經營論集, 第 43 卷 統合號 (2009年 12月)

벡터 공간 모델을 적용한 키워드 자동 추출에 대한 연구*

노 상 규**
조 원 진***

엄청나게 증가하는 정보의 홍수 속에서 사용자가 원하는 정보를 신속하게 얻기 위해서는 모든 자료를 자세히 읽고 검색하기보다는 가장 핵심이 되는 문서의 내용인 키워드를 통해 보다 빠른 시간에 문서의 내용을 효과적으로 이해할 수 있다. 그러나 아직까지 키워드가 할당되지 않은 많은 문서들이 존재하며, 키워드 할당을 수작업으로 처리하는 것은 많은 시간과 비용을 요하는 고된 작업이다. 이를 위해 문서를 대표할 수 있는 주요 키워드를 자동으로 추출하는 방법에 관한 연구들이 진행되어 왔다. 본 연구에서는 자동으로 키워드를 추출하기 위해 기존에 제안된 방법들을 소개하며, 입력된 질의에 대해서 가장 유사한 문서를 찾아주는 검색분야의 대표적인 알고리즘인 벡터 공간 모델(vector space model)을 역으로 적용하여 키워드를 자동적으로 추출하는 새로운 방법을 소개한다.

I. 서 론

컴퓨터와 인터넷의 발달로 날마다 증가하는 정보의 홍수 속에서 살아가고 있다. 그러나 오늘날 우리가 접하는 대부분의 정보는 체계적으로 분류되고 정리되어 있기보다는 정보 제공자들의 관심도와 편의에 의하여 여러 가지 형태로 제공되기 때문에 우리가 원하는 정보를 얻기까지 많은 시간을 소모하게 되는 경우도 있다. 이와 같이 여러 곳에 분산되어 있고 정돈되지 않은 정보원으로부터 필요한 정보를 신속하고 정확하게 검색하기 위해서는 모든 자료들을 자세히 읽고 검색하기 보다는 가장 핵심이 되는 문서의

*본 연구는 서울대학교 경영대학 경영연구소의 일부 지원에 의해 수행되었음.
**서울대학교 경영전문대학원 교수
***서울대학교 경영대학 박사과정

내용, 즉 키워드를 추출함으로써 빠른 시간 내에 문서 전체의 내용을 효과적으로 이해할 수 있는 능력이 요구된다. 보통 키워드는 잠재적 독자가 특정 문서가 관련되었는지의 여부를 판단하는 것이 용이하도록 문서에 대해 다소 상위수준의 설명을 제공하여 준다. 이와 같은 키워드들은 문서들을 매우 간결하게 요약해주기 때문에 이를 바탕으로 문서들 간의 관련성이나 유사도를 매우 적은 비용으로 판단할 때 사용되기도 한다.

이를 위해, 보통 학술적 연구 같은 경우 첫 페이지를 보면 초록 이외에 키워드 항목으로 연구내용을 대표할 수 있는 3~4개의 단어가 포함되어 있다. 이들 키워드는 논문을 검색하거나 브라우징 하는데 사용될 수 있으며, 논문 내용에서 중요한 단락을 찾아내거나 논문을 분류, 클러스터링 하는 요소로도 사용될 수 있다. 일반적으로 저널에서는 논문의 저자에게 키워드 항목의 제출을 요구하고는 있으나 필수항목이 아닌 경우가 많이 때문에 모든 논문에서 키워드 정보를 제공하고 있지는 않다. 키워드를 할당함에 있어 수작업으로 키워드 정보를 추출하여 키워드를 할당하는 작업은 내용에 기반하여 사람에 의해 심도 깊게 이루어지기 때문에 그 결과는 매우 효과적일 수 있으나, 그것을 판단하는 사람의 주관에 의해 사람마다 키워드 할당에 다소 차이가 있을 수 있다. 또한 키워드를 추출하고 할당하는 문제는 비단 학술연구와 같은 논문뿐만 아니라 신문기사, 도서 등 다양한 문헌들에 대해서도 존재한다. 이와 같은 기하급수적으로 늘어나는 수많은 문서들의 키워드를 수작업으로 할당하는 것은 굉장한 시간과 노력, 그리고 비용을 요하는 힘든 작업임이 분명하다. 그 결과, 자동으로 주요 키워드를 추출하는 방법에 관한 연구들이 진행되어 왔다(Turney, 2000; Frank et al., 1999; 조태호와 서정현, 2000; 이말례와 배환국, 2002).

자동 키워드 추출(automatic keyword extraction)이란 "문서 내용을 정확하게 표현하는 단어는 반드시 문서 중에 출현한다(内山惠三 and 中村正規, 1991; 이태헌과 박기홍, 2002)"라는 논리에 기초하여 문서가 포함하고 있는 여러 단어들 중에서 문서의 내용을 잘 표현할 수 있는 단어나 구로 표현되는 키워드를 자동적으로 추출하는 것이다. 자동적인 추출을 위해 특히 기계학습 및 통계적 기법을 이용한 방법이 주로 사용되어 왔다. 이와 같은 자동 키워드 추출은 문서검색, 웹페이지검색, 문서클러스터링, 문서요약, 텍스트마이닝 등을 위한 중요한 기반을 제공하기도 한다(Matsuo and Ishizuka, 2004).

많은 양의 문서들로부터 키워드를 추출하기 위해서는 추출방법에 대한 정확성과 효율성이 동시에 요구된다. 이를 위해 기계학습과 통계이론에 기초한 자동 키워드 추출을

위한 연구가 진행되어 왔으나, 대부분이 문서요약(Document summarization) 및 문서범주화(Document categorization)의 부문 연구로서 고려되어 왔으며, 키워드 자동 추출 자체의 시스템 개발을 위한 연구는 아직 미비하다. 본 연구에서는 기존에 키워드 자동 추출을 위해 제안된 방법론과 시스템을 소개하고, 입력된 질의에 대해서 가장 유사한 문서를 찾는데 적용되는 대표적인 검색 알고리즘인 벡터 공간 모델(vector space model)을 역으로 적용하여 타깃 문서에 대해서 관련된 키워드들을 자동 추출할 수 있는 새로운 접근법을 제시한다.

