용어 클러스터링을 이용한 단일문서 키워드 추출에 관한 연구*

A Study on Keyword Extraction From a Single Document Using Term Clustering

한 승 희(Seung-Hee Han)**

목 차

초 록

이 연구에서는 용어 클러스터링을 이용하여 단일문서의 키워드를 추출하는 알고리즘을 제안하고자 한다. 단락단위로 분할한 단일문서를 대상으로 1차 유사도와 2차 분포 유사도를 산출하여 용어 클러스터링을 수행한 결과, 50단어 단락에서 2차 분포 유사도를 적용했을 때 가장 우수한 성능을 나타냈다. 이후, 용어 클러스터링 결과를 이용하여 단일문서의 키워드를 추출하기 위해 단순빈도와 상대빈도의 조합을 통해 다양한 키워드 추출 공식을 도출, 적용한 결과, 단락빈도(pf)와 단어빈도×역단락빈도($tf \times ipf$) 조건에서 가장 우수한 결과를 나타냈다. 이 결과를 통해, 본 연구에서 제안한 알고리즘은 좋은 키워드가 가져야 할 두 가지 조건인 주제성과 고른 빈도분포라는 측면에서 단일문서를 대상으로 효과적으로 키워드를 추출할 수 있음을 확인하였다.

ABSTRACT

In this study, a new keyword extraction algorithm is applied to a single document with term clustering. A single document is divided by multiple passages, and two ways of calculating similarities between two terms are investigated; the first-order similarity and the second-order distributional similarity. In this experiment, the best cluster performance is achieved with a 50-term passage from the second-order distributional similarity. From the results of first experiment, the second-order distribution similarity was also applied to various keyword extraction methods using statistic information of terms. In the second experiment, pf(paragraph frequency) and $tf \times ipf$(term frequency by inverse paragraph frequency) were found to improve the overall performance of keyword extraction. Therefore, it showed that the algorithm fulfills the necessary conditions which good keywords should have.

키워드: 용어 클러스터링, 키워드 추출, 단일문서, 2차 분포 유사도, 텍스트 마이닝
Term Clustering, Keyword Extraction, Single Document, Second-order Similarity, Text Mining

---

* 이 논문은 2008학년도 서울여자대학교 사회과학연구소 교내학술연구비의 지원을 받았음.
** 서울여자대학교 사회과학대학 문헌정보학과 조교수(hanshee@swu.ac.kr)
논문접수일자: 2010년 7월 19일 최초심사일자: 2010년 8월 2일 게재확정일자: 2010년 8월 11일
한국문헌정보학회지, 44(3): 155-173, 2010. [DOI:10.4275/KSLIS.2010.44.3.155]

1. 서 론

모든 문서는 그 문서를 대표할 수 있는 키워드를 포함하고 있다. Turney(2000)에 따르면, 인간의 주관적인 평가에 의해 작성된 키워드는 평균적으로 80% 정도 유용하고 한다. 그러나 인간이 전지구상에 존재하는 모든 문서의 키워드를 직접 추출한다는 것은 물리적으로 불가능한 일이다. 특히나 오늘날과 같이 정보의 홍수 속에서 살고 있는 이 시대에는 더욱 불가능한 일이 되었다. 정보의 양이 급증하면서 연구자들은 인간의 지적 노력을 통해 키워드를 추출하는 방법 대신 컴퓨터를 통해 자동으로 문서를 대표할 수 있는 키워드를 추출하는 방법에 대해 지속적으로 연구해왔다.

일반적으로 좋은 키워드가 갖추어야 할 두 가지 전제조건이 있다. 첫째, 좋은 키워드는 주제성을 반드시 내포해야 한다. 둘째, 좋은 키워드는 통계적으로 고빈도어나 저빈도어가 아닌 중간빈도어일 가능성이 높다. 두 번째 조건은 단어의 문헌식별력에 의한 자동색인의 원리를 처음으로 발견해 낸 Luhn의 아이디어에서부터 현재의 정보검색 연구자들에게까지 변함이 없는 이론적 배경이다. 그렇기 때문에 문서를 대표할 수 있는 키워드를 자동적으로 추출하기 위해 문서 내에 출현하는 단어들을 일정한 기준으로 순위화하여 문서의 가장 대표적인 단어를 찾아내는 것이 모든 정보검색 연구의 가장 기본적인 단계라고 할 수 있다.

용어 클러스터링(term clustering)은 서로 관련 있는 용어들을 일정한 기준에 따라 모아서 여러 개의 용어 클래스를 형성하는 것을 말한다(정영미 1993). 기존의 용어 클러스터링 연구는 주로 탐색용 시소러스의 자동 구축을 목적으로 하였다. 많은 연구자들이 용어 클러스터링 기법을 이용하여 자동으로 구축된 탐색용 시소러스를 이용하여 탐색자의 초기 질의와 관련된 용어를 새로운 질의에 자동으로 추가함으로써 부적절한 질의어로 인해 발생하는 검색성능의 저하를 막고 정보검색의 효율을 높이고자 하였다(Sparck Jones 1971; 서은경 1984; Lewis and Croft 1990).

용어 클러스터링에 관한 최근의 연구동향을 살펴보면, 정보검색의 효율성을 기본으로 하되, 지식에 대한 주제적이고 구조적인 이해를 돕는 도구로서의 응용가능성을 분석하는 형태로 연구가 이루어지고 있다. 예를 들어, Liu et al.(2007)은 문장요약의 연구에 김수연, 정영미(2006)는 연관용어의 선정을 위해 용어 클러스터링 기법을 활용하였다. 이러한 유형의 연구에서는 클러스터링 자체가 목적이 되는 것이 아니라 클러스터링이 다른 2차적 정보처리의 기초적인 방법론으로 사용되고 있다. 이러한 동향에 착안하여, 이 연구에서는 단일문서를 대상으로 한 자동 키워드 추출에 용어 클러스터링을 적용해보고자 한다. 자동 키워드 추출에 용어 클러스터링을 이용할 수 있는 근거는 다음과 같다. 용어 클러스터링은 문헌이 나타내고 있는 주제(용어)를 대상으로 군집을 형성하고, 그 군집은 문헌의 구조나 내용을 나타낸다. 그리고, 이 구조는 키워드를 효율적으로 추출할 수 있는 기반을 마련해준다.

한편, 키워드 추출(keyword extraction)에 대한 연구는 자동색인에 대한 연구가 처음 시작된 1950년대 이후로 많은 연구가 진행되어 왔다. 키워드 추출 연구의 의의는 인간의 지적 노

력이 수반되는 키워드 생성 작업을 기계로 하여금 대신하게 하는 기술을 개발하는 것에 있다. 기존의 연구에서는 대량의 말뭉치를 이용하여 문서집단 전체를 대표하는 키워드를 추출하는 것이 주를 이루었으나, 최근 들어, 전자문서의 급증으로 인해 자동적으로 처리해야 할 문서의 규모가 확대되면서, 키워드 추출에 관한 연구도 문서집단 전체를 대상으로 하는 것보다는 단일문서를 단위로 한 연구에 관심이 높아지고 있다(Yatsuo and Ishizuka 2004).

