문서지문기법을 이용한 웹 문서의 자동 분류

김진화
서강대학교 경영학과
jinhwakim@sogang.ac.kr

Abstract

As documents in webs are increasing explosively due to the rapid development of electronic documents, an efficient system classifying documents automatically is required. In this study, a new document classification method, which is called Document Finger Print Method, is suggested to classify web documents automatically and efficiently. The performance of the suggested method is evaluated alone with other existing methods such as key words based method, weighted key words based method, neural networks, and decision trees. An experiment is designed with 10 documents categories and 59 randomly selected words. The result shows that the suggested algorithm has a superior classifying performance compared to other methods. The most important advantage of this method is that the suggested method works well without the size limits of the number of words in documents.

1. 서 론

1.1 연구 동기

인터넷의 발달로 인해 정보의 양은 폭발적으로 증가하고 있다. 컴퓨터를 통해 접하게 되는 대량의 전자문서를 효과적으로 분류한다는 것은 쉽지 않은 일이다. 이처럼 문서의 내용을 습득하는 과정을 원활히 하기 위해서는 문서에 대한 분류 작업이 필요하며, 방대한 문서의 분류작업을 수동으로 하는 것은 비효율적인 일이 되었으며 자동 문서 분류 방법의 필요성이 대두되고 있다. 방대한 문서를 수작업으로 여러 사람이 분류할 경우 개개인의 지식기반이 다르게 때문에 같은 문서를 상이하게 분류 할 수 있다. 이러한 객관적이지 못한 방법으로 문서를 분류한다는 것은 사용자로 하여금 혼란을 일으킬 수 있다. 일반적인 문서 분류 작업은 각 분야에 대한 전문가를 두어 해결하려 하지만 이 방법도 효율적이지 못하다. 반면 자동 문서분류 방법에 의해 문서를 분류하면 작업 비용을 최소화 시킬 수 있으며 객관적인 기준으로 문서를 분류함으로 서 문서분류 문제에 대한 효과적인 해결책을 제시할 수 있다.

문서 분류란 문서의 내용을 작업자가 읽고 문서를 미리 정의한 범주로 분류하는 작업이다. 일반적으로 분류라고 하면 수작업을 통해 분류하는 것을 말한다. 수작업을 통해 문서를 분류하는 과정에서 작업자의 지식기반에 따라 다르게 분류 될

수 있다. 이와 같은 분류 작업을 자동으로 하기 위해서는 자연어 이해 및 처리기술이 필수적이다. 그러나 현재의 자연어 처리 기술로는 만족할만한 분류 결과를 얻기는 어려운 실정이다. 자동 문서 분류란 이러한 작업들을 수 작업이 아닌 컴퓨터를 이용하여 자동으로 문서를 분류하는 것을 말한다. 자동 문서 분류나 자동 문서 요약에 쓰이는 기법들은 정보검색에서 쓰이는 기법들을 많이 이용하고 있다. 정보검색(Information Retrieval)이란 정보 항목들에 대한 표현, 저장, 조직, 접근을 다루는 것을 말한다(Hayes and Weinstein, 1991).

자동 문서 분류 시스템에서 단어의 빈도수와 가중치를 이용해 연구되어 온 것 중에서 가장 보편적인 방법이 TFIDF 를 이용하여 문서를 자동으로 분류하는 방법이다. TFIDF 란 문서 내의 해당 단어에 대한 출현 빈도(term frequency)와 출현하는 문서의 개수(document frequency)를 이용하는 것으로, 이 방법은 카테고리 정보를 충분히 이용하고 있지 않아 정확도 면에서는 떨어지지만, 분류문제에 대해 간단하고 쉽게 접근할 수 있도록 문제를 단순화 시킨 것이다(Yang and Pedersen, 1997). 본 논문은 TFIDF 보다 발전된 방법으로 문서지문기법을 이용하여 간단하면서 정확도도 높은 문서 분류 방법을 제안한다.

1.2 연구 방법

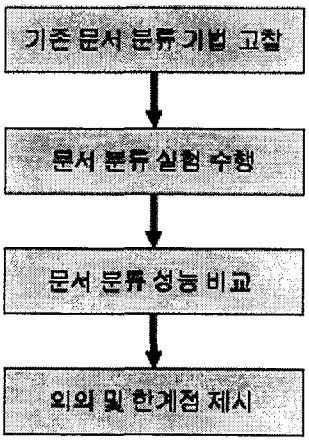
문서의 자동분류에서는 일반적으로 기계학습을 이용하여 미리 학습해 둔 범주 중 하나로 문서를 분류하는 처리이며, 이때 사용하는 기계학습은 개개의 사례를 분석하여 일반적은 규칙이나 함수를 찾아내는 귀납학습법이 보편적이다. 이를 위해 사용될 수 있는 데이터 마이닝은 많은 양의 데이터를 분석하여 의미있고 이해 가능한 패턴(pattern), 즉 지식을 찾아내는 분야로서, 특별히 텍스트 형태의 데이터를 분석하는 것을 텍스트 마이닝(Text Mining)이라고 한다(Trybula, 1999; Berry and Linoff, 2000).

텍스트로 된 문서들은 문서의 성격상 문서의 양이 늘어남에 따라 테이터 처리 시간이 기하급수적으로 증가하게 된다. 그러므로 문서 분류 알고리즘은 문서를 얼마나 정확하게 분류하는가 하는 적중률 못지 않게 문서를 얼마나 빠른 시간 내에 만족할만한 수준으로 분류해 내는가 하는 효율성도 중요하다. 따라서 단어의 양이 많아지더라도 분류성능도 적당하면서 빠른 시간내에 분류할 수 있는 새로운 분류 알고리즘이 필요하게 되었다.

본 연구에서는 연결빈도행렬(Connection Frequency Matrix)을 이용한 문서지문기법을 문서자동 분류 방법으로 제안한다. 이 제시된 알고리즘의 성능 비교를 위한 방법으로 미리 제시된 50 개의 단어를 이용하여 경제, 정치, 사회, 문화, 예술, 건강, 과학, 레져, 교육, 학문의 10 개 분야에 걸쳐 각 분야에 맞게 피 실험자 들에게 문장을 작성토록 하였으며, 작성된 문장이 사용된 알고리즘을 이용해 원래 문서의 내용과 의도대로 분류가 되는지 측정하였다. 기존의 신경망 방법과는 달리 본 연구에서는 보다 더 인간의 뇌를 시뮬레이션 할 수 있도록 하였다. 비교되는 문서 분류 방법으로는 색인어 방식의

Keyword Matching, Weighted Keyword Matching 방식과 의사결정나무, 그리고 인공신경망 방식이며 실험은 크게 세 단계로 나누어진다.

첫 번째 단계에서는 변수를 선정하게 되는데 독립변수로는 사전에 무작위로 선정된 50 개의 단어(word)이며 종속변수는 도서 분류법에 의한 주제별 분류방법을 사용하여 10 개 분야 선정하였는데, 선정된 10 개 분야는 인터넷 신문사의 기사 분류와 검색 사이트 야후(www.yahoo.co.kr), 네이버(www.naver.com), 심마니(www.simmani.com)등의 대표적인 주제별 검색 엔진에서 사용되고 있다. 자료 수집을 위한 두 번째 단계에서는 설문조사를 통하여 10 개 분야(category) 별로 각 분야의 성격에 맞는 문장을 작성하게 하며, 제시된 50 개의 단어 중 5 개 이상을 사용하여 작성토록 하였다. 세 번째 단계에서 문서의 분류실험을 수행하였으며 비교 대상인 알고리즘과 성능을 비교하였다. 연구 절차를 도식화 하면 <그림 1>과 같다.



<그림 1> 연구 절차

2. 이론적 배경

문서들은 전형적으로 자연 언어 형식으로 쓰여진 텍스트를 나타내는 캐릭터들의 순서열로 저장된다(Lewis, 1998). 정보검색 분야에서 문서를 표현하는 캐릭터 스트링을 변형하는 방법에 대한 많은 연구가 진행되어 왔다. 텍스트를 표현하기 위한 많은 통계학적이고 언어적이며 지식 기반 기술을 사용한 연구들이 정보 검색 분야에서 진행 되었다(Lewis, 1998). 그러나 언어적 분석이나 지식기반이 없는 단순한 문서 표현 방법은 다른 방법들과 비교해 비슷한 성능을 보여준다(Mehran and Sahami, 1998). 따라서 복잡한 전처리 과정을 거치지 않는 단순한 텍스트 표현법을 사용하더라도 시간과 계산의 복잡성 측면에서 이득이 되면서 다른 복잡한 처리를 가지는 방법들과 비슷한 효율을 얻을 수 있기 때문에 대부분의 정보 검색과 문서 분류 시스템에서는 단순한 단어 모델을 사용한다.