본 연구의 구성은 다음과 같다. 2장에서는 기존의 키워드 추출과 관련된 연구를 살펴보며, 이어서 3장에서는 벡터 공간 모델에 대한 소개와 함께 이것이 키워드 추출을 위해 어떻게 적용될 수 있는지에 대한 아이디어를 소개한다. 그리고 마지막 결론에서는 본 연구의 시사점과 한계를 제시하고 마무리한다.

II. 이론연구

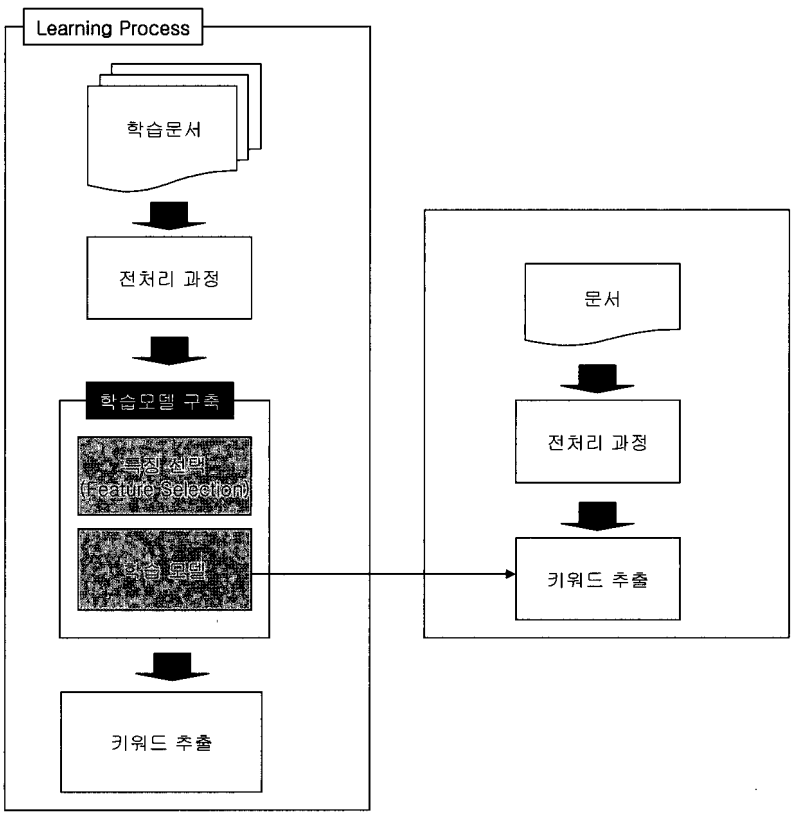
키워드를 추출하기 위한 연구는 우선 키워드 추출 기법(keyword extraction approach)과 키워드 할당 기법(keyword assignment approach)으로 구분할 수 있다. 키워드 추출 기법의 경우 문서 집합 내에서 키워드를 추출하는 것을 기본으로 한다. 반면 키워드 할당 기법은 정제된 어휘사전을 이용하여 문서에 적합할 것으로 판단되는 용어를 키워드로 할당하는 기법이다. 키워드 할당 기법의 경우 정제된 사전을 계속적으로 관리하는 것이 현실적으로 어렵고, 일반적으로 키워드는 복합어의 형태로 표현되는데다, 복합어는 신조어일 경우가 많기 때문에 빠른 업데이트가 쉽지 않아 키워드 추출 기법에 대한 적용이 보다 많이 이루어지고 있다(Frank et al., 1999).

앞에서 언급되었듯이, 이와 같은 키워드 추출을 위한 연구는 다른 분야에 비해 크게 활성화 되지 못한 것이 사실이다. 아직까지는 문서요약과 문서범주화를 위한 부수적인 단계로 키워드 추출과정에 대한 연구가 실험적으로 이루어지고 있으며 키워드 자동 추출 자체의 본질적인 성능향상을 위한 연구는 일부 연구에 지나지 않는다. 본 절에서는 키워드 추출을 위한 방법론과 시스템을 개발한 선행연구에 대해서 살펴본다. 기존 키워드 추출을 위한 접근법은 기계학습에 기초한 방법과 통계적 기법에 기초한 방법으로 구

분할 수 있다.[1] 각각의 접근법에 대한 간단한 소개와 함께 대표적인 연구가 제시된다.

1. 기계학습 기반 접근법(Machine-learning based approach)

기계학습 기법을 적용하여 키워드를 추출하는 문제는 일종의 분류(classification) 문제와 동일하게 간주된다. 즉, 특정 문서에 포함되어 있는 용어가 "키워드"인지, "No키워드"인지로 이진분류문제와 동일하게 간주하여 문제를 해결하고자 한다. 문서에 있는



〈그림 1〉 기계학습을 이용한 키워드 추출 시스템의 전체 구성도

1) 기계학습 역시 일종의 통계이론에 기초하고 있기는 하나, 본 연구에서는 감독학습(supervised learning)을 통해 학습 모델을 구축하여 키워드를 추출하는 경우 기계학습 기법으로, 그 외의 통계적 이론에 기초하여 키워드를 추출하는 경우 통계 기반 접근법으로 구분했다.

각 용어들(기계학습에서는 examples)이 키워드인지 아닌지에 대해서 올바르게 분류하고자 하는 것이 목표인 기계학습 기법 중심의 키워드추출 알고리즘들의 전체적인 프로세스는 다른 기계학습 알고리즘을 학습시키는 과정과 동일하며, 〈그림 1〉에 도식화 되어 있다. 각각의 단계에 대해서 살펴보면 다음과 같다.