대용량 문서집단을 대상으로 연구된 기존의 키워드 추출 기법이 단일문서의 정보처리에 적용될 경우의 의의는 다음과 같다. 첫 번째, 단일문서에 대한 이용자의 이해를 돕는다. 단일문서를 대상으로 추출된 키워드의 경우 해당문서를 대표하는 핵심어들로 구성되므로, 이용자는 검색결과의 적합성 판정을 위한 대용물로 이를 활용할 수 있다. 두 번째, 저자에 의한 키워드와 조합하여 이용자의 접근점을 풍부하게 할 수 있다. 기존의 저자 키워드라든지, 웹 2.0 환경에서 널리 응용되고 있는 태그 등과 같은 기존의 키워드와 단일문서의 본문에서 자동 추출된 키워드를 함께 이용하면 이용자의 검색 접근점이 풍부해짐으로써 검색 효율성을 높일 수 있다. 이러한 내용을 근거로, 본 연구에서는 기존의 말뭉치 등과 같은 대량의 문서집단에 적용되어 오던 텍스트 마이닝 기법인 용어 클러스터링과 키워드 추출기법을 단일문서에 적용하는 알고리즘을 제안하고자 한다. 제안되는 알고리즘은 단일문서 단위 정보처리와 검색 효율성의 향상에 유용하게 활용될 수 있을 것이다.

2. 용어 클러스터링과 자동 키워드 추출

2.1 분포 유사도 기반 용어 클러스터링

2.1.1 통계적 연관성 측정을 위한 두 가지 접근법

일반적으로 문헌이나 용어를 자동분류하기 위해서는 문헌-용어 행렬을 구성하여 문헌간 유사도나 용어간 유사도를 산출하게 되는데, 특히 용어 클러스터링은 용어간의 통계적 연관성 분석에 기초한다. 이 때, 용어간의 통계적 연관성을 측정하기 위한 접근법에는 두 가지가 있다.

하나는, 용어의 동시출현빈도를 이용하는 방식으로, 이 방식은 용어 A와 용어 B가 많은 수의 문헌에 함께 출현하였다면 두 용어가 서로 관련이 있다고 보고 같은 클래스에 포함시키는 원리를 이용하여 용어쌍 간 유사도를 측정한다. 분류대상물간의 통계적 연관성을 측정하는 유사계수는 크게 거리계수와 유사계수로 나누어 볼 수 있는데, 일반적으로 텍스트 데이터의 통계적 연관성을 측정하는 데 있어서는 거리계수보다 유사계수가 더 적합한 것으로 나타났으며(Strehl, Ghosh, Mooney 2000), 특히 코사인 유사계수나 피어슨 상관계수가 가장 널리 활용되고 있다.

또 다른 접근법으로는 두 확률분포 사이의 차이를 측정하여 거리나 유사성을 판단하는 분포 유사도(distributional similarity) 방식을 들 수 있다. 이 방식은, 예를 들어, 용어 A와 용어 B가 다른 용어와의 동시출현 분포가 유사하다면, 용어 A와 용어 B는 서로 관련이 있다고 보

고 같은 클래스에 포함시키는 원리를 근거로 용어간 유사도를 측정한다(이재윤 2007).

일반적으로 분포 유사도 척도는 다이버전스(divergence) 공식을 일컫는다. 다이버전스는 확률분포간의 차이를 측정하는 방법으로서 정보이론에서 출발한 Kullback-Leibler 다이버전스(KL-Divergence, Kullback 1968)가 가장 대표적이며, 그 공식은 다음과 같다.

$$D(q \parallel r) = \sum_{y} q(y)(\log q(y) - \log r(y))$$

일반적으로, 용어 클러스터링에는 Jensen-Shannon 다이버전스를 사용한다. Jensen-Shannon 다이버전스는 두 확률분포를 직접 비교하지 않고 개별 지점마다 두 확률분포의 값을 평균하여 얻어낸 분포와 각 확률분포 사이의 KL 다이버전스를 산출한 후 두 다이버전스의 평균을 구하는 방법이며, 그 공식은 다음과 같다(Lin 1991).

$$JS(q,r) = \frac{1}{2}[D(q \parallel avg(q,r)) + D(r \parallel avg(q,r))]$$

이론적으로는 두 확률분포가 가까울수록 둘을 평균한 분포와 각 확률분포 사이도 가까울 것이기 때문에 Jensen-Shannon 다이버전스는 KL 다이버전스와 비례하며, 두 확률분포의 순서를 바꾸더라도 결과가 같은 대칭 공식이다(이재윤 2007).

다이버전스 공식은 특히, 용어 클러스터링 분야에서 우수한 성능을 보이는 것으로 나타났다(Dagan and Lee 1999; Lee 1999; Lin 1991; Pereira, Tishby, and Lee 1993; Weeds 2003). Lee(1999)는 신문기사에 출현한 1000개의 용어를 대상으로 여러 유형의 다이버전스 공식을 적용하여 용어 자동분류 실험을 수행하였고, 이를 통해 Jensen-Shannon 다이버전스와 스큐 다이버전스(Skew Divergence)가 용어 클러스터링에 유용하게 쓰일 수 있음을 증명하였다. 또한 Weeds(2003)는 어휘의 분포 유사도(lexical distributional similarity)를 측정하기 위한 모델을 제시하기 위해 Jensen-Shannon 다이버전스와 스큐 다이버전스를 포함하는 다양한 유사도 측정 기법들을 비교·분석하였다.

2.1.2 2차 분포 유사도 기반 용어 클러스터링

2차 분포 유사도란 1차적으로 산출된 유사도 행렬에서 각 행(혹은 열)간의 유사도를 다시 산출한 것이다. 이는 White와 Griffith(1981)가 저자동시인용분석에서 저자동시인용빈도행렬로부터 상관계수행렬을 산출한 것과 같은 방식이다(이재윤 2007).

용어 클러스터링을 위해서는 용어간 유사도를 산출하여 1차 유사도 행렬을 산출한 다음, 이 행렬에 대해서 다시 피어슨 상관계수를 적용하여 2차 유사도를 산출하게 된다. 이와 같은 2차 유사도도 두 용어간의 유사도 분포를 비교하므로 일종의 분포 유사도에 해당한다. 예를 들면, 용어 A와 용어 B가 각각 유사하고 유사하지 않은 용어가 서로 비슷하다면 용어 A와 용어 B 사이에도 깊은 연관성이 있다고 볼 수 있다. 이를 달리 생각하면, 동일한 주제의 상이한 측면을 다룬 두 용어는 1차 유사도가 높게 나타나지 않지만, 제 3의 용어와의 유사한 정도가 비슷함에 따라 가까운 문헌이라 판단할 수 있게 된다.

이재윤(2007)은 문헌을 대상으로 1차 분포 유

사도와 2차 분포 유사도 행렬을 도출하여 문헌 클러스터링을 수행하고 결과를 비교, 분석한 연구에서 2차 분포 유사도가 전반적으로 더 우수한 클러스터링 성능을 보이는 것을 증명하였다.

2.2 키워드 추출 기법

키워드 추출이란 텍스트 마이닝의 한 분야로, 자동 용어 인식(automatic term recognition), 자동 색인(automatic indexing), 자동 키워드 추출(automatic keyword extraction)이라고도 하며, 텍스트로부터 텍스트의 주제를 대표하는 키워드 또는 대표어(representative term)를 자동으로 추출해내는 것으로, 이용자로 하여금 문헌에 대한 이해와 문헌간의 관계를 쉽게 파악하도록 한다. 또한 수작업 색인이 갖는 시간, 비용 등의 여러 가지 문제를 해결하는 동시에, 정보검색에 가치를 더함으로써 정보탐색과 접근의 본질을 향상시키는 것을 목표로 한다. 일반적으로 활용되고 있는 키워드 추출 기법을 정리하면 다음과 같다.

