성을 추정하면 식(5)와 같다.

$$R_{nm} = -1n(P_{wc}^{Dc} \times (1-P_{wc})^{D-Dc} \times {}_{D}C_{Dc} \quad (5)$$

4. 클래스 프로파일

두 단어의 문서 전체에서의 연관성은 식(5)에 의하여 계산되며, <표 1>은 이러한 각 단어들간의 연관 관계를 예를 들어 표로 구성한 것이다.

〈표 1〉 단어들간의 연관 관계 표 예제

	Genetic	crossover	function	···	network
Genetic					
crossover	3.1				
function	1.7	4.1			
:	:	:	:		
network	1.4	3.7	1.8	2.1	

4.1 프로파일 생성

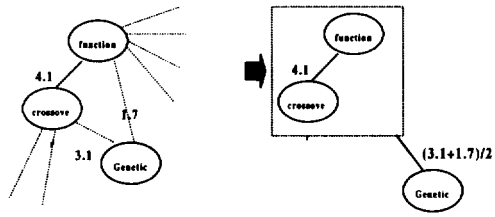
<표 1>에서 보여지는 단어들간의 연관관계는 프로파일을 구성하는 중요한 값으로서 이 표의 값 중에서 최대값을 기준으로 새로운 집합을 구성 하여 하나의 프로파일 후보를 만들어 나간다. 우선 모든 단어들간의 연관관계 값을 계산하며, 그 단어들은 모두 프로파일 후보가 되어 다음의 과정을 수행한다.

(1) 현재의 프로파일 후보간에 연관관계를 검색하여, 가장 높은 연관관계를 지닌 두 후보를 하나의 집합으로 구성한다. 새로 생성된 집합은 다시 하나의 프로파일 후보로서 간주된다.

(2) 병합의 과정에 따라 단어는 단어의 집합이 되며 이러한 집합은 프로파일 후보로서 간주된다.

(3) 병합과정에 의해 프로파일 후보의 소속 멤버가 추가됨에 따라 주위와의 연결값에 대한 갱신이 필요하다. 각각 n개와 m개의 소속 단어를 가진 프로파일 후보 A, B간의 연관도는 다음 식에 의하여 계산 및 갱신된다.

$$R_{AB} = \frac{\sum_{i=1}^{n}\sum_{i=1}^{m} R_{nm}}{n \times m} \quad (6)$$



(그림 2) 프로파일 후보간의 병합과정

(4) 현재 남아있는 연관관계의 최대값이 일정한 값 이하로 떨어질 때까지 [1]과 [2]의 과정을 반복한다.

<표 1>의 연관 관계 테이블에서 위 과정을 수행한 결과는 앞의 (그림 2)와 같다

위 과정에서 만들어진 후보 프로파일을 이용하여 다음과 같은 과정으로 최종적인 프로파일을 생성한다.

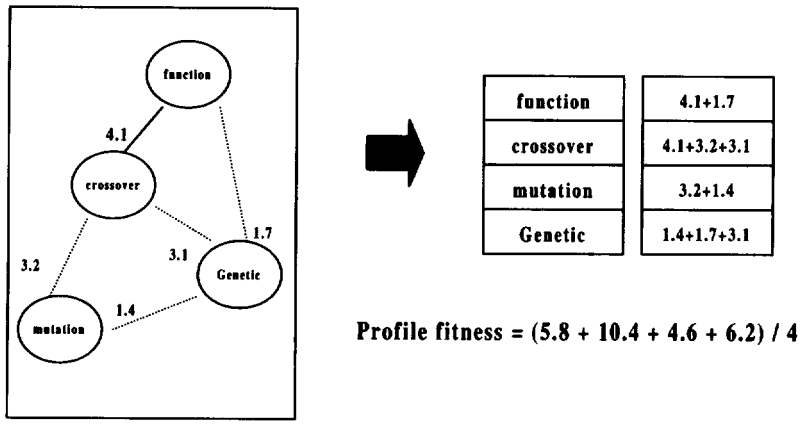
(1) 선택된 프로파일 후보내의 단어간의 원래의 연관관계를 <표1>을 이용하여 모두 연결한다.

(2) 각 단어별로 프로파일 후보내의 다른 단어로의 연관 관계를 나타내는 값을 합한 후에 이를 프로파일에서 그 단어의 가중치로 계산한다.

(3) 식(7)을 이용하여 적합한 프로파일인가를 평가한다.

프로파일의 적합도=단어의 가중치의 총합 / 단어의 개수 (7)

위 과정에 의하여 최종적인 프로파일을 생성하는 과정은 (그림 3)과 같다.