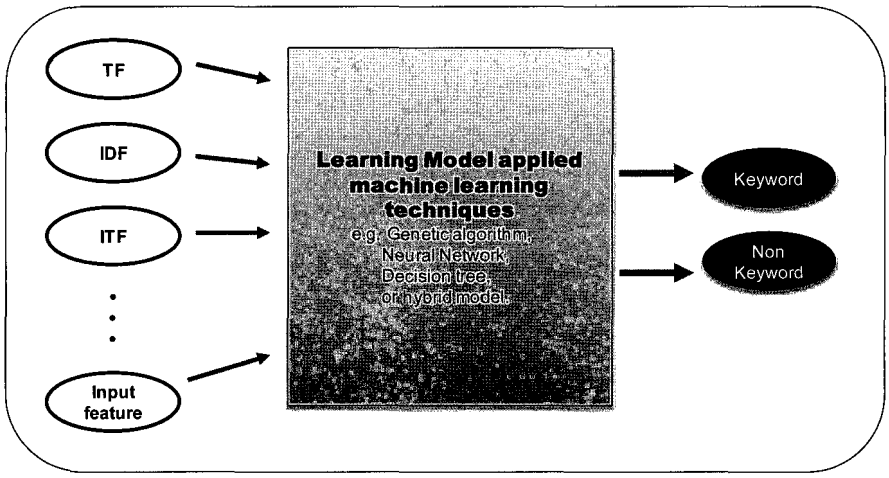
1) data 전처리

다양한 학습기법을 적용하기 전에 훈련데이터들이 우선 전처리되어야 한다. 즉, 학습 모델에 인풋으로서 사용하기 적합한 형태로 변형된다. 먼저 일반 사용자가 사용하는 문서는 파싱처리되어 의미있는 용어들의 집합으로 만들어진다. 파싱될 때 보통 구두점(punctuation mark), 대시(dashes), 괄호(brackets), 숫자(numbers) 같은 것에 따라 단어들이 구분되어 인식되며, 기호나 숫자, 조사 같은 부속어[2]들을 제거한 단어들의 리스트들이 최종 산출된다. 산출된 단어 리스트를 가지고 그 단어의 다양한 특징(feature) 벡터로서 표현된다. 대표적인 특징 벡터로 용어빈도(*tf*), 역문서빈도(*idf*), 역용어빈도(*itf*) 등이 주로 사용된다(조태호와 서정현, 2000).

2) 키워드 추출을 위한 학습 모델 구축

전처리된 데이터를 사용하여 다음의 〈그림 2〉와 같이 학습 모델을 만든다. 학습 모델(learning model)은 다양한 기계학습 기법을 적용하여 만들어지며, 특히 분류 문제(classification problem)에 좋은 성능을 보이는 의사결정나무(decision tree), 인공신경망(artificial neural network)과 같은 기법이 많이 적용된다. 학습 모델을 구축함에 있어서 사람에 의해 미리 할당된 키워드들을 갖고 있는 훈련데이터(training data)와 평가데이터(test data)를 이용하여 과적합(over fitting)되지 않도록 적절한 특징 벡터들이 우선 선택(feature selection)되며, 선택된 특징 벡터를 기준으로 최적의 성능을 갖도록 학습 모델이 구축된다.

2) 한글의 경우, 조사, 어미, 접속사, 대명사 등이 이에 해당되며, 영어의 경우 a, an, the와 같은 관사류, that, this와 같은 지시대명사, what 등의 의문사, is, are와 같은 be동사, 조동사 등이 이에 해당된다.



〈그림 2〉 키워드 추출을 위한 학습 모델

3) 키워드 추출

구축된 학습모형을 새로운 검증데이터를 이용하여 평가해봄으로써 실제 활용도에 대한 평가가 객관적으로 이루어진다. 그 결과 좋은 성능을 보이는 시스템은 실제 키워드 추출 프로세스에 적용되어 활용된다.

이상과 같이, 기계학습 기법을 키워드 추출하는 데 적용하는 과정에 대해 간단히 살펴보았다. 다음은 기계학습 분야에서 자동 키워드 추출과 관련된 연구가 많이 이루어지진 않았지만 기계학습 기법을 키워드 추출에 적용하는 것과 관련된 개념적 아이디어를 제공하여주는 대표적인 연구들을 간단히 소개한다.

Witten et al.(1999)은 Kea(an algorithm for automatically extraction keyphrases from text)라는 키워드 추출 시스템을 개발했다. 뉴질랜드 Waikato 대학에서 개발된 이 시스템은 키워드를 문서에서 추출해내는 시스템 중 우수한 성능을 보이는 것으로 보고되고 있다(이병희, 2005). Kea는 우선 다른 키워드 추출 기법과 유사한 방법으로 후보 키워드들을 추출한다. 후보 키워드로 추출되기 위해서는 용어들 사이에 불용어가 있어서는 안 된다는 조건만 만족하면 된다. 다음 단계로 Kea는 이미 사람들에 의해 키워드가 할당된 문서집합을 이용하여 학습을 통해 모델을 구축한다. 이때 학습 알고리즘으로 나이브 베이지안 기계학습 알고리즘을 사용하였으며, 모델 구축을 위해 사용한 특징들로

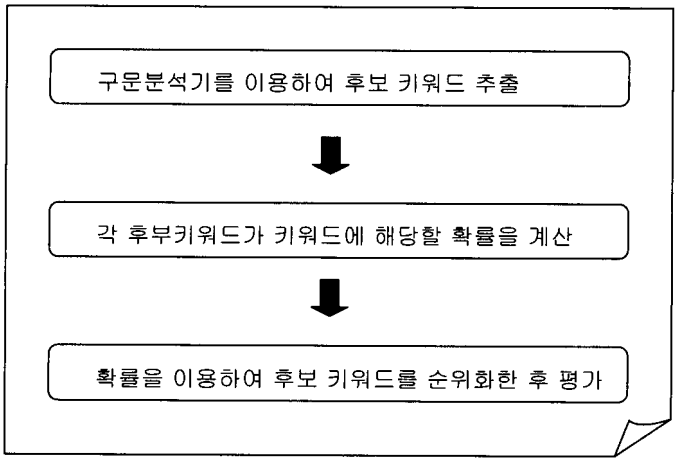
tfidf 가중치와 Distance를 사용하였다. 여기서 Distance는 해당 키워드가 문서의 어느 부분에서 처음으로 나타났는가를 의미하는 것으로 보다 일찍 출현한 용어일수록 그 문서를 표현함에 있어 영향력이 클 것으로 기대한다. 이 시스템은 현재 뉴질랜드 디지털 도서관(http://www.nzdl.org)에 구현되어 있으며 다른 키워드 추출 시스템에 비하여 상당히 효율적인 시스템으로 평가 받고 있다.