가장 단순하면서도 널리 사용되는 문서 표현법은 텍스트를 단어들의 집합으로 간주하는 방법이다. 이러한 방법에서 고려해야 할 부분은 문서들로부터 단어들을 추출하는 방법이다. 따라서 단어들을 추출하기 위해 기존의 형태소, 어근 추출(word stem) 법 등의 방법들을 쉽게 사용할 수 있다. 한국어 문서 처리에서 대표적으로 사용되고 있는 방법은 형태소 분석법이다. 자연언어 처리 기법에서 개발된 형태소 분석 기법은 한국어 문서를 표현하고 이해하는데 효과적이다. 단어 벡터 모델은 형태소 분석의 복잡한 부분인 파싱(parsing)을 통한 의미 분석을

생략하고 명사만을 추출하여 구성할 수 있다.

2.1 텍스트 마이닝

지식기반 사회가 도래함에 따라 대량의 지식 정보에 대한 체계적인 관리와 효율적인 검색 기능이 필요하게 되었다. 이에 따라 최근에 대량의 전자 문서로부터 의미 있는 지식 정보를 효과적으로 발견하기 위한 KDD 시스템에 대한 연구가 활발하게 전개되고 있다. 데이터마이닝은 KDD 시스템에서 가장 핵심적인 역할을 수행하는 요소이다.

2.1.1 데이터마이닝

데이터 마이닝 이란 대량의 데이터로부터 유용한 패턴이나 모델을 발견하기 위한 다양한 기법을 말한다. 데이터마이닝에 대한 기존의 연구 기법은 크게 분류 기법, 연관규칙 탐사, 순차패턴, 클러스터링 등으로 구분된다(Aggarwal and Yu, 1998). 분류 기법은 과거의 데이타로부터 정보를 추출하여 미리 정해진 카테고리로 분류하기 위한 규칙을 생성하는 기법으로 통계학이나 신경망 분야에서 연구되고 있다(Mehta and Agrawal, 1996; Shafer 등, 1996). 연관규칙 탐사는 어떤 사건이 동시에 발생하는 연관성에 관한 것으로 지지도(support)와 신뢰도(confidence)를 바탕으로 하여 각 항목간의 연관성을 찾는 기법이다(Rarawal and Imielinski, 1993). 순차패턴은 연관규칙에 시간이라는 개념을 포함하여 순차적으로 발생할 가능성이 큰 항목집합을 발견하는 기법으로 AprioriAll 과 AprioriSome 등과 같은 알고리즘이 연구되었다(Agrawal 등, 1995). 클러스터링은 모집단에 대한 사전 정보가 없는 경우 관측값들 사이의 거리를 계산하여 전체를 소집단으로 분할하는 기법이다.

최근에 전자 문서의 수가 급격하게 증가함에 따라 데이타마이닝 기법을 문서에 적용하기 위한 텍스트마이닝에 대한 연구가 활발하게 전개되고 있다. 텍스트마이닝이란 대량의 문서로부터 패턴을 탐사하여 유용한 지식 정보를 찾기 위한 데이터마이닝 기법의 하나이다. 그러나 문서의 비정형화, 특징의 불규칙성, 일반용어의 과다 출현 그리고 문서 길이의 불규칙성 등과 같은 문제점들로 인해 지식 발견에 많은 어려움을 겪고 있다. 텍스트마이닝 기법은 크게 분류 기법과 클러스터링으로 구분된다. 분류 기법은 문서의 내용에 따라 사전에 정해진 카테고리에 문서를 할당하기 위한 방법으로 k-Nearest Neighbor(Yang and Pedersen, 1997), 의사결정 트리(Lewis and Ringuett, 1994) 등과 같은 통계학적 방법과 Support Vetor Machine(Joachims, 1998), 신경망(Wiener, 1995) 등과 같은 기계학습 기법이 연구되고 있다. 클러스터링은 문서 내용의 유사도에 따라 소집단으로 분할하는 방법으로 계층적 클러스터링과 분할 클러스터링으로 나뉘어진다. 계층적 클러스터링 기법은 클러스터의 유사도에 따라 가장 유사한 클러스터 쌍을 병합하는 단계를 반복하는 Agglomerative 방법과 하나의 클러스터에서 출발하여 개별 클러스터로 분할하는 단계를 반복하는 Divisive 방법이 있다(Hamdouchi, 1989). 분할 클러스터링은 Centroid 의 반복 계산을 통하여 초기 값으로 주어진 k 개의

평면적인 클러스터로 분할하는 방법으로 k-Means 알고리즘이 대표적인 기법이다(Kaufman, 1990).

2.1.2 문서 자동 분류 시스템

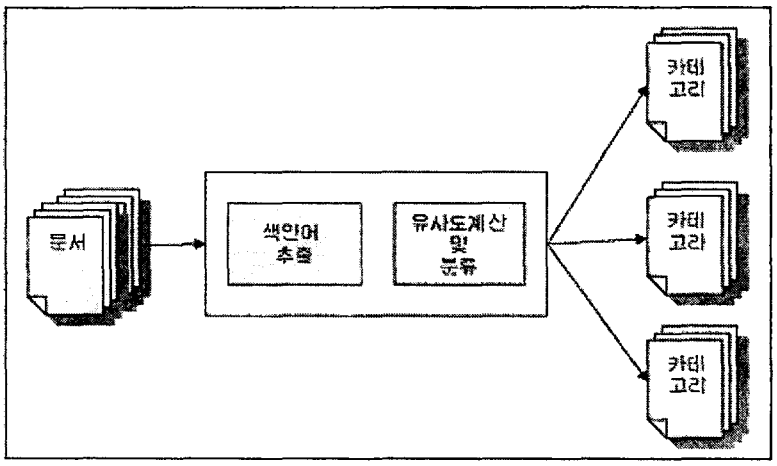
문서자동분류 시스템은 평면 또는 계층 구조에서 문서를 유사한 카테고리로 분류하여 대량의 문서를 체계적으로 관리하기 위한 시스템이다. 문서자동분류 시스템은 일반적으로 분류 기법을 이용하는 방법과 클러스터링을 이용하는 방법으로 나누어진다. 분류 기법은 사전에 정의된 분류 체계에 따라 문서를 유사한 카테고리로 배치하는 방법이다. 그리고 클러스터링은 사전 분류 체계 없이 문서간의 유사도에 따라 유사한 문서를 집단화하는 방법이다(정영미, 1987).

그 동안 대량의 문서를 관리하기 위해 사전에 정의된 카테고리에 따라 문서를 분류하는 방법을 많이 사용해 왔다. 특히 계층 구조에서는 하위 카테고리일수록 세부적인 정보를 포함하기 때문에 보다 정확한 정보 검색이 가능하다.

계층 구조를 사용하는 대표적인 예로는 Yahoo(http://kr.yahoo.com)를 들 수 있다. Yahoo 에서는 "건강과 의학", "비지니스와 경제", "컴퓨터와 인터넷" 등과 같은 카테고리와 하위의 세부 카테고리들로 계층 구조를 이루어 문서를 관리하고 있다. 여기서 각 카테고리의 문서는 사람이 직접 카테고리를 결정하여 등록한 것이다. 그러나 사람이 직접 수작업으로 대량의 문서를 분류하기 어려운 관계로 자동화된 문서 분류시스템이 필요하다.

2.1.3 문서 자동 분류 시스템의 일반적인 구성

일반적으로 평면 구조의 문서 분류 시스템은 색인어 추출 과정, 유사도 계산 그리고 분류 과정으로 구성된다. 색인어 추출 과정은 각 카테고리의 특징을 대표하는 색인어 집합을 찾아내는 과정이다. 그리고 유사도 계산 및 분류 과정은 문서와 카테고리간의 유사도를 계산하여 해당 문서를 가장 유사한 카테고리로 분류하는 과정이다..



<그림 2> 문서 자동 분류 시스템의 일반적인 구성도

계층 구조에서도 계층별로 색인어 추출과 유사도 계산 및 분류 과정을 반복한다. <그림 2>는 문서자동분류 시스템의 일반적인 구성도이다.