2.2.1 통계기반 접근법

키워드 추출을 위한 통계기반 접근법의 핵심은 '주제성이 있는 단어를 어떻게 통계적으로 식별하는가'에 있다. 통계기반 접근법은 단어빈도(tf)와 역문헌빈도(idf) 등과 같이 비언어적이고 통계적인 속성을 기반으로 키워드를 추출하는 방법을 의미한다. 용어에 대한 통계정보는 문헌 내의 키워드를 식별하는 데 쓰일 수 있다.

역문헌빈도는 문헌빈도가 낮은 단어, 즉 적은 수의 문헌에 출현한 단어에 높은 중요도를 부여하는 것으로서 많은 문헌에 출현한 단어는 문헌들을 식별하는 능력이 낮다는 가설에 기초한다. 즉, 역문헌빈도는 하나의 문헌에서가 아니라 전체 문헌집단 내에서 특정한 단어가 갖는 문헌식별능력을 측정하는 가중치이다(정영미 2005).

이 접근법의 가장 큰 장점은 다른 접근법에 비해 시스템의 계산 복잡도를 줄일 수 있는 동시에 일반적으로 성능이 우수하다는 점에 있다. 가장 대표적인 통계기반 접근법으로는 Sparck Jones(1972)가 제안한 $tf \times idf$ 공식이 있으며, 다음과 같다.

$$tf \times idf = tf \times \log \frac{N}{df}$$

(N= 전체 문헌 수, df= 문헌빈도)

2.2.2 언어학적 접근법

품사 및 구문분석 등을 통해 키워드를 추출하는 언어학적 접근법은 단어, 문장, 문헌 등의 언어학적 자질을 활용한다. Plas et al.(2004)은 EDR과 WordNet에 있는 어휘간 의미관계(IS-A 또는 PART-OF)를 이용하여 키워드를 추출하는 연구를 수행하였고, Hulth(2003)는 기본적인 통계적 접근법에 기반하여 해당 키워드의 품사를 분석하여 기존의 통계적 기법에 근거한 키워드 추출법의 결과를 향상시켰다. 이러한 언어학적 접근법은 단독으로 활용되기보다는 단어빈도나 역문헌빈도 등과 같은 통계적 접근법과 함께 이용되는 경우가 일반적이다.

2.2.3 기계학습 기반 접근법

기계학습 기반 접근법은 사례를 통한 지도학습을 기반으로 키워드를 추출하는 방식이다. 이러한 시스템은 수작업 키워드 등과 같은 학

습문서집단을 이용하여 주제범주를 학습한 후, 학습내용을 기반으로 시스템에 들어오는 입력문서의 키워드를 추출하는 방식을 의미한다.

Witten et al.(1999)은 KEA(Keyphrase Extraction System) 시스템을 고안하였는데, 이 시스템에서는 핵심문장을 추출하기 위해 나이브 베이즈 분류기를 이용한 알고리즘을 활용하였다. Suzuki et al.(1998)은 라디오 뉴스에서 키워드를 추출하기 위해 백과사전과 신문기사를 이용한 기계학습 기반 키워드 추출 시스템을 제안하였다. 이 시스템에서는 먼저, 학습집단인 백과사전과 신문기사로부터 자질 벡터를 생성하고, 이를 이용하여 라디오 뉴스를 자동분류한 후 학습집단의 키워드를 기반으로 라디오 뉴스의 키워드를 추출하는 방법을 이용하였다.

2.3 선행연구

키워드 추출은 자동색인이 처음 연구되던 1950년대부터 많은 연구자들의 관심을 받아왔으며, 최근에는 기존의 $tf \times idf$ 모델을 확장, 변형하거나, 폭소노미와 자동 추출 키워드를 비교・분석하거나, 대량문서보다는 단일문서를 대상으로 하는 형태로 연구가 진행되고 있다. 키워드 추출의 최근 연구들을 살펴보면 다음과 같다.

Matsuo and Ishizuka(2004)는 단일문서를 대상으로 키워드를 추출하는 연구를 수행하였는데, 특정 용어의 x^2 통계량을 계산하여 이 값이 높은 용어를 키워드로 추출하였다. 이 연구에서는 일정 빈도 이상의 용어를 추출한 후, 문장단위로 용어쌍 간의 동시출현빈도를 계산하고, 특정 용어가 특정 용어군과 함께 출현하면 해당 용어는 중요한 의미를 가지고 있다는 가정하에, 이를 계산하기 위해 x^2 통계량을 활용하였다. 그 결과, $tf \times idf$ 모델을 이용하여 추출한 키워드에 비해 정확률이 높은 것으로 나타났다. 이 연구결과는 단일문서를 대상으로 키워드를 추출하였고, 용어쌍 간의 동시출현빈도 정보를 이용했다는 점에서 본 연구에서 진행하고자 하는 방법과 유사하나, 단일문서를 처리하는 방법에 있어 문장단위의 분할만을 고려하였다는 점에서 차이가 있다.

Al-Khalifa and Davis(2006)는 키워드 자동추출에 대한 대안으로 폭소노미 태그의 활용방안에 대해 연구하였다. 이 연구에서는 자동으로 추출된 키워드와 폭소노미의 특징을 비교하고, 폭소노미가 자동 추출 키워드보다 의미 있는 결과를 나타낼 것이라고 가정한 후, 웹 문서를 대상으로 폭소노미와 자동 추출 키워드의 중복도를 계산하고, 평가자에 의해 두 유형의 키워드의 품질을 평가하는 연구를 수행하였다. 연구결과, 웹 문서에 대해서는 폭소노미 태그가 자동 추출 키워드에 의해 의미수준의 메타데이터를 추출할 수 있다는 것을 밝혀냈다. 그러나, 이 연구의 결과는, 학술문서 집단을 대상으로 했을 때 폭소노미를 활용하는 것이 상대적으로 어려운 한계점이 있어, 일반화하기 어렵다.

키워드 추출에 관한 최근의 국내 연구로는 이성직, 김한준의 연구(2009)가 있다. 이 연구에서는 대용량 뉴스문서집합을 대상으로 키워드 추출을 수행하여 분야별 주제를 제시할 수 있는 키워드를 추출하는 기법을 제안하였는데, 기본적인 $tf \times idf$ 모델을 기반으로 6가지 변형된 공식을 마련하여 각 분야별로 후보 키워드를 추출하였다. 또한 분야별로 추출된 단어들의 분야별 교차비교분석을 통해 의미없는 단어

를 제거함으로써 그 성능을 높이는 방법을 사용하였다. 이 방법은 기존의 $tf \times idf$ 모델을 다양한 형태로 변형, 적용했다는 데에 의의가 있으나, 대량의 문서집합을 대상으로 전체 문헌집단을 대표하는 키워드를 추출한 것이므로 본 연구가 지향하는 단일문서의 키워드 추출실험과는 그 관점이 다르다.

또한, 이주호, 김학수(2009)는 단어간 의존관계를 이용하여 단일문서의 키워드를 추출하는 연구를 수행하였는데, 단어간 의존관계는 Google의 PageRank 알고리즘을 응용하였다. 연구 결과, 대량의 말뭉치를 사용하지 않고도 기존에 연구된 방법들보다 높은 성능으로 키워드를 추출할 수 있었다. 그러나, 이 연구는 단일문서를 활용해 전체 문헌집단을 대표하는 키워드를 추출하는 방식으로, 단일문서를 활용해 해당문서의 핵심 키워드를 추출하는 본 연구와는 그 특징이 다르다.