Turney(2000)는 C4.5 의사결정나무(decision tree)를 5개의 다른 문서군을 형성하는 총 652개의 문서들에 적용하였다. 실험결과 75%의 성공률을 보였다. C4.5를 이용하여 키워드를 추출하는 실험을 하는 과정에서 C4.5와 같은 일반적 목적의 학습 알고리즘보다 더 좋은 precision을 달성할 수는 방법에 대해 고민하는 과정에서 유전적 알고리즘(genetic algorithm)을 함께 적용한 하이브리드 모델을 개발했다. 이것은 GenEx라는 키워드 추출을 위한 시스템으로, 유전적 알고리즘을 사용하여 미세조정된 일련의 휴리스틱 규칙에 기반하여 주요구문을 추출한다. 유전적 알고리즘은 규칙들의 파라미터들을 조정하여 훈련문서들에서 올바르게 식별된 키워드들의 수를 최적화한다. Turney는 단순히 C4.5만을 적용하여 구축한 학습모델보다 GenEx가 보다 우수한 키워드 추천 성능을 가진다고 실험을 통해 결론지었다.

조태호와 서정현(2000)은 문서의 내용을 반영하는 중요한 단어, 즉 키워드를 추천하는 시스템을 위해 신경망(Neural network)을 이용하여 학습모델을 구축했다. 기존에는 단순히 단어의 빈도와 역문서 빈도를 특징으로 사용하여 키워드 추천을 위한 모델을 만들었다. 그러나 조태호와 서정현의 연구에서는 실제로 문서내 빈도와 역문서빈도 뿐만 아니라 제목에 포함여부, 단어의 위치 등도 중요한 고려요인으로 판단하여 새롭게 특징으로 추가하여 이것으로 특징 벡터를 형성하고 이를 학습하여 키워드를 선택하는 신경망 모델을 이용한 접근법을 제안했다. 실험결과 단순히 단어빈도와 역문서빈도를 이용하여 키워드를 추출하는 방법보다 보다 좋은 성능을 보였다.

2. 통계 기반 접근법(Statistical based approach)

지도학습 이외의 통계이론에 기반하여 키워드를 추출하는 경우는 대부분 *tfidf*에 기반한 가중치를 부여하여 키워드를 추출하는 연구가 주종을 이루고 있다. 물론 *tfidf*의 가중치 부여 방법은 기계학습을 위한 특징 벡터의 일부를 형성하기도 하지만, 순수하게



〈그림 3〉 통계적 방법을 이용한 키워드 추출의 일반적 단계

용어들의 *tfidf*값에 기초하여 높은 순위를 갖는 용어들을 추출하는 경우는 앞서 설명한 기계학습에 기반한 키워드 추출 방법들과는 차이가 있다. 통계에 기반한 방법으로 *tfidf*, 그리고 이창범 등(2002)에 의해 새롭게 제안된 주성분 분석(Principal components analysis)을 적용하여 키워드를 추출하는 방법에 대해서 소개하고, 이어서 이외의 통계 이론에 기반하여 키워드를 추출하는 연구들을 제시한다.

통계적 방법을 사용한 키워드 추출 기법은 일반적으로 〈그림 3〉과 같은 단계를 거친다.

- *tfidf*(Term Frequency Inverse Document Frequency)

통계적 기법에 의한 자동 키워드 추출을 위한 기준으로 주어진 문서에서 특정 단어가 얼마나 자주 사용되었는가 하는 빈도수 정보가 사용된다(이말례와 배환국, 2002). 단어의 사용빈도 수는 산출 방식에 따라 단순 빈도수와 상대빈도 수로 구분된다. 단순 빈도수는 문서의 크기나 분석 대상이 되는 텍스트의 길이 또는 단어의 출현빈도수를 전혀 고려하지 않은 것으로 이것을 기준으로 키워드로 판단하기에는 부적절한 면이 있다. 상대빈도수는 이와 같은 문제점을 고려한 것으로써 키워드를 선택하는 기준으로 사용되기에 적합한 형태이다. 상대 빈도수는 단어 빈도수를 문서의 크기 등으로 나누어 줌으로써 빈도수를 정규화한 것이다. 상대빈도수를 활용한 대표적인 것이 *tfidf*이다. *tfidf*

(Term Frequency Inverse Document Frequency) 방식이란 문서군들이 있다고 할 때, 어떤 낱말이 어떤 한 문서에서 얼마나 중요한 것인지를 나타내는 통계적 수치로 다음의 수식으로 표현할 수 있다(Salton and McGrill, 1983).

$$tfidf(w_i, d_j) = tf(w_i, d_j) \times \log\left(\frac{n}{df(w_i)}\right)$$

$tf(w_i, d_j)$ = 문서d_j에 단어 w_i가 나타나는 횟수
$df(w_i)$ = 단어 w_i가 들어가는 문서의 총 수
n: 코퍼스의 전체 문서의 개수

tf(term frequency)는 해당 낱말이 문서에 얼마나 많이 나오는지를 나타낸다. 해당 낱말이 문서에 많이 나올수록 그 문서에서 중요하다고 할 수 있다. 다음의 〈표 1〉은 특정 문헌 내 용어의 출현빈도에 따라 가중치를 부여하는 *tf*의 이형으로서 이를 정규화하거나 일부 변형시켜 다양한 가중치 부여 기법들이 종합되어 있다(김판준, 2008).