2.2 키워드 집합 (Keyword Set)을 이용한 분류 방법

2.2.1 색인어 추출

그 동안 연구된 대표적인 색인어 추출 기법은 DF(Document Frequency), IG(Information Gain), MI(Mutual Information), x^2 Statistic, TS(Term Strength), TF*ICF 등이 있다(조광제, 김준태, 1997). DF 기법은 특정 용어가 출현하는 문서의 빈도수에 따라 색인어로 추출하는 방법으로 간단하지만 성능이 우수한 편이다. IG 기법은 이전 분류 모델에서 주로 사용되는 방법으로 카테고리에 속하는 문서에서 특정 용어의 출현 혹은 부재에 대한 정보를 이용하는 방법이다. MI 기법은 용어 및 카테고리에 대한 분할표를 이용하여 용어와 카테고리간의 종속성 여부를 판별하는 기법으로 용어 연관성에 관한 통계적 언어 모델링 분야에서 널리 사용되고 있다. x^2 Statistic 은 전통적인 통계적 기법으로 MI 기법처럼 분할표를 이용하여 용어와 카테고리간의 종속성 여부를 판별하는 방법이다. 또한 x^2 값은 정규화된 값이므로 MI 기법과는 달리 동일한 카테고리내의 다른 용어들과 직접 비교가 가능하다. x^2 Statistic 의 단점은 출현 빈도가 낮은 용어에 대해서는 신뢰도가 떨어지는 점이다(Dunning, 1999). TS 기법은 관련된 문서에서 공통적인 사용 정도에 따라 용어의 중요도를 측정하는 방법이다. 즉 학습 문서간의 코사인 계수를 계산하여 유사 문서 집합을 구성하고 유사 문서 집합 내에서 공통적으로 사용된 용어를 찾아내는 방법이다. TF*ICF 기법은 특정 카테고리 내에서 용어가 출현하는 빈도(TF, Term Frequency)와 용어가 출현하는 카테고리 수의 역빈도(ICF, Inverse Category Frequency)를 이용한 방법으로 TF*IDF 기법과 유사하다. 그러나 TF*IDF 기법은 문서 내에서 용어의 중요도를 나타내는 반면 TF*ICF 기법은 카테고리 내에서 용어의 중요도를 나타낸다.

2.2.2 TFIDF (term frequency inverse document frequency) 알고리즘

TFIDF 학습기법은 문서에서 단어들을 추출하여 단어목록과 가중치로 구성된 테이블을 만들고, 이것을 이용하여 문서를 분류하는 방법이다. TF(term frequency)는 한 키워드가 속해있는 문서에서 나타나는 횟수를 말하며, IDF(inverse document frequency)는 DF 의 역으로서, DF 는 키워드가 발견된 문서들로부터 몇 개의 문서에서 나타나는가를 측정한 수치이다. IDF 의 수치가 클수록 변별력이 크다는 것을 의미며, 한 키워드의 가중치를 구하는 식은 <표 1>과 같다(백혜정, 박영택, 윤석환, 1999).

앞서 많은 문서들 중에서 그 문서들을 대표할 수 있는 특징을 추출하기 위해 단어의 빈도수(term frequency)를 많이 이용한다고 언급했다. 그러나 단어의 빈도수가 높은 것이 그 문서를 정확히 대표하는 단어가 된다고 확신할 수 는 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*IDF 는 역 문서 빈도수(inverse document frequency)를 단어의 빈도수와 같이 적용함으로 서 그 문서를 대표하는 단어들을 효율적으로 찾을 수 있는 알고리즘이다.

<표 1> TFIDF 의 키워드 가중치

$$W = tf \cdot idf$$

W : 키워드의 가중치

tf : 현재문서의 키워드 빈도수

idf : 키워드가 포함된 문서들의 빈도수의 역

2.2.3 키워드 집합 (Keyword set)

키워드 집합(Keyword set)에 의한 방법은 패턴 매칭을 사용하는 방법으로 패턴 매칭에 사용할 문장 형태를 단어들의 집합으로 표현한다. 패턴의 정의에 사용되는 어휘들은 W_i 라 하면 키워드 집합에 의한 분류 규칙의 형태는 <표 2>와 같다.

<표 2> Keyword 분류 규칙

Keyword Set Rule: { $W_1, W_2, \ldots, W_n$ } → C_K

- W_i : 사용되는 어휘
- C_K : 분류된 카테고리

이 방법은 한국어 문장이 많은 경우에 자유로운 어순이 가능하다는 점과, 특정 구문 표현만 분류의 단서가 되는 것이 아니라 특정 단어들이 한 문장 안에 동시에 나타나기만 하면 단어들이 멀리 떨어져 있는 경우에도 단서가 될 수 있다는 점을 고려한 것이다.

<표 3> 키워드 매칭(Keyword Matching)

문장 1 : "빗길을 과속으로 달리던 승용차가 중앙선을 넘어 앞에 오던 화물차와 정면 충돌하였다."

문장 2 : "어제 밤에 일어난 고속도로 정면 충돌 사건의 원인은 과속 주행하던 승용차의 중앙선 침범 때문인 것으로 밝혀졌다."

키워드 집합에서는 집합내의 단어 수는 분류의 정확도 및 분류율에 영향을 미친다. 집합내의 단어 수가 많아질수록 그 키워드 집합과 매칭이 되는 문장 수는 적어져서 분류율은 감소하며, 더 엄격한 매칭으로 인해 분류의 정확도는 증가하게 된다.

2.3 신경망 (Neural Network)

신경망은 인간의 두뇌에 있는 대규모의 뉴런(neuron)들의 상호 연결되어 있는 구조를 모델링한 것이다. 신경망 모델은 사람의 뉴런과 같은 많은 처리단위가 서로 연결되어 외부로부터 입력되는 여러 정보를 동적인 상황하에서 처리할 수 있는 지능적인 시스템이다.

신경망모형은 인간이 경험으로부터 학습해 가는 두뇌의 신경망 활동을 흉내내어 자신이 가진 데이터로부터의 반복적인 학습 과정을 거쳐 패턴을 찾아내고 이를 일반화함으로써 특히 향후를

예측(Prediction)하고자 하는 문제에 있어서 유용하게 이용되는 기법 매우 복잡한 구조를 가진 데이터들 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형(Flexible nonlinear Model)의 하나이다. 주로 Supervised data 에 적용되어 결과변수(target)에 대한 예측(Prediction)이나 분류(Classification)를 목적으로 감춰진 패턴을 찾고 이를 일반화하는데 이용, 혹은 Unsupervised data 에서 코호넨 맵(Kohonen maps)을 이용하여 데이터의 클러스터링 작업을 수행하는데 쓰인다.

인공지능 기법을 통한 문서분류는 키워드를 이용한 문서분류 방법보다 정확도가 높다. 허나 단어의 분류에 이용되는 단어의 개수가 200 개 이상이 되면 실질적으로 인공지능을 사용하여 문서를 분류하는 일이 불가능하다. 인공지능을 이용한 문서분류를 할 경우 한나의 단어가 한 개의 변수로 설정이 되어야 한다.

3. 연구 설계

3.1 자료수집과 변수선정

3.1.1 자료 수집

본 연구에 사용된 자료는 설문 작성을 통해 얻었다. 인터넷 신문에서 사용하고 있는 10 개 분야(category)를 정했으며 이 10 개 분야에서 사용된 단어 중 사용 빈도수가 높은 50 개를 선정하였다. 수집된 data 는 500 개의 record 로 구성되었으나, 최종적으로는 470 개의 record 가 최종 선택 되었다. 그 중에서 370 개는 train data 로 100 개는 test data 로 구성하였다.