3. 실험 설계

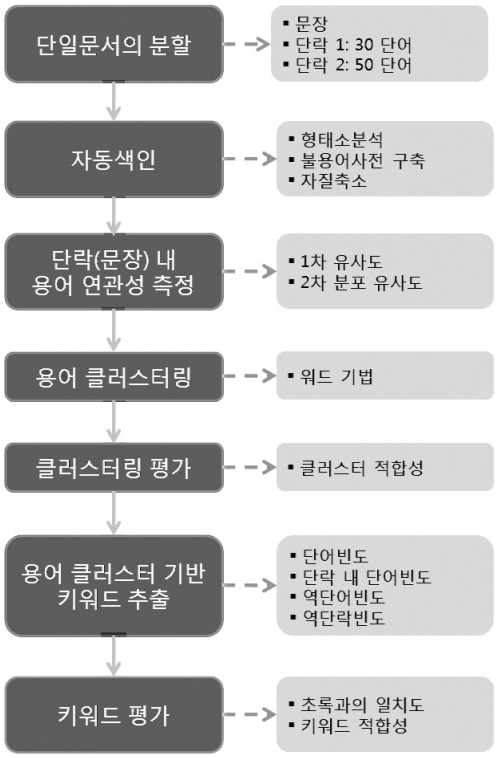
3.1 실험 개요

3.1.1 용어 클러스터링

대부분의 정보검색 실험에서는 코퍼스에 포함된 복수의 문서를 대상으로 다양한 기법들을 적용한다. 그러나 이 연구의 목적은 단일문서를 대상으로 키워드를 추출하는 것이기 때문에 실험대상의 특성이 기존의 정보검색 실험과는 다르다. 이 연구에서는 최근 5년간 발표된 문헌정보학 분야 국내 학술지 논문 20편을 대상으로 실험집단을 구축하였다. 이 실험에서 텍스트 처리의 기본 단위는 단락과 문장이다. 그러므로 이 실험에서 사용한 단어빈도(tf), 단락 내 단어빈도(term frequency in passage, 이하, tfp), 단락빈도(passage frequency, 이하 pf)는 개념적으로 각각 장서빈도(cf), 단어빈도(tf), 문헌빈도(df)와 같다. 실험 과정은 〈그림 1〉과 같다.

문서를 단락으로 분할하는 방법 중 가장 손쉬운 것은 저자가 나눈 문단이나 문장을 그대로 이용하는 것이다. 그러나 일반적으로 모든 단락의 길이가 유사하지 않으므로 단락의 길이를 정규화할 필요가 있다. 단락검색과 지역적 질의확장 분야를 중심으로 단락길이의 정규화 방법에 대해서 많은 연구가 있었다. Callan(1994)은 단락검색에서 문단이 너무 짧거나 길어지는 것을 막기 위해 문단의 크기를 결정한 후 긴 문단은 일정 크기 이상이 되면 분할하고 짧은 문단은 합치는 방법을 제안하였다. 또한 Zobel et al.(1995)은 단락분할에 휴리스틱을 적용하여 30~300단어를 기준으로 단락을 결정하였다. 이러한 연구에서, 특정 크기의 단어 창(word window)을 만들어 문헌을 고정길이 단락으로 분할하는 것이 저자에 의한 문단으로 나누는 것보다 성능이 우수하는 것이 공통적인 결과로 나타났기 때문에, 이 실험에서도 단락의 길이를 30개 단어와 50개 단어로 고정하여 분할하는 방식을 사용하였다. 또한, 단락 이외에도 문장 내에서의 용어간 연관성을 이용하여 클러스터링 실험을 수행하기 위해 문헌을 문장단위로도 분할하였다.

단락 수준이 결정된 후에는 단락을 기준으로 색인어를 추출하였다. 색인 과정에서는 한 글자짜리 단어와 숫자, 기호, 그리고 $tf = 3$ 이하의 저빈도어를 불용어로 제거하였다. 특히 저빈도어를 제거함으로써 효과적으로 자질을 축소할 수 있었다.



〈그림 1〉 실험 개요

단락 혹은 문장 내 출현한 용어들간의 연관성은 코사인 유사계수를 활용한 1차 유사도와, 이에 다시 피어슨 상관계수를 적용하여 얻은 2차 분포 유사도를 이용하였다. 특히 2차 분포 유사도는 앞에서 설명한 것처럼, 용어 클러스터링에 유용한 분포 유사도의 일종으로 볼 수 있기 때문에 1차 유사도를 활용한 클러스터링과의 비교를 통해 어떠한 유사도 측정 방식이 용어 클러스터링에 더욱 효과가 있는지를 확인할 수 있다. 용어 x와 용어 y에 대해 x_i는 단락(문장) i에 출현한 용어 x의 가중치이며, y_i는 단락(문장) i에 출현한 용어 y의 가중치일 때, 코사인 유사계수 $\cos(x,y)$와 피어슨 상관계수 $r(x,y)$의 공식은 다음과 같다(Sneath and Sokal 1973).

$$\cos(x,y) = \frac{\sum_i (x_i y_i)}{\sqrt{(\sum_i x_i^2)(\sum_i y_i^2)}}$$

$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

$$(\bar{x} = \frac{1}{n}\sum_i x_i,\ \ \bar{y} = \frac{1}{n}\sum_i y_i,\ \ i = 1...n)$$

단락 및 문장 내 용어 연관성 측정을 형성된 용어 - 용어 행렬을 통해 워드 클러스터링 기법(Ward's method)을 수행하여 각 문서별로 10개의 용어 클러스터를 생성한 후 각 조건에서 생성된 클러스터 결과를 평가하여 최적의 용어 클러스터 생성조건을 결정하였다.

3.1.2 용어 클러스터 기반 키워드 추출

최적의 클러스터링 기법으로 생성된 10개의 클러스터를 대상으로 키워드 추출 실험을 수행하였다. 클러스터링을 통해 키워드를 추출하는 이론적 근거는 다음과 같다.

클러스터 표현(cluster representation)은 내부적 표현(internal representation)과 외부적 표현(external representation)으로 나뉜다(Tombros 2002). 내부적 표현은 클러스터 기반 검색 등과 같이 클러스터를 가지고 2차적인 응용을 할 때 클러스터의 내용을 요약하는 클러스터 센트로이드를 형성하는 것이고, 외부적 표현은 클러스터의 내용을 텍스트나 그래픽으로 표시하여 이용자가 클러스터의 구조적 이해를 돕도록 지원한다. 용어 클러스터에 외부적 표현기법을 적용한다면, 각 용어 클러스터의 대표어나 키워드 추출을 통해 클러스터의 내용을 표현함으로써 해당 문헌의 구조적이고 주제적인 이해를 도울 수 있다(한승희, 정영미 2004).

본 연구에서는 생성된 개별 클러스터를 대상으로 통계적 특성에 기반한 키워드 추출 기법을 적용하여 각 클러스터를 대표하는 용어를 추출하고, 이를 키워드로 간주하였다. 즉, 단일문서 당 생성된 10개의 클러스터에서 각각 하나의 키워드가 추출되면 한 문서 당 10개의 키워드를 추출하게 된다. 이 연구에서 적용한 통계기반의 키워드 추출 기법은 〈표 1〉과 같다.