그러나 문서군에 있는 다른 문서에서도 이 낱말이 많이 나타날수록 이것은 그 낱말이 흔하게 나오는 낱말이라는 것을 의미한다. 이것을 *df*(document frequency)라고 하는데, 흔하게 나오는 낱말들을 걸러주기 위하여 역수를 취하여 *idf*(inverse document frequency)라고 한다. 그 결과, 문서군을 구성하는 문서에 따라서 *idf* 값이 달라진다. 예를 들어 "온톨로지"라는 낱말은 세상에 있는 모든 문서들 사이에서는 잘 나오지 않아서

〈표 1〉 *tf*와 이것의 다양한 변형들

약어	공식	설명/출처
bin	1 or 0	*tf*=1, if *tf*>0
atf	0.5 + (o.5 *(*tf*/max *tf*))	Augmented *tf*
itf	1 –(1/(1+*tf*))	Inverse Term Frequency
Itf	log(1+*tf*)	log *tf*

이 낱말이 어떤 문서의 핵심어로 채택될 수 있는 가능성이 높지만, 온톨로지에 대한 문서들로 구성된 문서군의 경우 온톨로지란 이 단어는 상투어가 되어 각 문서들을 세분화하여 구분할 수 있는 다른 낱말들이 높은 가중치를 얻게 된다.

즉, *tfidf*는 어떤 단어에 대한 중요도는 그 단어가 문서에 나온 횟수에 비례하고, 그 단어가 있는 모든 문서의 총수에는 반비례한다는 것이 기본 아이디어이다. 이 방식을 이용하면 하나의 문서 중에서 *tfidf* 값이 가장 높은 순으로 키워드로 채택된다.

이말례와 배환국(2002)은 Anchor Text의 단어들이 키워드로 적합한지를 *tfidf*를 이용하여 테스트하였다. Anchor Text는 어떤 웹문서의 내용을 사람이 직접 요약해 놓은 것으로 그 문서의 키워드를 추출하는데 중요한 정보가 될 수 있다. 그들은 Anchor Text가 실제 얼마나 유용한 정보를 포함하고 있는지를 수치적으로 증명해 보였으며, 그 결과는 가중치가 높아서 키워드로 적합하지 않는 단어들이 있는가하면, Anchor Text는 웹문서 자체에 있는 문자열들이 아니므로 아예 문서에 나오지도 않는 단어가 있어 키워드로 적합하지 않은 단어도 추출되기도 했다. 이를 해결하는 새로운 지능형정보추출 에이전트 시스템을 제시하였다. 이 시스템은 Anchor Text에 *tfidf*를 적용하여 키워드를 추출하는 부분과 Anchor Text에 있는 각각의 단어들이 문서의 키워드로서 얼마나 적합한가를 평가하는 부분으로 나누어진다. 여기서 비교 분석을 하는 이유는 Anchor Text에서 추출한 단어들의 가중치가 그 문서에 있는 단어들의 가중치와 비교해 볼 때 상대적으로 작은 가중치의 것들이라면 키워드로 부적합할 수도 있기 때문이다. 즉, 비교분석을 한 후 그 결과를 토대로 좀 더 나은 키워드 추출 방법을 제안했다.

- 주성분 분석(Principal Component Analysis)

주성분 분석은 여러 개(p>=2)의 반응 변수에 대하여 얻어진 다변량 자료를 분석대상으로 다차원적인 변수들을 축소, 요약하는 차원의 단순화와 더불어 일반적으로 서로 상관되어 있는 반응 변수들 간의 복잡한 구조를 분석하는데 그 목적을 두고 있다. 이를 위하여 주성분 분석은 반응변수들을 선형 변환시켜 주성분이라고 부르는 서로 상관되어 있지 않은, 혹은 독립적인 새로운 인공변수들을 유도한다. 이때 각 주성분이 보유하는 변이의 크기를 기준으로 그 중요도 순서를 생각할 수 있는데, 그들 중 첫 소수 몇 개의 주성분이 원래 자료에 내재하는 전체 변이 중 가능한 많은 부분을 보유하도록 변환시킴으로써 정보의 손실을 최소화하는 차원의 축약을 기대할 수 있다. 결국, 주성분 분

석을 이용한다면 문서의 내용을 나타내기 위해 문서에 출현하는 모든 단어를 사용하는 대신에, 정보의 손실을 최소화하면서 소수의 몇 개의 단어로 문서의 내용을 표현할 수 있다. 즉 그 문서의 키워드를 추출할 수 있다.

이창범 등(2002)은 문서의 내용을 대표할 수 있는 키워드를 추출하는데 있어 다변량 통계 분석 기법 중의 하나인 주성분 분석을 이용하는 모델을 제안했다. 고유값과 고유벡터를 이용하여 문서자체내의 단어의 흐름을 정량화하고, 그 정보를 이용하여 문서 자체 내에서의 발생 빈도와 공기정보(co-occurrence)를 이용하여 주제어를 추출하였다. 실험결과, 단순히 단어의 출현빈도만을 이용한 방법에 비해 제안한 모델이 더 좋은 성능을 보임을 증명했다.

- 그외 통계적 접근법을 사용한 연구들

Matsuo and Ishizuka(2004) 는 여러 문서로 구성된 코퍼스가 아닌 하나의 문서에 적용할 수 있는 새로운 키워드 추출 알고리즘을 제시했다. 우선 빈발 용어가 추출되고, 그 다음은 각 용어와 그 빈발 용어가 같은 문장에 존재하는 co-occurrence가 계산된다. Co-occurrence 분포는 다음과 같이 문서에서 한 용어의 중요성을 보여준다. 만약 용어 a와 그 빈달 용어들 사이의 co-occurrence의 확률 분포가 빈발 용어들의 특정 서브셋에 바이어스되면, 용어 a는 키워드가 될 수 있는 것으로 판단한다. 분포의 바이어스 정도는 카이제곱에 의해 평가된다. 그들이 제시한 알고리즘은 코퍼스를 사용하지 않고도 *tfidf*를 적용한 결과와 유사한 성능을 보였다.