3.1.2. 변수선정

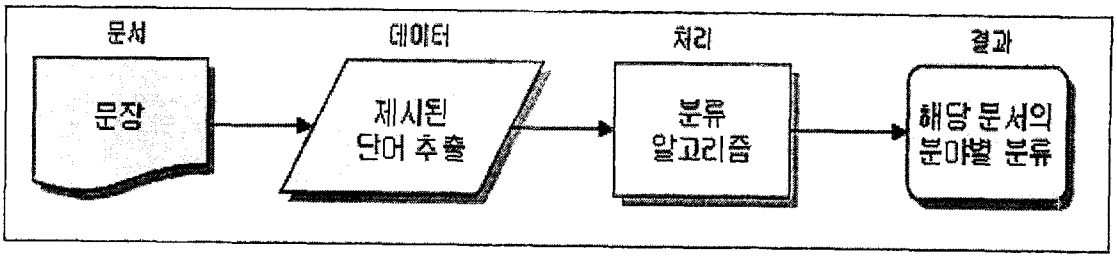
독립변수는 50 개 단어(word)로서 w1, w2, w3, . . w50 이다. 이 변수들은 무작위로 다양한 단어를 선정하였다. 종속변수는 10 개 분야(category) 이며 각 변수는 c1, c2, c3, . . c10 로 구성된다. 변수에 대한 설명은 첨부 1 에서 설명된다.

문장 내에 어떠한 단어가 사용되었으면 1, 사용되지 않았으면 0 으로 표시하였다. 예측 변수로 사용된 분야(category) 는 도서 분류법에 의한 주제별 분류법을 사용되는 10 개 분야로 선정하였다. 선정된 10 개 분야는 인터넷 신문사의 기사 분류 및 검색 사이트 야후(www.yahoo.co.kr), 네이버(www.naver.com), 심마니(www.simmani.com) 등 대표적인 주제별 검색 엔진에서 사용되고 있다.

3.2 실험수행 방법

3.2.1 연결빈도행렬(CFM; Connection Frequency Matrix)

실험은 문장에서 추출된 단어를 각 알고리즘을 이용해 분류하고 그 분류 정확도를 측정한다. 연결빈도행렬(Connection Frequency Matrix)의 실험 수행 방법은 <그림 3>과 같다. 정렬된 데이터는 50 개의 tuple 과 470 개의 record 로 이루어져 있다. 여기서 tuple 은 단어를 나타내는 독립변수이며, record 는 설문을 통해 해당 분야(category)에 맞게 작성된 하나의 문장이다.



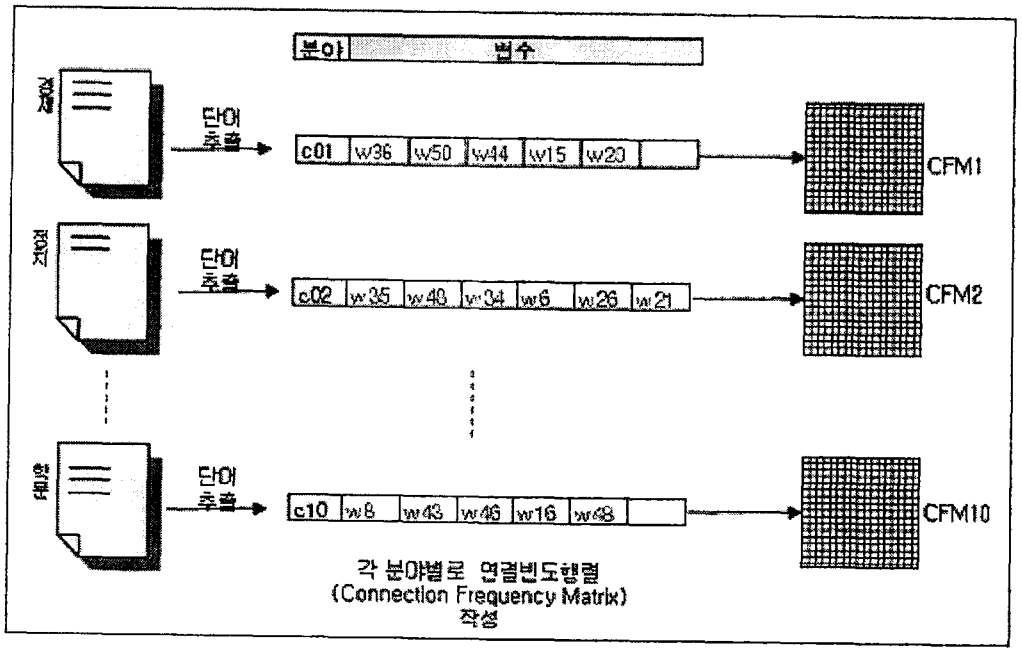
<그림 3> 실험 수행 방법

이 record 는 해당 분야에 맞게 작성된 문장이므로, 실험 편의상 특정 분야에 속한 문서로 가정하고 문서 분류 성능에 대한 예측치로 사용하였다. 문장 내에서 사용된 변수는 1 로, 사용되지 않은 변수는 0 으로 표시하였다.

3.2.2 데이터 분류와 학습

연결빈도행렬(CFM) 알고리즘의 분류 성능 평가를 위해 470 개의 record 중에서 370 개는 training data 로 100 개는 test data 로 나누었다.

Training set 은 문서 특징을 지문형태로 나타내기 위한 CFM 작성을 위해 사용되고, test set 는 예측정확도를 비교하기 위해 사용된다. 이 training 데이터는 c01 부터 c10 까지 각 카테고리 별로 10 개의 연결빈도행렬에 그 정보가 저장된다. <그림 4>에서 보여 지듯이 각 문서에서 추출된 단어들은 각각의 카테고리를 위해 마련된 연결빈도행렬에 그 단어들 사이의 연결정보가 저장된다.