통계기반의 키워드 추출 기법은 크게 단순빈도를 이용한 방법과 상대빈도를 이용한 방법으로 나누어 볼 수 있다. 단일문서를 대상으로 했을 때 좋은 색인어란, 여러 단락에 고르게 출현하는 단어보다는 일부의 단락에 집중적으로 출현하는 단어가 주제식별력을 가질 것이다. 이러한 특징을 나타내는 용어를 식별하기 위해,

〈표 1〉 실험에서 사용한 통계기반 키워드 추출 기법

빈도 특성	기법	설명/공식
단순빈도	tf	문헌 전체에서의 단어의 출현 빈도(단어빈도)
	pf	단어가 출현한 단락의 수(단락빈도)
	tfp	단락 내에서의 단어의 출현빈도(단락 내 단어빈도)
상대빈도	$tf \times \ln(tf)$	단어빈도와 단어빈도의 자연로그의 곱
	$tf \times itf$	단어빈도와 역단어빈도(itf)의 곱 $tf \times \log_2 \frac{N}{n}$ (N= 문서 내 총 단어 수, n = 용어 i의 출현빈도)
	$tf \times ipf$	단어빈도와 역단락빈도(ipf)의 곱 $tf \times \log_2 \frac{N}{n}$ (N= 총 단락 수, n = 용어 i의 단락빈도)
	$tfp \times \ln(tf)$	단락 내 단어빈도와 단어빈도의 자연로그의 곱
	$tfp \times itf$	단락 내 단어빈도와 역단어빈도(itf)의 곱 $tfp \times \log_2 \frac{N}{n}$ (N= 문서 내 총 단어 수, n = 용어 i의 출현빈도)
	$tfp \times ipf$	단락 내 단어빈도와 역단락빈도(ipf)의 곱 $tf \times \log_2 \frac{N}{n}$ (N= 총 단락 수, n = 용어 i의 단락빈도)

단순빈도인 tf와 tfp의 값을 정규화한 $\ln(tf)$, itf(역단어빈도), ipf(역단락빈도)를 적용하였다. $\ln(tf)$와 itf는 문서 전체를 대상으로 용어의 단순빈도를 문서 전체를 대상으로 전역적으로 정규화하기 위해 사용되었고, ipf는 단락 내 용어의 출현빈도를 단락 내에서 지역적으로 정규화하기 위해 사용되었다.

3.2 평가 방법

다양한 조건의 클러스터를 생성한 후, 어떠한 조건에서 용어 클러스터링이 가장 효과적으로 이루어지는지를 확인하기 위해 클러스터링 결과를 평가하였다. 한승희, 정영미(2004)는 클러스터링 기법이 주제적으로 연관성 있는 용어들을 효과적으로 군집화하는지를 측정하기 위해서 클러스터 적합도를 제안하였으며, 그 공식은 다음과 같다.

$$\text{클러스터 적합도} = \frac{\text{클러스터 대표주제에 적합한 용어 수}}{\text{클러스터에 속한 용어 수}}$$

또한, 본 연구에서 제안한 키워드 추출 기법이 얼마나 효과적인가를 살펴보기 위해 다양한 기법으로 추출된 키워드를 대상으로 초록과의 일치도와 키워드 적합도를 평가하였다. 본 연구의 실험대상이 된 학술논문은 저자에 의한 키워드를 가지고 있으나, 시스템에서 추출한 키워드의 수와 저자에 의한 키워드의 수가 일치하지 않아 객관적으로 평가하기 어렵다. 그러므로, 본 연구에서는 추출된 키워드가 본문의 주제를 얼마나 충실하게 표현하고 있는지를 평가하기 위해 학술논문의 내용을 대표하는 초록에 출현한 용어와 자동으로 추출된 키워드를 비교하고, 또한 추출된 키워드가 키워드로서 얼마나 적합한지를 살펴보기 위해 키워드 적합도를 평가하였다. 키워드 적합도의 평가는 본 연구자에 의해 수행되었다. 추출된 키워드의 평가 공식은 다음과 같다.

$$\text{초록과의 일치도} = \frac{\text{초록에 출현한 키워드 수}}{\text{문서당 추출된 키워드 총 수}}$$

$$\text{키워드 적합도} = \frac{\text{키워드로 적합한 용어 수}}{\text{문서당 추출된 키워드 총 수}}$$

4. 실험 결과의 분석

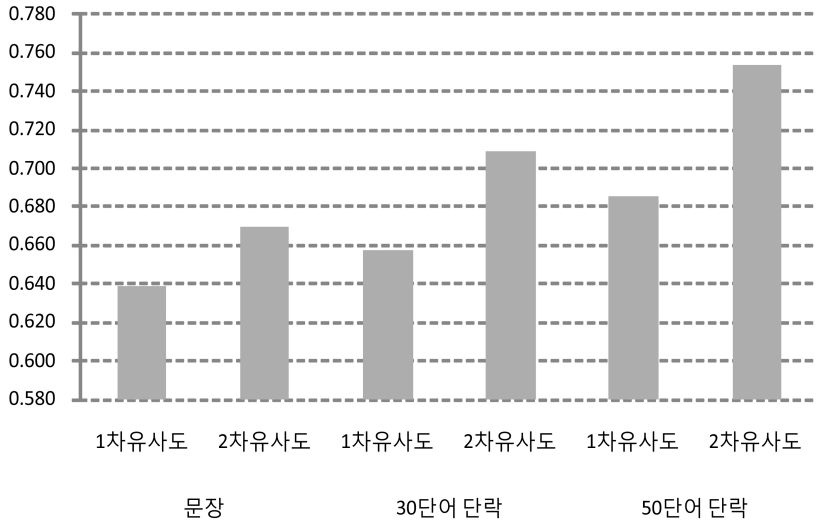
4.1 용어 클러스터링

다양한 용어 클러스터 생성조건 중에서 최적의 조건을 선정하기 위해 용어 클러스터 적합도를 평가한 결과, 〈표 2〉 및 〈그림 2〉에서 보는 바와 같이 50단어 단락을 기준으로 2차유사도를 적용했을 때 가장 우수한 성능을 보이는 것으로 나타났다. 또한, 일반적으로 클러스터의 크기가 작고 균일한 것이 좋은 클러스터라고 하는데, 본 연구에서 제시한 클러스터 생성조건이 좋은 클러스터링 결과를 가져왔는지 확인하기 위해 각 조건별로 클러스터당 소속 용어 수의 평균 표준편차를 계산하였다. 〈표 3〉에서 보는 바와 같이, 클러스터링 성능이 우수할수록 클러스터 당 소속 용어 수의 표준편차도 낮은 것을 알 수 있다.