(그림 3) 프로파일 생성 과정

4.2 문서 분류

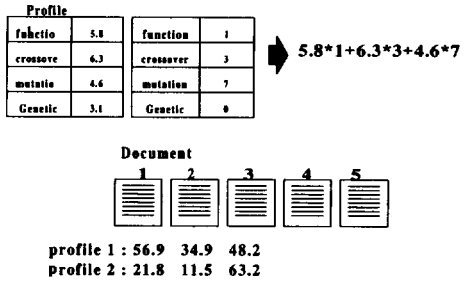
위에서 제안한 프로파일 생성과정을 이용하면 여러 개의 프로파일을 생성할 있다. 이러한 각각의 프로파일은 서로간의 밀집성을 지닌 단어들의 묶음으로서 구성된다. 생성된 프로파일을 기준으로 어느 문서가 어느 프로파일에 포함되는가를 찾음으로써 문서를 분류 할 수가 있다. 하나의 문서가 소속될 프로파일을 결정하기 위해서는 어떤 프로파일이 그 문서에 나오는 단어와 높게 일치하는 가를 계산하여야 한다. 이를 문서에 대한 프로파일의 영향력으로 정의하며 각각의 문서가 포함되는 프로파일을 찾아가는 과정은 다음과 같다.

(1) 각 프로파일의 단어에 대한 가중치와 문서에서 그 단어의 빈도수를 곱한 값을 합하여 (그림4)와 같이 문서에 대한 프로파일의 영향력을 계산한다.

$$\text{프로파일 영향력} = \sum_{i=1}^{n} \text{각 단어의 가중치 X 문서에서 각 단어의 빈도} \quad (8)$$

(2) 각각의 문서에 대한 프로파일의 영향력을 나타낸 값 중에서 최대값을 지니는 프로파일을 그 문서의 대표 프로파일로 선정하여 이 문서는 그 프로파일에 의해 생성된 그룹에 소속된다.

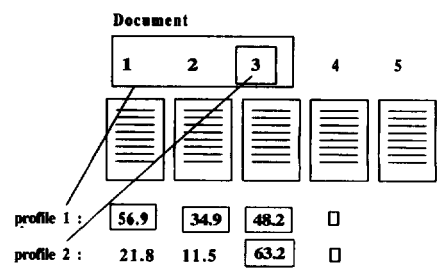
(3) 수 많은 실험 결과 프로파일에 대한 영향력이 15 이하일 경우 어떤 그룹에도 소속되기 어려운 문서로 판명되었으며 이로 인하여 영향력이 15 이하인 문서는 분류 불가능 문서로 취급하였다.



(그림 4) 문서에 대한 프로파일의 연관성

(4) 각각의 프로파일에 최대값을 갖는 문서들의 영향력 값의 평균을 구하고 최대값이 아니면서 평균값 이상인 값들을 가진 문서들도 대표 값으로 선정한다.

(5) 각각의 프로파일에 영향을 주는 대표 값을 갖고 있는 문서들을 (그림 5)와 같이 분류하고 값의 순서에 의해 문서들을 정렬 함으로서 사용자의 관심이 높은 문서부터 보여줄 수 있도록 한다.



(그림 5) 문서 분류

5. 실험 및 평가

본 연구에서 사용한 확률을 이용한 자동 문서 분류기법은 기존의 분류 방식과는 근본적인 차이가 있다. 기존의 방법은 문서와 문서간의 유사도를 키워드 벡터를 사용하여 추출하는 방법이었다. 그러므로 분류된 문서를 기준으로 프로파일을 생성한다.

문서간의 유사도 계산-문서의 분류-분류된 문서의 프로파일 생성

그러나 본 연구에서 사용한 방법은 키워드 간의 유사도를 이용해 키워드 프로파일을 먼저 생성하고 이를 기준으로 문서를 분류하는 방법이다.

키워드간의 유사도 계산-키워드 프로파일 생성-문서의 분류

문서의 분류 알고리즘의 중요한 기준은 크게 2가지로 볼 수 있다.

- 알고리즘의 정확도 : 오류 즉, 잘못된 분류가 적을수록 효율적인 알고리즘이다. 이는 두 부류의 문서뿐만 아니라 여러 부류의 문서에 대해서도 적용되어야 한다.
- 분류 항목의 정확성 : 각 주제별 적절한 분류로 나뉘어 져야 한다.

이를 위해서 본 실험에서는 전문가(사람)가 분류한 경우와 Salton이 제시한 자동 문서 분류기법[3, 7, 10]에 의한 분류, 그리고 본 연구에서 제시한 방법에 의한 분류를

1) 분류 항목이 얼마나 잘 선정되었는가? ,
2) 분류 항목간의 연관도 및 계층에 관한 관계가 성립되는가?
3) 각 분류 항목에 포함되는 문서는 적절한가?