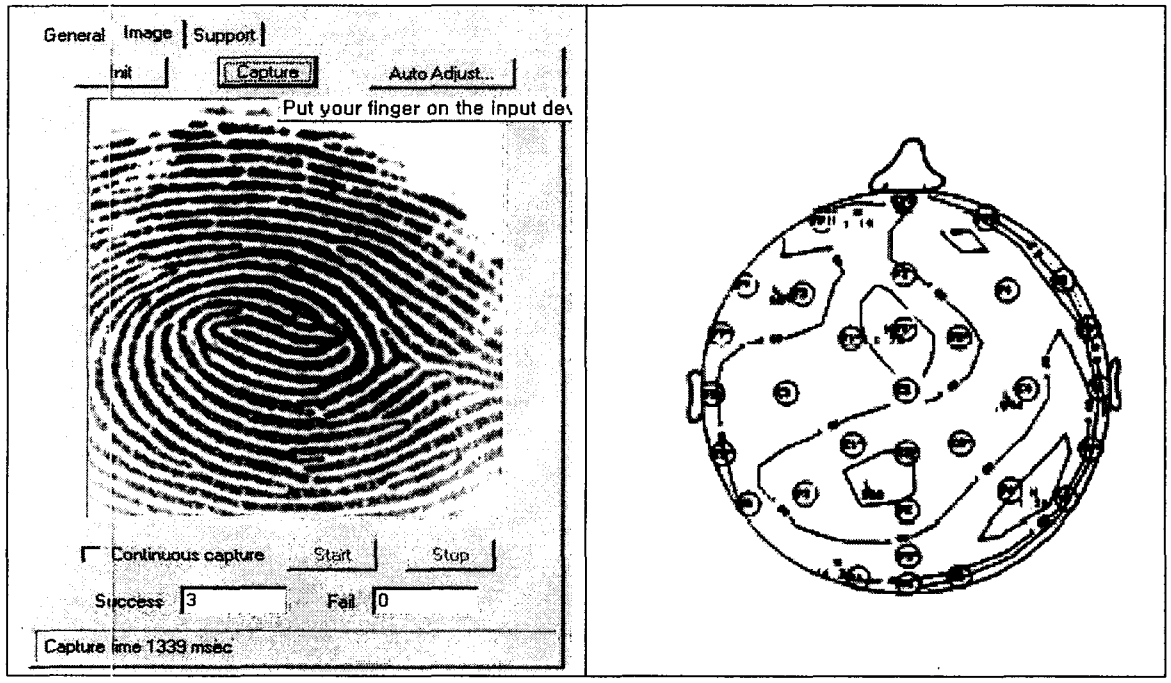


<그림 4> 데이터의 분류

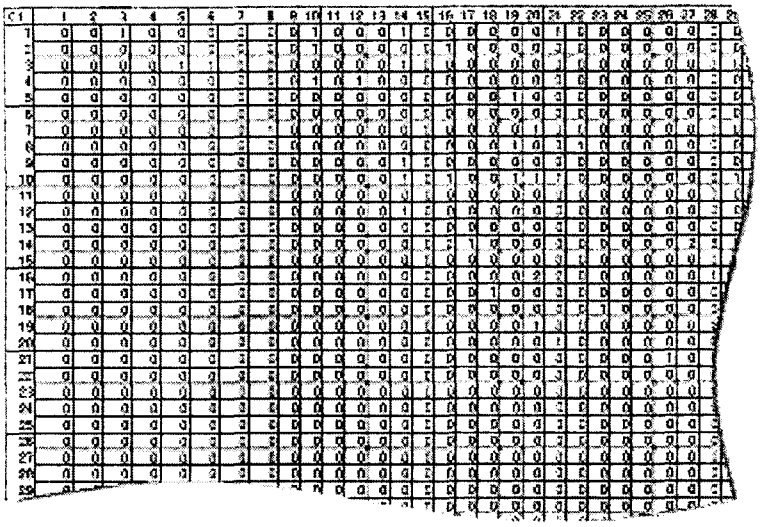
3.2.3. 문서지문 matrix 의 작성

인간 각자에게는 각각의 사람을 구별할 수 있는 개개인만의 지문이 존재한다. 근래에 등장한 지문을 이용한 지능형 door lock 시스템이 그 예라고 할 수 있다. 지문은 땀샘이 융기되어 만들어진 융선과 융선과 융선사이의 골로 구성되어 있다. 지문인식 시스템은 이러한 융선과 골의 모양에 따른 지문의 중심점, 분기점, 단점 등의 특징을 추출하여 비교하여 개개인을 구별한다. 각 개개인이 <그림 5>의 왼편 그림과 같이 그 개인 고유의 특징을 가진 지문이 있듯이 각 문서의 분류는 각각의 고유의 문서 지문을 만들 수 있다. 지문인식에서 융선과 그 융선에 나타난 특징으로 개개인을 구분하듯이 <그림 6>에서 보여 지듯이 단어와 단어 사이의 조합 발생빈도를 2 차원적인 matrix 로 나타낼 수 있다.

이 발생 빈도를 나타내는 2 차원 matrix 는 인공지능의 neural networks 와 그 기능, 학습, 이용 면 에서 아주 유사하다 할 수 있다. 허나 neural networks 의 학습의 개년이 input 과 output 을 가장 잘 설명하는 weights 값을 찾아내는 수학적 기법인데 반해 인간의 학습은 보다 단순한 정보의 반복, 누적이다. 본 연구에서 제시하고 있는 문서지문기법은 이러한 면에서 보다 더 인간의 학습과 가까운 기법이라고 할 수 있다. 어떠한 문장에서 특정 단어의 집합이 사용되면 단순히 이 문서에 이들 단어가 사용 되었다는 정보 외에도 이들 단어가 연관을 갖고 이 문장에 사용 되었다는 관계(relationship)에 관한 정보도 동시에 저장되어야 한다.



<그림 5> 지문인증과 Brain Mapping



<그림 6> 정치 분야 Matrix

<그림 6>은 정치로 분류되는 문서의 지문을 정치로 분류되는 문장들에서 사용된 50 개의 단어로 나타내었다.

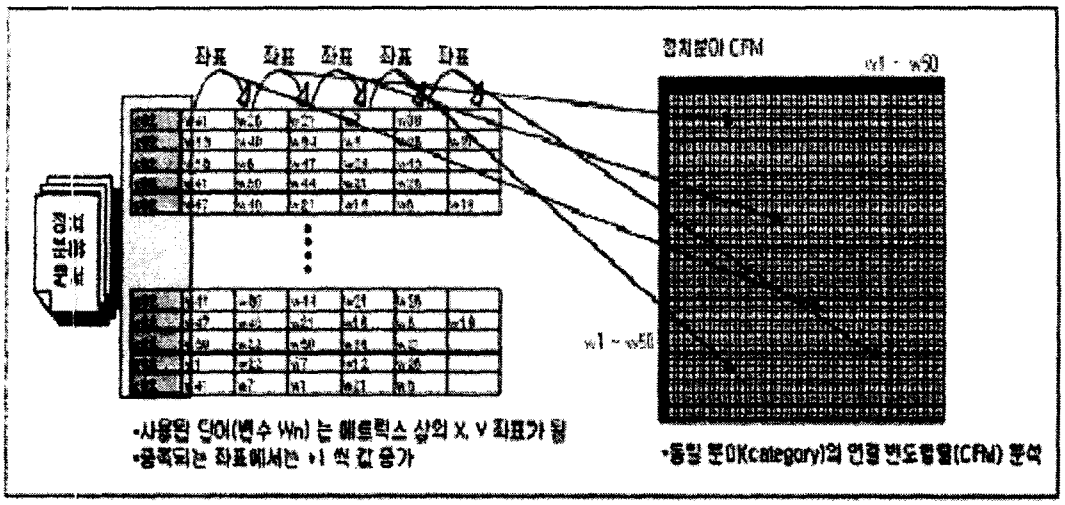
문서지문 Martix 의 x, y 축은 <그림 6>에서 처럼 사용된 단어를 나타내는 변수 w1 부터 w50 까지 순서대로 나열되어 있다. 문서에서 사용된 단어와 그들의 관계는 사용된 빈도수 만큼 각 카테고리 matrix 의 x, y 에 누적된다. 이 matrix 를 통해 단어 간의 연관성을 측정할 수 있으며, 각 카테고리 마다 의 CFM 을 이용하여 문서를 분류한다.

3.2.4 연관성 측정

<그림 7> 같이 경제분야(c01)로 분류된 어떤 문서에서 추출된 단어가 {w3, w37, w44, w50, w35}와 같은 순서로 사용되었다면 matrix 상의 (x, y) 좌표는 (w3, w37), (w3, w44), (w3, w50), (w3, w35), (w37, w44), (w37, w50), (w37, w35), (w44, w50), (w44, w35), (w50, w35) 가 된다.

#2								#2						
c1	경마	전세	통상	흥작	일본		2	c1	w3	w37	w44	w50	w35	
c2	일본	군대	항공기	연료	컴퓨터		2	c2	w35	w6	w47	w29	w43	
c3	일본	지방자치	주차	준수	보험		2	c3	w35	w41	w39	w40	w20	
c4	종교	전문가	로마	유출	종교		2	c4	w38	w36	w13	w34	w38	
c5	미모	배달	여행	로마	보행		2	c5	w17	w18	w28	w13	w19	
c6	딱	예살	내장	담배	분노		2	c6	w12	w15	w9	w10	w21	
c7	오리	고기	예살	딱	미모		2	c7	w31	w4	w15	w12	w17	
c8	일본	로마	항공기	여행	보험		2	c8	w35	w13	w47	w28	w20	
c9	지방자치	선발	종교	컴퓨터	기초과학	유료	2	c9	w41	w22	w38	w43	w8	w32
c10	일본	기초과학	선발	선생	등선	컴퓨터	2	c10	w35	w8	w22	w23	w45	w43

<그림 7> CFM 상의 변수 좌표



<그림 8> 연결빈도행렬(CFM)에서의 변수 누적

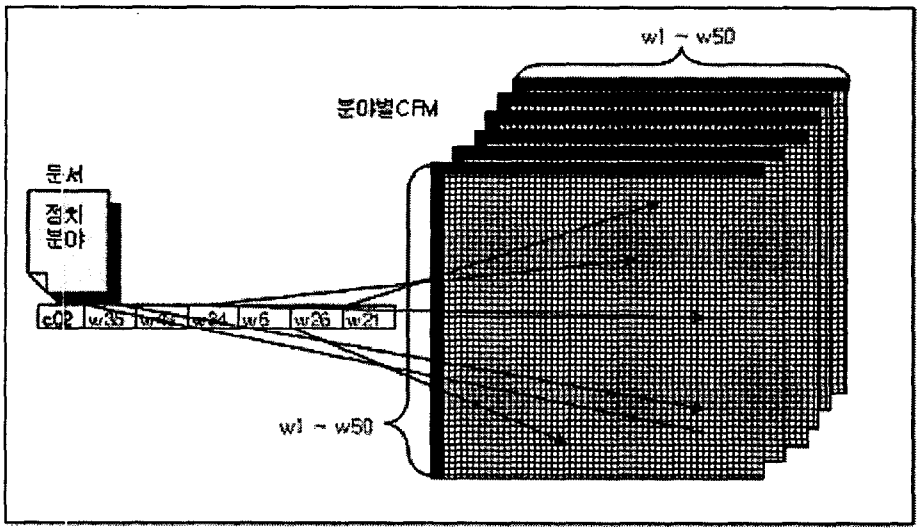
위에서 의해 얻어진 좌표는 <그림 8>에서와 같이 matrix 상에 기록되며, 동일 좌표일 경우 값은 누적된다. 변수의 사용 빈도가 높아질수록 matrix 상의 값도 높아지며 이는 해당 단어의 연관성이 높다는 것을 의미한다. 만약 문장에서 추출된 변수가 n 개 라면 CFM 상의 (x, y) 좌표 개수는 아래와 같다.