클러스터 생성조건 중 문서의 분할단위를 대상으로 클러스터링 결과를 평가해보면, 문장단

〈표 2〉 용어 클러스터 적합도 평가 결과

조건 / 문서	문장		30단어 단락		50단어 단락		평균
	1차유사도	2차유사도	1차유사도	2차유사도	1차유사도	2차유사도	
doc_1	0.700	0.700	0.644	0.711	0.678	0.689	0.687
doc_2	0.573	0.547	0.560	0.600	0.613	0.740	0.606
doc_3	0.630	0.658	0.753	0.733	0.712	0.781	0.711
doc_4	0.639	0.648	0.680	0.738	0.730	0.770	0.701
doc_5	0.725	0.758	0.791	0.802	0.791	0.769	0.773
doc_6	0.652	0.717	0.688	0.724	0.720	0.753	0.709
doc_7	0.521	0.620	0.555	0.648	0.633	0.759	0.623
doc_8	0.717	0.764	0.727	0.782	0.739	0.787	0.753
doc_9	0.522	0.618	0.577	0.682	0.602	0.732	0.622
doc_10	0.655	0.652	0.672	0.732	0.688	0.782	0.697
doc_11	0.572	0.633	0.598	0.652	0.605	0.713	0.629
doc_12	0.601	0.654	0.614	0.678	0.632	0.719	0.650
doc_13	0.712	0.723	0.736	0.778	0.754	0.812	0.753
doc_14	0.701	0.734	0.712	0.772	0.736	0.792	0.741
doc_15	0.615	0.692	0.642	0.712	0.678	0.766	0.684
doc_16	0.700	0.701	0.668	0.712	0.682	0.775	0.706
doc_17	0.595	0.591	0.612	0.654	0.652	0.712	0.636
doc_18	0.711	0.722	0.742	0.764	0.734	0.782	0.743
doc_19	0.645	0.623	0.592	0.599	0.677	0.689	0.638
doc_20	0.585	0.632	0.592	0.695	0.661	0.742	0.651
평균	0.639	0.669	0.658	0.708	0.686	0.753	0.686



〈그림 2〉 클러스터 생성조건별 용어 클러스터 적합도 평균

〈표 3〉 클러스터 당 용어 수의 평균 표준편차

조건 / 표준편차	문장		30단어 단락		50단어 단락	
	1차유사도	2차유사도	1차유사도	2차유사도	1차유사도	2차유사도
평균	9.875	8.284	7.425	6.671	6.661	4.536

위 분할보다는 단락단위로 분할한 경우가 더 나은 클러스터링 성능을 보였다. 문장단위 분할의 두 가지 조건인 30단어 단락과 50단어 단락 중에서는 50단어 단락의 경우가 성능이 우수한 것으로 나타났다. 문장단위 분할의 경우 단락단위에 비해 용어간 동시출현빈도가 상대적으로 낮기 때문에 용어간 유사성의 측정이 제대로 이루어지지 않았을 것으로 추측가능하다. 이러한 관점에서, 30단락보다 50단락의 경우에 성능이 우수한 것도 문서의 분할규모가 일정 수준 이상이 되어야 용어간 유사성이 효과적으로 측정될 수 있다고 해석할 수 있다.

용어간 연관성 측정 방식에 대해 분석해보면, 평균적으로 용어간 동시출현빈도를 기반으로 코사인 유사도에 기반한 1차 유사도에 비해 1차 유사도 행렬을 기반으로 피어슨 상관계수를 적용하여 생성된 2차 분포 유사도 기반 클러스터링이 더 나은 성능을 보였다. 이를 통해 분포 유사도가 용어 클러스터링에 적합하다는 기존의 연구결과가 단일문서를 대상으로 했을 때에도 유효하다는 것을 확인할 수 있다(표 4 참조).

4.2 키워드 추출

앞에서 언급했듯이, 좋은 키워드의 전제조건은 일반적으로 두 가지라 할 수 있다. 첫 번째는 주제성이 있어야 한다. 두 번째로는 주제성이 있는 단어는 고빈도어나 저빈도어가 아닌 경우가 일반적이다. 본 연구에서 제시한 알고리즘이 이러한 두 가지 조건을 만족시키고 있는지

〈표 4〉 50단어 단락 2차 분포 유사도 조건에서의 용어 클러스터링 결과(doc_1)

클러스터 1	ASIS&T, UF, 갱신, 관계, 관계정보, 디스크립터, 부여, 수행, 시소러스, 실험, 용어, 위키피디아, 의미관계, 적합, 주제, 주제영역, 추가, 추출, 평가, 평균, 포괄
클러스터 2	COLLECTIVE, INTELLIGENCE, THESAURUS, WIKIPEDIA
클러스터 3	ENGINE, SEARCH, VOSS, WEB, 구조, 링크, 문서, 분류체계, 상위, 상호, 수정, 시스템, 어휘, 연관, 연관관계, 카테고리
클러스터 4	KNOWLEDGE, LIBRARIES, MANAGEMENT, PUBLIC, SEMANTIC, 개념, 문헌정보학, 표현
클러스터 5	관리, 구조적, 내용, 리디렉션, 전문가, 지식관리, 품질, 협업
클러스터 6	구글, 구성원, 기능, 기술, 백과사전, 분류, 생산, 웹, 웹2.0, 이용, 이용자, 인터넷, 지성, 지식, 집단, 집단지성, 참여, 탐색
클러스터 7	구축, 비용, 시간, 유지, 태깅, 협력
클러스터 8	능력, 정보, 정보검색
클러스터 9	동등관계, 의미, 중복, 통제어휘
클러스터 10	언어, 커뮤니케이션

확인하기 위해, 단순빈도와 상대빈도를 이용하여 클러스터에서 키워드를 추출하고, 이를 초록과의 일치도 및 키워드 적합도로 평가하였다. 그 결과는 〈표 5〉, 〈표 6〉과 같다.

〈표 5〉의 초록과의 일치도 평가결과를 보면, 단순빈도 계열에서는 pf가, 상대빈도 계열에서는 $tf \times ipf$가 우수한 결과를 나타냈다. 〈표 6〉의 추출 키워드 적합도에서도 동일한 양상으로 pf와 $tf \times ipf$가 가장 우수한 키워드 추출기법으로 나타났다. 추출된 키워드의 평가기준별 기법의 성능 추이는 〈그림 3〉과 같다. 〈그림 3〉에서 보는 바와 같이, 두 가지 평가기준에서 모두 기법별로 거의 유사한 성능패턴을 보이는 것을 알 수 있다. 즉, 특정 기법에서 초록과의 일치도가 높으면 키워드 적합도도 높다.

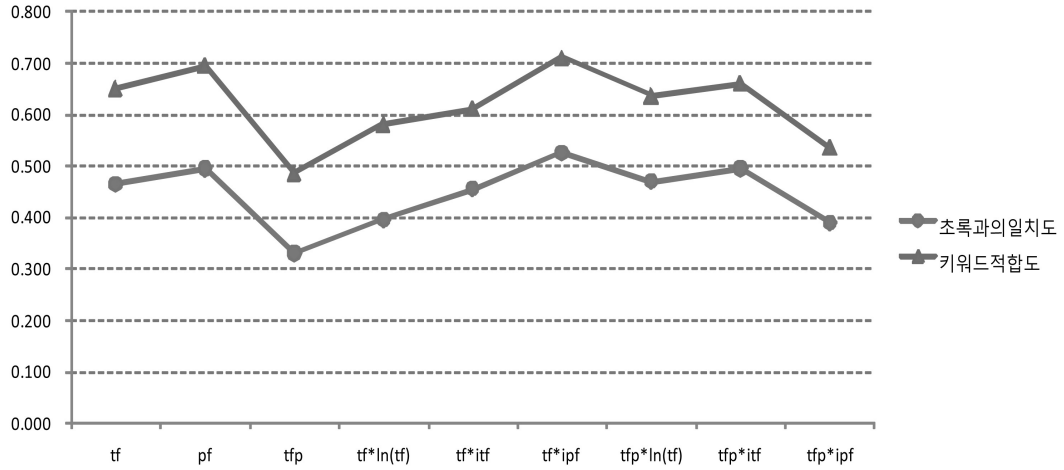
단순빈도 계열에서 pf가 키워드 추출에 있어 높은 성능을 보인 이유는 단락 단위의 문서분할 환경에서 특정 용어의 단락빈도가 용어의 주제적 특성을 잘 반영했기 때문이라고 볼 수 있다. 또한 상대빈도 계열에서 $tf \times ipf$가 우수한 성능을 보인 것은 ipf가 tf를 정규화시킴으로써 단어빈도나 단락 내 단어빈도의 편차를 축소하면서 이들이 갖는 빈도특성을 효과적으로 나타낸 것으로 해석할 수 있다. 반면, $tf \times \ln(tf)$와 $tfp \times ipf$ 기법이 다른 기법에 비해 성능이 낮은 이유는 동일한 요소를 이용하여 두 번 정규