라는 항목을 이용하여 비교, 평가하였다.

실험의 대상은 다음과 같다.

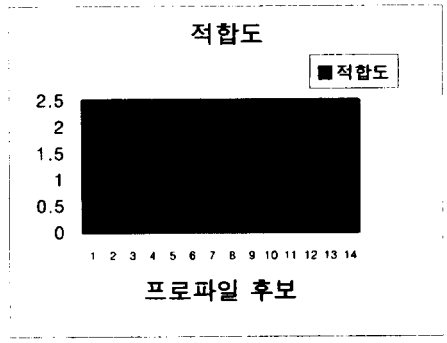
(1) 인공지능 관련문서 5개와 임의의 주제를 지닌 문서 5개를 실험
(2) 이것은 인공지능에서 널리 사용되는 두 가지 분야로서 Neural Network 과 Genetic Algorithm에 관련된 문서를 각각 9개씩 , 두 가지 주제를 동시에 다루는 문서를 다시 2개를 포함시켜 실험
(3) 각각 신경망, 네트워크 라우팅, GIS, 3차원 그래픽 4가지의 주제를 지닌 문서를 각각 5개씩 실험

5.1 실험1

1) 전문가(사람)에 의한 문서의 분류
   인공지능 관련문서 : 1, 2, 3, 4, 5
   독립적 주제를 지닌 문서 : 6, 7, 8, 9, 10
2) Salton의 방법에 의한 문서의 분류
   분류 1 : 1, 4, 5
   분류 2 : 2, 3
   분류 3 : 6, 7 ,9
   분류 4 : 8, 10
3) 단어의 일치도를 이용한 분류(프로필의 영향력이 15이하면 분류로부터 제외시킴)

문 서	1	2	3	4	5	6	7	8	9	10
프로필1	55	34	31	51	28	4	12	5	9	1

본 논문에서 제안한 단어의 일치도를 이용한 분류에 의해 생성된 프로파일 후보의 숫자는 24개였다 이들 중 10개는 하나의 단어로 이루어진 프로파일 후보로서 고려대상에서 제외되었다. 이를 제외한 나머지 프로파일들의 적합도는 다음과 같다.



(그림 6) 생성된 프로파일들의 적합도

5.2 실험 2

1) 전문가에 의한 문서의 분류
   신경망에 관련된 문서 : 1,2,3,4,5,6,7,8,9
   유전자 알고리즘에 관련된 문서 : 12, 13, 14, 15, 16, 17, 18, 19, 20

두가지에 동시에 관련된 문서: 10, 11

2) Salton의 방법에 의한 문서의 분류

분류 1: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11

분류 2: 12, 13, 14, 15, 16, 17, 18, 19, 20

3) 단어의 일치도를 이용한 분류(프로필의 영향력이 15이하면 그 분류로부터 제외시킴)

문 서	1	2	3	4	5	6	7	8	9
프로필 1	**24**	**53**	**23**	**34**	**42**	**45**	**39**	**45**	**52**
프로필 2	13	4	9	0	4	7	8	8	3

문 서	12	13	14	15	16	17	18	19	20
프로필 1	14	3	8	0	4	9	0	4	7
프로필 2	**55**	**34**	**31**	**23**	**34**	**42**	**23**	**34**	**42**

문 서	10	11
프로필 1	**25**	**54**
프로필 2	**38**	**34**

5.3 실험 3

1) 전문가에 의한 문서의 분류

신경망에 관련된 문서 : 1, 2, 3, 4, 5

네트워크 라우팅에 관련된 문서 : 6, 7, 8, 9, 10

GIS에 관련된 문서 : 11, 12, 13, 14, 15

3차원 그래픽에 관련된 문서 : 16, 17, 18, 19, 20

2) Salton의 방법에 의한 문서의 분류

분류 1 : 1, 2,

분류 2 : 3, 4, 5

분류 3 : 6, 9, 10

분류 4 : 11, 15

분류 5 : 12, 13, 14

분류 6 : 16, 17, 18, 19, 20

3) 단어의 일치도를 이용한 분류(프로필의 영향력이 15이하면 분류로부터 제외)

문 서	1	2	3	4	5	6	7	8	9	10
프로필 1	**26**	**47**	**41**	**42**	**52**	0	4	7	8	8
프로필 2	9	2	0	0	3	**53**	**23**	**34**	**42**	**45**
프로필 3	4	9	0	4	9	0	4	7	9	0
프로필 4	11	4	12	0	7	7	5\	8	2	3

문 서	11	12	13	14	15	16	17	18	19	20
프로필 1	1	3	8	0	4	9	0	4	7	2
프로필 2	7	12	0	12	0	7	7	5	8	2
프로필 3	**31**	**23**	**55**	**34**	**31**	3	9	0	4	8
프로필 4	3	4	9	0	4	**23**	**34**	**23**	**34**	**42**

6. 결 론

본 연구에서 제시한 방법은 기존의 방법에 비해 크게 두 가지 장점을 지닌다.