경우의 수 = ${}_nC_2$

(n 은 사용된 변수의 개수)

3.2.5 CFM 분류 정확도 측정

위의 순서를 통해 train set 에서는 종속변수인 카테고리의 문서 특징이 CFM 을 통해 나타내게 된다. CFM 의 분류 성능을 측정하는 방법은 아래와 같다. <그림 9>은 정치(c02) 카테고리 CFM 의 예측 방법을 나타내고 있다.



<그림 9> 정치 분야 CFM 정확도 측정

<표 4> 연결빈도행렬의 예측 정확도

문서	연관도										Category	Prediction (%)	Match	정확도 (%)
	CFM 1	CFM 2	CFM 3	CFM 4	CFM 5	CFM 6	CFM 7	CFM 8	CFM 9	CFM 10				
c01	8	4	2	5	1	0	6	0	3	4	c01	O	1	1
c01	2	0	2	0	0	3	0	1	0	0	c01	X		
c01	7	2	2	1	1	2	7	0	2	5	c01	O	2	0.5
c02	3	8	4	4	5	1	0	1	3	1	c02	O	1	1
c02	3	5	4	1	1	5	1	2	1	2	c02	O	2	0.5
c02	1	2	3	1	3	0	0	0	4	0	c02	X		
c02	3	3	3	0	0	1	0	0	2	0	c02	O	3	0.333
c02	1	8	3	2	5	1	0	0	0	0	c02	O	1	1
c03	5	2	4	4	2	1	14	0	4	11	c03	X		
c07	1	3	3	3	4	0	18	0	6	18	c07	O	2	0.5
c07	2	4	1	1	4	0	16	0	11	20	c07	X		
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
c07	0	1	0	0	2	1	21	1	5	13	c07	O	1	1
c07	0	0	2	0	0	2	2	1	0	0	c07	O	3	0.333
c10	6	7	0	3	2	1	20	0	3	22	c10	O	1	1
c10	3	5	4	5	7	1	33	0	8	32	c10	X		
합												57		51.667

Test set 문서들에 사용된 변수를 10 개의 CFM 에 모두 매칭 시켜 단어의 연관도가 가장 높은 CFM 이 해당 문서의 카테고리가 된다. 변수 매칭 방법은 "3.2.4 연관성 측정" 의 방법과 같다. 즉 문서 내에 사용된 변수의 좌표로 매칭을 시켜서 연관도를 구한 다음, 각 카테고리의 예측치와 결과치를 비교하여 예측 정확도를 구한다.

<표 5> CFM 의 예측 정확도

예측 정확도(%) = $\sum$ (1 / Match)
- Match : 중복 적중된 예측

Test set record 는 모두 100 개였으나 <표 4>에서는 지면의 한계로 일부 결과만을 표시하였으며, 51.667 % 예측

정확도는 100 개 record 의 정확도가 모두 포함된 계산 결과이다(<표 5> 참조).

3.3 기존 문서 분류 알고리즘의 성능 비교

3.3.1 키워드(Keyword Matching)

키워드 매칭은 패턴 매칭 사용하며 단어들의 집합으로 표현되는 방법이다. 즉 유사 단어로 구성된 문장을 공유하는 문서일수록 같은 카테고리로 분류될 확률이

<표 6> 카테고리별로 카운트된 변수

	w1	w2	w3	w4	w5	…	w45	w46	w47	w48	w49	w50
c01	4	1	4	2	2	…	0	0	7	0	8	5
c02	4	3	0	2	0	…	0	0	2	0	1	1
c03	6	0	1	2	0	…	0	1	5	3	7	3
c04	4	2	2	3	1	…	1	2	5	3	3	2
c05	7	0	1	1	6	…	3	0	1	5	1	1
c06	1	1	3	15	2	…	0	2	1	1	3	0
c07	3	4	0	0	0	…	2	4	15	3	0	0
c08	3	0	16	1	4	…	4	2	11	10	0	0
c09	7	2	1	2	3	…	0	3	3	1	3	0
c10	0	1	0	1	2	…	0	5	4	0	1	1

높고 집합내의 단어 수가 분류 정확도에 영향을 미친다.

이런 특징 때문에 키워드 매칭의 성능 평가는 다음의 순서로 이루어진다. 먼저, test set 의 문장 내에 사용된 변수의 횟수를 카운트 하여 패턴을 찾아 분류 하고, 다음으로 train set 과 비교해서 그 적중률을 비교한다.

문서 내에 사용된 변수의 횟수를 카운트 하기 위해 <그림 8> 에서 사용된 분야별 CFM matrix 를 활용하였다. <표 6>에서는 각 변수가 카테고리 별로 사용된 횟수가 카운트 되어 있다(지면상 일부 records 생략). c01, c02 . . c10 은 카테고리 이고 w1, w2, w3 . . w 50 은 사용된 단어의 횟수를 카운트 한 것이다.

지면상 일부 record 는 생략하였고, Test set 예측율과 비교한 모든 records 가 포함된 예측 정확도는 <첨부 2>와 같다. 분류 정확도를 계산할 때에는, 문장이 속한 카테고리와 키워드매칭 알고리즘이 예측한 카테고리가 적중했다 하더라도 1 개 이상의 카테고리에 중복 되었다면 해당 문서의 분류 적중률은 1/n 이 된다. 키워드 매치의 문서별 적중률은 21.288% 로 매우 낮게 나타났다. 각 문서의 키워드 매칭은 분류 정확도가 비교적 낮게 나타났으나, 한국어 문서의 경우 처럼 자유로운 어순이 가능하다는 장점이 있다.

3.3.2 가중 키워드(Weighted Keyword Matching)

가중 키워드 방법을 이용한 문서 분류 방법도 키워드 매칭 방법과 마찬가지로 수행하며, 키워드 매칭과의 차이점은 문서 내에 사용된 변수가 n 개일 경우 사용된 빈도를 카운트 하고, 사용된 변수가 W1, W2, ... Wn 일 때 해당 단어에 가중치를 부여한다. 가중 키워드의 알고리즘에서는 변수가 문장 내에서 사용되었다 할지라도, 한번 사용된 단어는 사용되지 않은 것으로 간주하여 제외하였다. 그 이유는 한번 사용된 변수는 변별력의 차이를 알 수 없기 때문이다.

<첨부 3>은 가중 키워드 매칭의 문서 분류 정확도를 나타낸다. 키워드 매칭과 마찬가지로 일부 record 는 생략 되었으며, 문장이 속한 카테고리와 가중 키워드 매칭 알고리즘이 예측한 카테고리가 적중했다 하더라도 1 개 이상의 카테고리에 중복 되었다면 해당 문서의 분류 적중률은 1/n 로 계산하였다. 가중 키워드 매칭의 문서 분류 적중률은 31.5% 나타났다.

3.3.3 신경망(Neurual Network)

신경망 기법을 사용하기 위해 사용한 툴은 데이터 마이닝 솔루션인 Clementien 8.0 이며 솔루션에 포함된 신경망 모델을 이용하여 분류 정확도 측정하였다. 이 테스트에서의 SPSS 데이터는 연결빈도행렬에서 정렬한 데이터이며, 신경망을 모델링하여 만들어진 Neural Network 모델을 교차테이블에서 독립변수를 입력변수로 하고 종속변수를 출력변수로 하는 교차표를 출력하는 방법으로 문서 분류에 대한 성능 측정을 수행하였다. 여기서 독립변수는 w1 부터 w50 까지의 50 개의 단어이며 종속변수는 c01 부터 c10 까지의 10 가지의 카테고리이다. 실험은 디폴트 옵션으로 수행하였으며 출력 결과는 다음과 같다.