〈표 5〉 추출 키워드의 초록과의 일치도

기법 / 문서	단순빈도			상대빈도					
	tf	pf	tfp	$tf \times \ln(tf)$	$tf \times itf$	$tf \times ipf$	$tfp \times \ln(tf)$	$tfp \times itf$	$tfp \times ipf$
doc_1	0.5	0.5	0.3	0.4	0.5	0.5	0.6	0.5	0.4
doc_2	0.5	0.6	0.3	0.4	0.4	0.5	0.4	0.4	0.4
doc_3	0.4	0.5	0.4	0.5	0.4	0.5	0.5	0.6	0.5
doc_4	0.6	0.6	0.4	0.3	0.5	0.6	0.4	0.6	0.3
doc_5	0.6	0.6	0.5	0.6	0.6	0.7	0.4	0.4	0.3
doc_6	0.5	0.7	0.4	0.3	0.3	0.6	0.4	0.5	0.5
doc_7	0.4	0.4	0.4	0.3	0.4	0.5	0.6	0.5	0.4
doc_8	0.5	0.5	0.3	0.4	0.5	0.5	0.5	0.4	0.4
doc_9	0.5	0.5	0.3	0.5	0.3	0.6	0.5	0.3	0.5
doc_10	0.4	0.4	0.3	0.5	0.6	0.7	0.4	0.5	0.5
doc_11	0.4	0.3	0.2	0.5	0.3	0.5	0.4	0.3	0.3
doc_12	0.5	0.4	0.4	0.4	0.4	0.5	0.3	0.4	0.3
doc_13	0.4	0.5	0.3	0.3	0.4	0.6	0.6	0.5	0.4
doc_14	0.5	0.5	0.3	0.3	0.5	0.6	0.6	0.6	0.5
doc_15	0.5	0.5	0.3	0.2	0.5	0.6	0.5	0.6	0.4
doc_16	0.3	0.6	0.4	0.3	0.6	0.4	0.4	0.5	0.3
doc_17	0.5	0.6	0.2	0.4	0.4	0.4	0.4	0.6	0.3
doc_18	0.4	0.4	0.3	0.5	0.5	0.5	0.3	0.6	0.4
doc_19	0.5	0.4	0.2	0.4	0.5	0.3	0.6	0.6	0.3
doc_20	0.4	0.4	0.4	0.4	0.5	0.4	0.6	0.5	0.4
평균	0.465	0.495	0.330	0.395	0.455	0.525	0.470	0.495	0.390

〈표 6〉 추출 키워드의 적합도

기법 / 문서	단순빈도			상대빈도					
	tf	pf	tfp	$tf \times \ln(tf)$	$tf \times itf$	$tf \times ipf$	$tfp \times \ln(tf)$	$tfp \times itf$	$tfp \times ipf$
doc_1	0.7	0.6	0.5	0.5	0.5	0.7	0.5	0.7	0.5
doc_2	0.8	0.7	0.4	0.5	0.6	0.7	0.6	0.6	0.5
doc_3	0.7	0.6	0.3	0.5	0.5	0.6	0.6	0.7	0.6
doc_4	0.7	0.7	0.5	0.5	0.6	0.8	0.5	0.7	0.5
doc_5	0.6	0.7	0.4	0.7	0.6	0.9	0.6	0.5	0.7
doc_6	0.5	0.7	0.3	0.6	0.5	0.6	0.8	0.6	0.5
doc_7	0.5	0.5	0.5	0.6	0.6	0.6	0.7	0.7	0.5
doc_8	0.6	0.6	0.6	0.7	0.6	0.7	0.6	0.6	0.6
doc_9	0.7	0.6	0.6	0.5	0.7	0.8	0.7	0.7	0.4
doc_10	0.6	0.8	0.5	0.6	0.7	0.7	0.5	0.6	0.6
doc_11	0.7	0.7	0.4	0.5	0.6	0.7	0.5	0.7	0.6
doc_12	0.5	0.8	0.6	0.6	0.5	0.8	0.6	0.8	0.4
doc_13	0.6	0.7	0.5	0.6	0.7	0.7	0.8	0.6	0.5
doc_14	0.6	0.8	0.4	0.6	0.5	0.6	0.7	0.7	0.6
doc_15	0.7	0.7	0.6	0.5	0.6	0.5	0.7	0.7	0.5
doc_16	0.8	0.7	0.5	0.7	0.6	0.8	0.7	0.6	0.6
doc_17	0.8	0.7	0.4	0.6	0.7	0.6	0.6	0.5	0.5
doc_18	0.7	0.8	0.6	0.6	0.7	0.7	0.6	0.8	0.6
doc_19	0.6	0.7	0.6	0.6	0.7	0.9	0.7	0.7	0.4
doc_20	0.6	0.8	0.5	0.6	0.7	0.8	0.7	0.7	0.6
평균	0.650	0.695	0.485	0.580	0.610	0.710	0.635	0.660	0.535



〈그림 3〉 키워드 추출 기법별 성능의 추이 변화

〈표 7〉 추출된 키워드와 단락빈도 특성

(doc_1, doc_2, $tf \times ipf$ 조건)

doc_1		doc_2	
term	pf 순위	term	pf 순위
시소러스	1	프로그램	124
THESAURUS	51	SERVICES	124
카테고리	15	이용제공	142
개념	13	교수안	21
리디렉션	27	기록	3
집단지성	6	가계기록	27
구축	20	RESEARCH	45
정보	20	기록정보	8
의미	34	정보조사제공	149
언어	83	연구안내자료	142

화함으로써 정규화의 효과가 감소했기 때문으로 분석된다. 즉, $tf \times \ln(tf)$는 두 번 모두 전역적 tf를 이용하였고, $tfp \times ipf$ 기법은 지역적 단락 정보를 반복해서 이용했기 때문에 효과적으로 빈도값을 정규화하지 못한 것으로 해석된다.

일반적인 정보검색 환경에서 고빈도어나 저빈도어는 기능어나 희귀어이기 때문에 문헌의 주제를 잘 나타내지 못하고 중간빈도어가 문헌의 주제를 잘 나타내는 것으로 알려져 있다. 이러한 이론이 클러스터링을 통해 추출된 키워드에도 적용이 되는지 살펴보기 위해 각 군집에서 추출된 키워드의 빈도 특성을 살펴보았다. 〈표 7〉은 전체 20개 문헌 중 2개 문헌에서 $tf \times ipf$ 조건으로 추출된 키워드와 이들의 단락빈도를 살펴본 것으로, 보는 바와 같이 고른 단락빈도 분포를 나타내고 있어, 이 연구에서 제시한 알고리즘이 특별히 고빈도어나 저빈도어를 선호하지 않고 고른 빈도분포를 갖도록 키워드를 추출할 수 있는 것으로 확인되었다.