첫째, 기존의 방법이 각 분류하는데 있어서 문서를 중심으로 함으로서 모든 문서는 각각의 독립된 분류에 소속되었던 것과는 달리 본 연구에서 제시하는 방법은 문서에 있어서 각 분류에 대한 소속의 정도를 표현할 수 있게 하였다. 그 결과 실험 2와 같은 유전자 알고리즘과 신경 망 양자를 동시에 언급하는 문서의 경우 이들이 양쪽의 성격을 동시에 지님을 보여줄 수 있었다. 실세계의 문서가 여러 가지 주제의 혼합을 다루는 경우가 많음을 고려할 때 더욱 효율적인 분류를 제시할 수 있게 된다.

둘째, 분류의 기준을 제시한다. 본 연구에서 제시하는 프로파일 적합도의 계산은 분류로서의 적합도를 표현하게 되는데 이는 문서와 독립된 키워드에 의한 분류를 형성하기 때문에 가능한 것이다. 이를 통해 적절한 기준의 분류항목을 구성할 수 있다. 그러나 문서의 유사도를 중심으로 한 기존의 분류기법은 각각 문서간의 유사도를 중심으로 분류를 수행한다. 그러므로 하나의 클래스를 형성하기에 적절한 문서의 범위를 규정하기 어렵다는 단점이 있다. (실험 1)의 경우 각각의 독립된 주제를 다뤘기 때문에 분류의 형성이 어려운 문서의 경우에 대해서도 인공지능 관련문서와 유사한 형태의 분류를 이루기 때문에 실제 사람이 생각하는 분류와는 다르다. 그러나 본 연구에선 어떤 프로파일에 대해서도 소속도가 낮은 문서일 경우(15이하인 경우) 분류가 힘든 문서로 규정했다.

참 고 문 헌

[1] J.S. Shin, J.H. Kwak and C.H. Lee, "Automatic Classification of Web Documents with Word Accordance of Degree using Probability Model," Proceedings of ICOIN 13, Jan., 1999, 6A-3.1-6A-3.4

[2] K. Sparck Jones, "A Statistical Interpretation of Term specificity and Its Application in Retrieval," Journal of Documentation, 29 : 4, pp.351-372

December, 1973.

[3] G. Salton and M.J. McGill, 'Introduction to Modern Information Retrieval,' McGraw-Hill, New York, 1983.

[4] C. Buckley, G. Salton and J. Allan, "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, 1994

[5] Thorsten Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," CMU-CS-96-118, March 1996.

[6] Marko Balabanovic and Yoav Shoham, "Learning Information Retrieval Agents: Experiments with Automated Web Browsing," Proc. 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, March 1995, AAAI press.

[7] Salton, G., and Buckey, "Term weighting approaches in automatic text retrieval," Technical Report pp.87-881, Cornell University, Department of Computer Science 1987.

[8] Amit Singhal, Chris Buckley, Mandar Mitra, "Pivoted Document Length Normalization," SIGIR pp.21-26. 1996.

[9] John Davies, Richard Weeks, Mike Revett & Andy McGrath, "Using Clustering in a WWW Information Agent." URL : http://www.labs.bt.com/jasper /html/jasclus2.htm

[10] Gerard Salton and Chris Buckly, "Improving Retrieval Performance by Relevance Feedback," 1990

[11] C.J. van Rijsbergen, 'Information Retrieval,' Butterworths, London. 2nd Edition, 1979.

신 진 섭

e-mail : jssjon@tjhealth.ac.kr

1986년 충남대학교 계산통계학과 졸업

1989년 건국대학교 대학원 전자계산학과 졸업(공학석사)

1995년 건국대학교 대학원 전자계산학과 박사과정 수료

1992년~현재 대전보건대학 사무자동학과 전임강사

관심분야 : 웹 에이전트, 지능형 정보검색, 전자 상거래

이 창 훈

e-mail : chlee@kkucc.konkuk.ac.kr

1975년 연세대학교 수학과 학사

1977년 한국과학기술원 전산과 석사

1993년 한국과학기술원 전산과 박사

1977년~1980년 고려 System(System Engineer)

1980년~현재 건국대학교 컴퓨터공학과 교수

관심분야 : 인공지능, 전문가 시스템, 에이전트 시스템, 정보검색