신형주 등(2000)은 텍스트 문서의 키워드를 추출하고 문서를 주제별로 분류하기 위해 확률적 그래프 모델을 사용하는 방법을 제안했다. 텍스트 문서 데이터를 문서와 단어의 쌍으로 표현하여 확률적 생성 모델을 학습하였다. 확률적 그래프 모델의 학습에는 정의된 likelihood를 최대화하기 위한 Expected Maximization 알고리즘을 사용하였다. 시스템이 추출한 키워드와 사람이 제시한 키워드가 유사한지를 실험적으로 평가하였으며, 학습한 결과가 사람이 문서에 대해 파악한 키워드와 유사한 결과를 보였다.

III. 벡터 공간 모델(Vector Space Model)을 이용한 키워드 추출

정보검색 이론(Information Retrieval Theory)은 다양한 형태의 비구조화된 정보를 담고 있는 문서와 필요한 정보에 대한 요구를 표현하는 질의를 각각 내용에 따라 특정한 형태로 표현하고 이를 이용하여 사용자가 필요로 하는 정보를 효과적으로 검색, 분류하는데 필요한 방법론을 연구하는 분야이다(허원창, 1999).

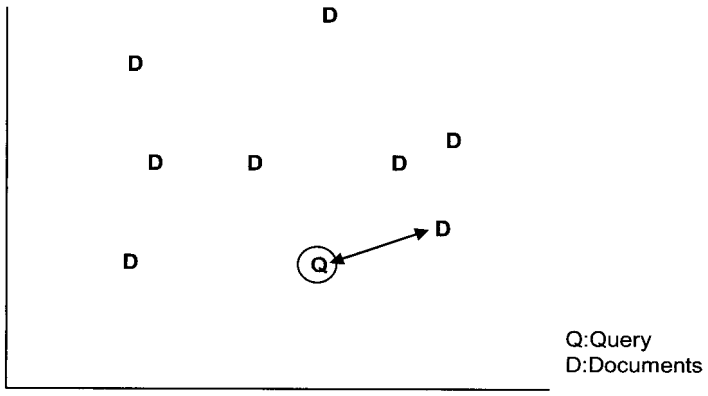
정보를 검색하기위해 주어진 질의와 문서 사이의 유사성을 평가하는 방법으로는 벡터공간모델, 확률검색, 불리안모델 등이 있으며, 다음 절에서는 본 연구에서 키워드 추출을 위해 수정하여 적용하고자 하는 벡터 공간 모델에 대해서 보다 자세히 살펴보고, 이어서 벡터 공간 모델을 키워드 추출을 위해 역으로 적용하는 방법에 대한 아이디어를 간단히 제시한다.

1. 벡터 공간 모델(Vector Space Model)

최근에는 많은 정보를 손쉽게 접할 수 있는 검색 서비스를 제공하는 사이트들이 많아지고 있으며, 원하는 정보를 보다 빠르고 쉽게 찾기 위한 사용자들의 이용 또한 나날이 증가하고 있다. 검색 서비스를 제공하는 사이트에서는 통상적으로 인터넷 상에 존재하는 원시데이터를 수집하여 해당 자료별로 키워드를 선정한 후 데이터베이스로 구축하고, 사용자들이 찾고자 하는 자료의 일부 키워드를 입력하여 검색을 요청하면 해당 키워드로 지정되어 있는 자료들을 사용자들에게 제공하는 형태로 서비스를 수행하고 있다.

이때 검색 서비스를 제공하는 사이트에서는 사용자들의 검색에 따라 제공되는 자료를 문서의 정확도, 중요도 등에 따라 상위의 문서들을 상위에 배치하여 사용자들에게 제공한다. 이와 같은 문서의 중요도를 분석하는 많은 방법 중에서 벡터 공간 모델(Vector Space Model)은 정보 필터링, 문서 내에서의 정보검색, 색인과 유사도를 계산하기 위한 수학모델로서, 다차원 선형공간에서의 벡터 정보를 이용하여 자연어를 포함한 문서의 중요도를 분석하기 위한 방법을 제시하고 있다.

"Term vector model"이라고도 불리는 벡터 공간 모델은 코넬 대학교의 Gerald Salton

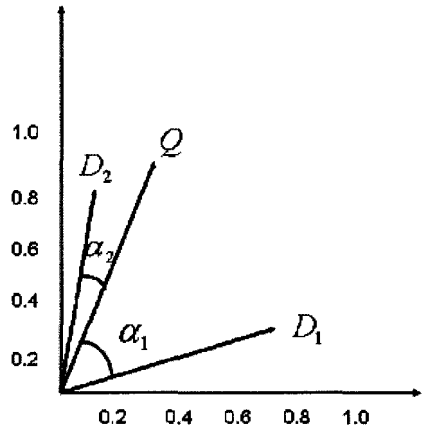


〈그림 4〉 벡터 공간 모델의 유사도 계산

이 SMART(System for the Mechanical Analysis and Retrieval of Text) 시스템(Salton and McGrill, 1983)을 벡터 공간 모델을 적용하여 개발한 이래, 벡터 공간 모델은 단순하고 빠르기 때문에 현재 내용기반 웹 정보 검색에서 가장 전형적인 검색 모델이라고 할 수 있다. 이것은 각 문서는 그 문서가 포함하고 있는 색인단어의 벡터로 나타낼 수 있고, 〈그림 4〉처럼 문서의 유사도는 벡터에 위치한 단어들 간의 거리로 계산해낼 수 있다라는 것이 이 모델의 대전제다.