<표 7> 신경망 문서 분류 정확도

CATEGORY	1	2	3	4	5	6	7	8	9	10
1	26	2	0	0	0	0	2	5	1	1
2	3	26	0	0	2	0	0	1	1	4
3	19	5	2	0	4	0	4	1	2	0
4	8	9	0	2	5	1	3	8	0	1
5	3	8	1	1	16	0	4	3	0	1
6	19	0	0	0	4	3	6	3	2	1
7	1	3	0	0	1	0	21	0	1	10
8	0	0	0	3	0	1	0	33	0	0
9	11	8	2	1	1	1	8	1	1	3
10	2	3	0	0	3	0	14	2	2	11

- 정확도 : 38.108 %
- 입력 레이어 : 50 개의 뉴런
- 숨김 레이어 1 : 3 개의 뉴런
- 출력 레이어 : 10 개의 뉴런

3.4 실험결과 분석

실험 결과에서 최초 예측 정확도를 단순 카운트 했을 때는 prediction 이 57% 로 높게 나타났다. 그러나 문서가 동일한 분류 정확도를 갖는 한 개 이상의 CFM 에 match 되었을 때 이는 정확한 예측률이라고 할 수 없다. 왜냐하면 실제로 하나의 문서가 두 개의 카테고리로 분류 될 수는 없기 때문이다. 따라서 prediction 이 적중한 결과 중 연관도가 동일하게 나온 각기 다른 CFM 이 존재할 때는<표 7> 과 같이 중복된 CFM 만큼 적중률을 나누어서 계산하였다. 이렇게 보정된 연결빈도행렬의 문서 분류 정확도는 51.667% 으로 측정되었다.

<표 8> 문서 분류 알고리즘 예측 정확도 비교

분류 방법	Keyword Matching	Weighted Keyword Matching	Neural Network	Connection Frequency Matrix
분류 정확도 (%)	21.28	31.5	38.108	51.66

문서 분류 알고리즘의 예측 정확도를 비교한 결과 본 논문에서 제시한 연결빈도행렬(CFM; Connection Frequency Matrix) 알고리즘은<표 8> 과 같이 51.66 %로 비교된 문서 분류 알고리즘 중 높은 예측 정확도를 보여주고 있다.

4. 결 론

4.1 연구결과 요약 및 시사점

빠르게 변하는 인터넷 환경 하에서 급증하는 정보를 효과적으로 구하는 방법이 절실히 요구 된다. 특히 전자문서의 경우 양적인 면에서 사람이 수작업으로 분류할 수 있는 범위를 벗어났다. 따라서 기계적인 처리를 통한 자동분류가 피할 수 없는 상황이다. 본 논문에서는 웹 상의 문서를 자동으로 분류하는 기법인 연결빈도행렬(CFM; Connection Frequency Matrix)에 대해 소개하였다. 이와 같은 연구를 수행한 본 논문의 의의는 다음을 들 수 있다.

첫째, 문서 자동 분류를 위한 새로운 기법이다. 예측 정확도 역시 기존의 분류 알고리즘과 비교가 가능하다. 활용 여부에 따라 문서분류 뿐만 아니라 문서내용 요약이나 맞춤 지식 추천에도 응용될 수 있을 것이다.

둘째, 변수가 되는 문서의 단어 수에 영향을 받지 않는다. 기존의 문서 분류 알고리즘은 저장된 문서의 양이 많아지면 그 수행시간이 급격히 증가한다는 단점을 갖고 있다(이재식, 2002). 그에 비해 연결빈도행렬(CFM)은 matrix 크기에 제한이 없기 때문에 빠른 문서 분류를 수행할 수 있다. 문서 내에서 사용된 단어가 많아졌을 때에는 matrix 크기만 키워 주면 되기 때문에 분류 처리 시간이 오래 소요되지 않는다.

셋째, 연결빈도행렬(CFM)은 신경망(Neural Network) 보다 더 비슷하게 인간 뇌의 신경(뉴런; neuron)의 상호 연결 구조를 시뮬레이션 하였다. 뇌의 뉴런들은 다른 뉴런들로부터 입력을 받아 하나의 출력을 생성하는데 한 뉴런의 출력은 다른 뉴런의 입력 정보가 된다. 이 과정에서

패턴 추출 및 예측은 출력 값이 나올때까지 반복 학습을 수행하게 되는데, 연결빈도행렬(CFM)에서는 Matrix 상에서 단어간의 연관도가 즉시 표현되므로 반복 학습이 필요하지 않다.

4.2 연구의 한계점 및 향후 연구 방향

본 논문의 실험에 사용된 자료 수집 과정이 작문을 통한 설문 조사 방법이었기 때문에 수집에 어려움이 있었으며 data set 의 사이즈가 작았다. 작성한 문장을 하나의 문서로 가정하고 얻은 실험 결과가 보다 현실적으로 적용되기 위해서는 실제 웹 문서와 데이터베이스를 이용해야 할 것으로 생각된다. 문서 분류 실험 결과 연결빈도행렬(Connection Frequency Matrix)에서 사용된 이진형 독립변수는 총 50 개로 적지 않은 수이지만 실제 웹 문서를 분류하기 위해서는 더 많은 입력 변수가 고려되어야 할 것으로 생각된다. 또한 실험 중 가중치를 주는 경우 구체적인 근거가 부족했고, 가중치의 기준이 되는 정보가 어떤 것으로 하느냐에 따라 분류 정확도가 달라지는 경우가 있을 수 있을 것이다.

웹 문서의 자동 분류 과정에서는 html tag 나 한글 조사 등 불필요한 부분을 제거하고 명사 단어를 추출해 해내는 전처리 과정이 필요한데 본 연구에서는 다루지 못하였다. 향후 연구에서는 연결빈도행렬(CFM) 알고리즘에 전처리 알고리즘이나 sw 를 추가한다면 보다 더 지능화된 문서 자동 분류 알고리즘으로 발전될 수 있을 것이다.

참고 문헌

김상범, 임해창, 윤덕호, 한광록, 이미영, "범주간 관계의 고려를 통한 자동 문서 범주화의 개선," HCI 2000 학술발표 논문집, 2000, pp.894-899.

신진섭, "단어들의 연관성을 이용한 문서의 자동분류," 한국정보처리학회, 정보처리논문지, 제 6 권, 제 9 호, 1999, pp.2422-2430.

신진섭, "웹 문서 분류를 위한 단어의 연관성 모델과 클러스터링 모델," 박사 학위 논문, 2000.

이재윤, "문헌 자동분류에서 용어 가중치 기법에 대한 연구," 한국정보관리학회 학술대회 논문집, 2000, pp.41-44.

정영미, "지식 자동분류를 위한 유사성 척도의 비교 평가," 데이타베이스진흥센터 제 2 회 디지털도서관 컨퍼런스 논문집, 1999, pp.87-97.

조광제, 김준태, "역 카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동 분류," 한국정보과학회 봄 학술발표 논문집, Vol. 24, No. 1, 1997, pp.507-510.

조태호, "신경망 또는 k-NN 에 의한 신문 기사 분류와 그의 성능 비교," 한국정보과학회 가을 학술발표논문집, 제 25 권, 제 2 호, 1998.

조태호, "텍스트 마이닝에 대한 소개와 기능," 한국정보처리학회 추계학술논문집, 1998, pp.27-29.

진훈, 김인철, "문서 분류를 위한 특징 선택," 학술발표논문집,

한국정보과학회, 제 28 권, 제 1 호, 2001, pp.262-264.

최정민, 진훈, 김인철, "웹 문서 분류법의 실험적 비교," http://dblab.kyungwon.ac.kr:5302/cd/cdl/thesis/J-internet%20applicable/2-jm%20choi.doc.

최종후, 한상태, 강현철, 김은석, 'AnswerTree 를 이용한 데이터마이닝 의사결정나무분석," SPSS 아카데미, 서울, 1998.

한광록, 선복근, 한상태, 임기욱, "인터넷 문서 자동 분류 시스템 개발에 관한 연구," 정보처리 논문지, 제 7 권, 제 9 호, 2000, pp. 2867-2875.

한국어 형태소 분석기-HAM (Hangul Analysis Module), http://nlp.kookmin.ac.kr.

한승희, 이재윤, "문헌 클러스터링을 위한 유사계수간의 연관성 측정," 한국정보관리학회 학술대회 논문집, 1999, pp.25-28.

한정기, 박민규, 김준태, "구문 패턴과 키워드 집합을 이용한 자동 문서 분류의 성능 향상," HCI 98 학술대회, 1998, pp.70-73.

허준희, 최준혁, 이정현, 김중배, 임기욱, "문서의 주제어별 가중치 부여와 단어 군집을 이용한 한국어 문서 자동 분류 시스템," 정보처리학회논문지, 제 5 호, 제 8-B 권, 2001, pp.447-454.