구체적으로 살펴보면, 벡터 공간 모델에서는 단어와 문서는 k차원 공간의 벡터로 인코딩된다(Berry, 1999). K는 독립적인 단어와 개념 또는 텍스트와 관련된 클래스의 수에 따라 결정된다. 그러므로 각각의 벡터 성분은 대응되는 단어, 개념 및 클래스의 중요성을 반영한다. 벡터공간 모델상에서는 각 문서들과 사용자 질의는 n차원 공간 속의 벡터들로 취급되며, 이때 각 차원들은 색인용어들로 표현된다. 이 기법에 의한 검색방법은 다음과 같다. 각 문서들은 사용자 질문에 대해서 그 유사성의 순위별로 출력되며, 이러한 과정은 코사인 상관도에 의해 계산되다. 예를 들어, 다음의 〈그림 5〉에서 보면, α_2가 α_1보다 크므로 D_2가 D_1보다 사용자 질의(Q)에 더 가까이 위치해 있으므로 D_2가 사용자 질의에 대해 보다 유사한 문서라고 할 수 있다. 이로써 벡터 공간 내에서 사용자 질의에 가장 근접해 있는 문서들을 직관적으로 검색해 낸다.

벡터 공간 모델을 사용하기 위해서는 문서의 벡터 공간에 있는 용어의 가중치(weights)를 계산하고 있어야 한다. 이를 위해서 앞의 2장에서 소개된 *tfidf*(term frequency inverse



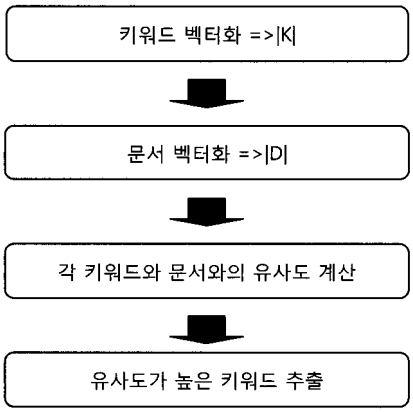
〈그림 5〉 질의와 문서들 사이의 유사도 비교

document frequency)가 주로 사용되고 있다.

벡터 공간 모델의 주요 장점은 용어의 가중치 기법이 검색 성능을 향상시키고, 질의 조건에 근접한 문서 검색이 가능하며, 거리 측정 기법 즉 코사인 순위화 등의 기법이 문서를 질의를 기준으로 정렬해 줄 수 있다는 점이다. 반면 문서가 포함하고 있는 주요 색인어가 상호 독립적이어야 한다는 가정이 단점으로 지적된다. 이를 극복하기 위해 온톨로지를 적용하여 주요 색인어를 분류하고 거리를 고려하는 시도도 있다.

2. 벡터 공간 모델을 이용한 키워드 추출방법

앞서 설명한 벡터 공간 모델은 사용자가 질의를 입력한 경우, 그 질의어와 가장 유사한 문서를 찾아주기 위해, 질의어와 각 문서들의 거리를 측정한다. 본 연구는 벡터 공간 모델을 적용하여 특정 문서에 적합한 키워드를 자동 추출하는 새로운 아이디어를 제안한다. 즉, 기존의 검색분야에서 벡터 공간 모델은 많은 문서들 중에 입력한 키워드와 가장 근접한 키워드를 찾는 데 적용되어 왔으나, 본 연구에서는 여러 키워드 중에서 특정 문서와 가장 유사한 키워드를 찾아내기 위해서 벡터 공간 모델을 적용하는 새로운 아이디어를 제시한다. 그 결과 문서에 입력된 단어들과 키워드들을 비교하여 각 문서에 포함된 단어에 기반하여 적합한 키워드를 추천할 수 있다.



〈그림 6〉 키워드 추출을 위한 벡터 공간 모델 프로세스

벡터 공간 모델을 적용하여 키워드를 추출하는 과정은 다음과 같다. 우선, 키워드 각각은 자신의 벡터 길이가 계산되며, 키워드 추천을 하고자 하는 타겟 문서는 그 문서가 포함하고 있는 단어들을 기반으로 벡터화된다. 문서를 벡터화 할 때 적용하는 용어가중치는 여러 가지가 있을 수 있다. 예를 들어, 초기에는 정보검색 분야의 대표적인 가중치 기법으로 문헌과 문헌집합 내 출현 정보를 조합한 *tfidf* 가중치 기법을 주로 사용해 왔으나, 최근에 와서 범주 정보의 사용을 적극적으로 검토하는 등 다양한 시도가 진행되고 있다. 즉, 가중치 산출의 기반이 되는 단서로 *tfidf* 외에 가중치 산출을 위한 새로운 정보원으로서 범주 요소를 사용하는 연구가 활발히 이루어지고 있다(김판준, 2008).

키워드와 타켓 문서가 벡터화되면, 이제 둘 사이의 유사도를 계산한다. 유사도를 계산할 수 있는 다양한 방법이 있으나 일반적인 벡터 공간 모델에서 사용하는 코사인 유사도 공식을 이용해 둘 사이의 유사도를 계산한다. 유사도 계산을 통해 유사도가 높은 상위 3~5개를 키워드로서 도출함으로써 역 벡터 공간 모델의 키워드 추출 프로세스는 종료된다. 역 벡터 공간 모델을 이용한 키워드 추출 과정은 〈그림 6〉처럼 정리되어 도식화될 수 있다.

IV. 결론

최근 디지털 도서관이 등장하고 인터넷이 폭 넓게 보급되어 온라인 상에서 얻을 수 있는 텍스트 정보의 양이 급증하면서 문서 관리의 효율과 검색의 성능 향상을 꾀하기 위해 이런 문서들에 키워드 할당의 필요성에 대해 의견을 같이하고 있다. 사람은 문서를 읽고 그 내용을 머릿속에서 개념적으로 정리하여 몇 개의 용어를 이용하여 키워들 인지한다. 그러나 이와 같은 과정을 사람에 의해 수작업을 진행하는 것은 시간적, 비용적으로 비효율적일 수 있다. 그 결과 문서를 대표할 수 있는 주요 키워드를 자동으로 추출하는 방법에 관해 연구들이 진행되어 왔다. 특히, 인간이 학습하는 과정을 모방하여 기계적 시스템으로 하여금 학습을 통해 추론, 연산, 판단하도록 하는 기계학습 분야에서 키워드 자동 추출에 관심을 갖고 몇몇 연구가 실험적으로 진행되기도 했다.

그러나 아직까지는 키워드 자동 추출 자체의 성능을 향상시키기 위해 키워드 추출 시스템에 초점을 맞춰 연구를 진행한 것은 일부 연구에 지나지 않는다. 대부분이 문서 범주화나 문서요약을 위한 부문연구로서 키워드 자동 추출에 대한 연구를 접근했다. 그러나 문서범주화나 문서요약을 위해 키워드를 추출하는 것은 궁극적인 문서범주화나 문서요약을 위한 시스템의 성능 향상을 위해 디자인되어, 실제 문서내용에 근거하여 논문이나 도서의 키워드를 추출하는데 적용할만한 키워드 자동 추출기의 성능에는 부합되지 못하는 경우가 많다. 즉, 다른 시스템을 위한 서브모듈로서 키워드 추출기가 아닌 본연의 목적 자체가 키워드 자동 추출을 위한 방법론 및 시스템에 대한 연구에 보다 관심을 가질 필요가 있다.

본 연구에서는 기존에 키워드 자동 추출을 위해 제안된 방법론과 시스템을 소개하여 키워드 자동 추출과 관련된 연구의 흐름을 파악하여 키워드 자동 추출과 관련된 개념적 이해를 높일 수 있도록 했다. 또한 입력된 질의에 대해서 가장 유사한 문서를 찾는 데 적용되는 대표적인 검색 알고리즘인 벡터 공간 모델(Vector space model)을 적용하여 타겟 문서에 대해서 관련된 키워드들을 자동 추출할 수 있는 새로운 접근법을 제시한다.

본 연구는 벡터 공간 모델을 사용하여 키워드 추출에 적용할 수 있는 방법을 제시했으며, 향후에는 본 연구에서 제시한 모델을 실험적으로 검증하는 과정이 뒤따를 것이

며, 벡터 공간 모델을 적용함에 있어 여러 가지 용어 가중치를 변형 적용하여 키워드 추출에 적합한 용어 가중치 설정에 대한 연구도 진행될 것이다.

참고문헌

김판준, "용어 가중치 기법을 이용한 로치오 분류기의 성능 향상에 관한 연구," *정보처리학회지*, 제25권, 제1호, 2008.

신형주, 장병탁, 김영택, "텍스트 문서의 주제어 추출을 위한 확률적 그래프 모델의 학습," *한국정보과학회 봄 학술발표 논문집*, Vol. 27, No. 1, 2000.

이병희, "복합어 및 주제어 추출을 이용한 개선된 정보 검색 시스템," 아주대학교 정보통신전문대학원 석사학위논문, 2005.

이말례, 배환국, "TFIDF를 이용한 키워드 추출 시스템 설계," *인지과학*, 제13권, 제1호, 2002.

이창범, 김민수, 이기호, 이귀상, 박혁로, "주성분 분석을 이용한 문서 주제어 추출," *소프트웨어 및 응용*, 제29권, 제 9-10호, 2002.

조태호, 서정현, "문서의 키워드 추출에 대한 신경망 접근," *한국정보과학회 가을 학술발표 논문집*, Vol. 27, No. 2, 2000.

허원창, "신경망회로망을 사용한 WWW 문서의 자동분류, 서울대학교 공과대학 석사학위논문, 1999.

Berry, M., Z. Dramc, and E. Jessup, "Matrics, Vector Spaces, and Information Retrieval," *SIAM Review*, Vol. 41, 335-362, 1999.

Frank, E, W. Paynter, I. Witten, C. Gutwin, and C. Nevill-Manning, "Domain-specific keyphrase extraction," *In proceedings of IJCAI'99*, 1999.

Ikonomakis, M. S. Kotsiantis, and V. Tampakas, "Text classification using machine learning Techniques," WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, 966-974, 2005.

Matsuo, Y. and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," International Journal on Artificial Intelligence

Tools, Vol. 13, No. 1, pp. 157-169, 2004.

Salton, G., and M. McGrill, *Introduction to modern information retrieval*, Mcgraw-Hill, New York, 1983.

Turney, P., "Learning algorithm for keyphrase extraction," *Information Retrieval*, 303-306, 2000.

Witten, I., G Paynter, E. Frank, C. Gutwin, and C. Nevill-Manning, "KEA: Practical Automatic Keyphrase Extraction," in *Proceedings of DL'99,* 1999.

內山惠三 and 中村正規, "중요 키워드 추출방식과 그 활용방법," *情處學DBS硏報*, 1991.

Automatic Keyword Extraction Using Vector Space Model

Sangkyu Rho*
Wonchin Cho**

Keywords are an important means of documents summarization, clustering, and topic search. Only a small minority of documents has author-assigned keywords, and assigning them manually is a tedious process that requires knowledge of the subject matter. Therefore, it is highly desirable to automate the keyword extraction process. In this research, we review existing researches related automatic keyword extraction. We grouped existing researches into two approaches, those are machine learning based approach and statistical based approach. Moreover, we propose a new technique for automatic keywords extraction. This is based on Vector Space Model that is a classical information retrieval technique. Automatic keyword extraction technique using Vector Space model, we propose, is potentially of great benefit.

Keywords: automatic keyword extraction, vector space model

*Professor of MIS, Graduate school of business, Seoul National University
**Ph.D. student in MIS, College of Business Administration, Seoul National University