홍진혁, 류중원, 조성배, "실세계의 FAQ 메일 자동분류를 위한 문서 특징추출 방법의 성능 비교," 2001 봄 학술발표논문집, 한국정보과학회, 제 28 권, 제 1 호, 2001, pp.271-273.

Aggarwal, C. C. and Yu, P. S., "Data Mining Techniques for Associations, Clustering and Classification," Lecture Notes in Computer Science 1574, 1998, pp. 13-23.

Apte, C. and Damerau, F., "Automated Learning of Decision Rules for Text Categorization," ACM TOIS, vol 12, no 3, 1994, pp.233-251.

Breiman, L., Friedman, J. H., Olshen, R. A., and C. J. Stone, "Classification and regression trees," Wadsworth, 1984.

Chuang, W. T., Tiyyagura, A., Yang, H. H., and Giuffrida, G., "A Fast Algorithm for Hierarchical Text Classification," Data Warehousing and Knowledge Discovery, 2000, pp.409-418.

Goldszmidt, M. and Sahami, M., "A Probabilistic Approach to Full-Text Document Clustering," Tech. Report ITAD-433MS-98-044, SRI International, 1998.

Guthrie, L. and Walker, E., "Document classification by machine: Theory and practice," Proceedings of COLING-94, 1993.

Han, E. H., Karypid, G., and Kumar, V., "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification," www-users.cs.umm.edu/~karypis/publications/ Papers/pdf/wknn.pdf.

Joachims, T., "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proceedings of Machine Learning, 1998, pp.137-142.

Joachims, T., "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization," Proceeding of the 14th International Conference on Machine Learning ICML97, 1997, pp.143-151.

John, G. H. and Langley, P., "Estimating continuous distributions in Bayesian classifiers," Proc. 11th Conf on Uncertainty in Artificial Intelligence, Montreal Canada, 1995, pp.338-345.

Khan, I. and Blight, D., "Categorizing Web Documents Using Competitive Learning," ICNN, 1997, vol 1, 1997, pp.96-99.

Lewis, D., "Evaluating Text Categorization," Proceedings of the Speech and Natural Language Workshop, Asilomar, 1991, pp.312-318.

Lewis, D. and Ringuette, M., "A Comparison of Two Learming Algorithms for Text Classification," Annual Symposium on Document Analysis and Information Retrieval, 1994, pp.81-93.

Li, Y. H. and Jain, A. K., "Classification of Text Documents," The Computer Journal, Vol. 41, No. 8, 1998, pp.537-546.

Loh, W. and Shih, Y., "Split selection methods for classification trees," Statistica Scinica, Vol. 7, 1997, pp.815-840.

McCallum, A. and Nigam, K., "A Comparison of Event Models for Naive Bayes Text Classification," www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf

Quinlan, J. R., "C4.5 Programs for machine learning," Morgan Kaufmann, San Mateo, 1993.

Quinlan, J. R., "Induction of decision trees," Machine Learning, Vol. 1, No. 1, 1986. pp.81-106.

Salton, G., "Automatic Text Processing," AddisonWesley.INC, 1989, pp.275-280.

Sasaki, M. and Kita, K., "Rule-Based Text Categorization Using Hierarchical Categories," IEEE SMC, 98, vol. 3, 1998, pp.2827-2830.

Smith, M., "Neural Networks for Statistical Modeling," International Thomson Computer Press, 1996.

Wiener, E., Pedersen, J,O., and Weigend, A. S., "A Neural Network Approach to Topic Spotting," Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval, 1995, pp.317-332.

Witten, I. H. and Frank, E., "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," Academic Press, 2000.

Witten, I. H. and Frank, E., "Data Mining,"
Morgan Kaufmann Publishers, 2000.

첨부

<첨부 1> 시험에 사용된 50 개의 단어와 10 가지 분류

	단어(word)		변수명		설명
독립변수	감독	안하무인	w1	w26	
	견본	어깨	w2	w27	무작위로
	경마	여행	w3	w28	선정된 단어
	고기	연료	w4	w29	
	고독	오른쪽	w5	w30	가나다 순
	군대	오리	w6	w31	
	기준	유료	w7	w32	문장내에
	기초과학	유예기간	w8	w33	사용되었으면
	내장	유출	w9	w34	1,사용되지
	담배	일본	w10	w35	않았으면 0
	동문	전문가	w11	w36	
	떡	전세	w12	w37	
	로마	종교	w13	w38	
	맞벌이	주차	w14	w39	
	매실	준수	w15	w40	
	미래	지방자치	w16	w41	
	미모	천연기념물	w17	w42	
	배필	컴퓨터	w18	w43	
	보행	통상	w19	w44	
	보험	풍선	w20	w45	
	분노	함수	w21	w46	
	선발	항공기	w22	w47	
	선생	해안선	w23	w48	
	손등	화재	w24	w49	
	실존	흉작	w25	w50	

	분야 (category)	변수명	설명
종속변수	경제	c01	10 개 분야 가나다 순
	정치	c02	
	사회	c03	
	문화	c04	
	예술	c05	
	건강	c06	
	과학	c07	
	레져	c08	
	교육	c09	
	학문	c10	

<첨부 2> 키워드 매칭의 문서 분류 정확도

문서	CFM 1	CFM 2	CFM 3	CFM 4	CFM 5	CFM 6	CFM 7	CFM 8	CFM 9	CFM 10	category	Prediction	Match	정확도 (%)
c01	4	2	4	4	2	3	3	2	3	2	c01	O	3	0.333
c01	5	2	4	5	1	1	1	2	2	1	c01	O	2	0.5
c01	5	1	3	2	2	4	2	2	4	1	c01	O	1	1
c02	5	3	4	3	3	2	2	2	4	3	c02	X		
c02	4	5	5	5	5	3	2	1	4	4	c02	O	4	0.25
c03	4	4	5	4	4	1	4	3	5	4	c03	O	2	0.5
c05	4	5	5	5	5	5	4	2	4	5	c05	O	6	0.167
c06	2	2	2	4	2	5	3	1	3	1	c06	O	1	1
c06	5	3	5	3	4	4	3	2	5	3	c06	X		
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
c06	5	3	5	3	5	5	5	5	4	5	c06	O	7	0.142
c09	3	3	4	4	4	5	4	4	5	4	c09	O	2	0.5
c09	4	3	4	3	4	2	5	3	4	4	c09	X		
c09	4	4	5	3	4	3	4	3	5	5	c09	O	3	0.333
c09	4	4	5	4	5	3	4	3	5	5	c09	O	4	0.25
c09	6	5	6	4	5	4	5	4	5	6	c09	X		
c09	3	4	4	4	4	2	4	1	4	4	c09	O	7	0.142
c09	3	2	4	3	3	4	3	3	3	2	c09	X		
합														21.288

<첨부 3> 가중 키워드 매칭의 문서 분류 정확도

문서	CFM1	CFM2	CFM3	CFM4	CFM5	CFM6	CFM7	CFM8	CFM9	CFM10	Category Prediction Result			정확도 (%)
3	48	11	28	19	8	26	24	20	19	26	c01			1
4	45	37	32	30	22	26	43	9	30	45	c01			0.5
6	50	37	24	32	19	30	40	26	26	48	c01			1
10	12	46	11	31	38	6	8	9	12	27	c04	X		
1	9	36	19	32	53	12	20	11	16	45	c05			1
2	13	13	15	22	26	22	2	36	2	8	c05	X		
10	27	44	27	34	31	30	47	9	26	55	c05	X		
1	25	20	18	25	15	54	37	13	17	36	c06			1
5	22	14	18	17	13	43	4	26	6	11	c06			1
6	46	21	28	25	19	44	43	55	21	50	c06	X		
7	20	18	32	17	14	21	10	2	18	13	c06	X		
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
9	13	2	10	15	17	55	8	4	6	0	c06			1
10	23	24	30	12	14	49	2	6	14	13	c06			1
1	14	16	16	17	15	12	67	19	28	59	c10	X		
2	19	28	18	23	26	3	56	17	50	60	c10			
3	22	57	11	28	29	8	14	11	19	18	c10	X		
5	35	34	29	32	19	14	84	15	33	84	c10			0.5
9	11	13	26	36	38	26	30	12	25	37	c10	X		
10	30	36	24	28	26	26	85	15	38	81	c10	X		
합														9.5