추출된 키워드를 분석한 결과, 키워드 추출 과정에서 대부분의 고빈도어가 누락된 것을 볼 수 있었는데, 그 주된 이유는 고빈도어간의 동시 출현빈도가 높았기 때문이다. 즉, 두 고빈도어간의 연관성이 높게 측정되어 한 클러스터에 속하게 되면서 빈도가 높음에도 불구하고 두 고빈도어 중 상대적으로 빈도가 낮은 용어는 키워드로 선정되지 못하고 누락된 것이다. 그렇기 때문에 클러스터링에 기반한 키워드 추출 방법은 키워드가 상대적으로 저빈도어를 선호하는 경향을 보이면서 고른 빈도분포를 갖게 하는데 영향을 주어 단일문서의 주제성 표현에도 긍정적인 영향을 미친다고 할 수 있다.

5. 결 론

키워드 추출은 정보검색, 자동분류, 요약, 주제탐지 등 텍스트 마이닝 분야에서 기반이 되는 기술이다. 문서로부터 추출된 키워드는 텍스트 마이닝을 위한 중요 속성으로 활용되어 문서의 브라우징, 주제탐지, 자동분류, 정보검색 시스템의 성능향상 등에 기여한다. 특히, 최

근에는 전자문서의 급증으로 인해 처리해야 할 문서의 수가 많아지면서, 전체 문서집단을 대상으로 하는 것 보다는 단일문서를 대상으로 키워드를 추출하고 이를 이용하여 검색 효율성을 향상시키는 것에 대해 연구자들이 관심을 나타내고 있다.

본 연구에서는 단일문서를 대상으로 용어 클러스터링을 이용하여 특별히 고빈도어를 선호하지 않고, 빈도분포가 고르면서, 주제성이 있는 키워드를 추출할 수 있는 알고리즘을 제안하였다. 먼저 단일문서를 단락단위로 분할한 후 이를 대상으로 용어간 연관성을 측정하였다. 용어간 연관성을 측정하기 위해서는 코사인 유사도에 근거한 1차 유사도와 더불어 이 1차 유사도를 피어슨 상관계수로 한 번 더 연관성 측정을 한 2차 분포 유사도를 활용하였다. 2차 유사도는 분포 유사도의 일종으로 빈도에 근거한 용어간 연관성 이외에 용어간의 숨은 관계를 밝혀내기 때문에 용어 클러스터링에 적합하다고 알려져 있다. 워드 기법으로 각 문서별 10개의 클러스터가 생성되도록 한 후, 이 결과를 클러스터 적합도로 평가한 결과, 50단어 단락 단위로 분할하고 2차 유사도를 이용했을 때 가장 우수한 클러스터링 성능을 보이는 것으로 나타났다.

다음으로는, 최적의 클러스터링 성능을 나타낸 조건 하에서 생성된 클러스터를 대상으로 키워드 추출 기법을 적용하였다. 주제성이 있으면서 고른 빈도분포를 갖는 키워드를 추출하기 위해 이 연구에서는 단순빈도와 상대빈도를 이용하여 클러스터에서 키워드를 추출하였다. 그리고 추출된 키워드가 좋은 키워드의 조건을 만족시키는지 확인하기 위해 초록과의 일치도 및 키워드 적합도로 주제성을 평가하였다. 평가 결과, 초록과의 일치도와 키워드 적합도에서 모두 pf와 $tf \times ipf$가 우수한 결과를 나타내어, 본 연구에서 제시한 방법이 어느 정도 주제성이 있는 키워드를 추출할 수 있음을 증명하였다. 또한 추출된 키워드가 고빈도어부터 저빈도어까지 고른 빈도분포를 보임으로써 일반적으로 알려진 좋은 키워드의 전제조건을 만족시키는 것을 확인하였다.

키워드 추출 연구는 검색효율성 향상을 위한 것이 목적이다. 그렇기 때문에, 이 연구에서 제안한 알고리즘으로 생성된 키워드가 실제로 검색효율성 향상에 긍정적인 영향을 미치는지에 대해 연구해볼 필요가 있다. 또한, 본 연구에서 제안한 기법을 다양한 유형의 문헌에 적용함으로써 이 기법의 성능을 일반화하는 후속연구가 필요하다.

참 고 문 헌

[1] 김수연, 정영미. 2006. 텍스트 마이닝 기법을 이용한 연관용어 선정에 관한 실험적 연구. 『정보관리학회지』, 23(3): 147-165.

[2] 서은경. 1984. 용어의 자동분류에 관한 연구. 『정보관리학회지』, 1(1): 78-99.

[3] 유사라. 1999.『정보학연구와 분석방법론』. 서울: 나남출판.

[4] 이성직, 김한준. 2009. TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법.『한국전자거래학회지』, 14(4): 59-73.

[5] 이재윤. 2007. 분포 유사도를 이용한 문헌클러스터링의 성능향상에 대한 연구.『정보관리학회지』, 24(4): 267-283.

[6] 이주호, 김학수. 2009. 의존관계를 이용한 단일문서의 키워드 추출.『2009 한국컴퓨터종합학술대회 논문집』, 36(1): 293-296.

[7] 정영미. 2005.『정보검색연구』. 서울: 구미무역.

[8] 정영미. 1993.『정보검색론』. 서울: 구미무역.

[9] 한승희, 정영미. 2004. 클러스터링 기법을 이용한 개별문서의 지식구조 자동 생성에 관한 연구.『정보관리학회지』, 21(3): 251-267.

[10] Al-Khalifa, Hend S., & Hugh C. Davis. 2006. "Folksonomies versus automatic keyword extraction: an empirical study." *Proceedings of IADIS Web Applications and Research*, 2: 132-143.

[11] Callan, James P. 1994. "Passage-level evidence on document retrieval." *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 302-310.

[12] Dagan, Ido, Lillian Lee, & Fernando Pereira. 1999. "Similarity-based models of cooccurrence probabilities." *Machine Learning*, 34(1-3): 43-69.

[13] Hulth, A., Jussi Karlgren, Anna Jonsson, Henrik Bostrom, & Lars Asker. 2010. "Automatic Keyword Extraction Using Domain Knowledge." *Lecture Notes in Computer Science*, 2004/2010: 472-482.

[14] Kullback, Solomon. 1968. *Information Theory and Statistics*, 2nd ed. New York: Dover Books.

[15] Lee, Lillan. 1999. "Measures of distributional similarity." *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics*, 25-32.

[16] Leweis, David D., & W. Bruce Croft. 1990. "Term clustering of syntactic phrases." *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 385-404.

[17] Lin, J. 1991. "Divergence measures based on the Shannon entropy." *IEEE Transactions on Information Theory*, 37(1): 145-151.

[18] Liu, M., Li, W., Wu Mingli, & Qin Lu. 2007. "Extractive summarization based on event term clustering." *Proceedings of the ACL 2007*, 185-188.

[19] Matzuo, Y., & M. Ishizuka. 2004. "Keyword extraction from a single document using word

co-occurrence statistical information." *International Journal on artificial Intelligence Tool*, 13(1): 157-169.

[20] Pereira, F., Naftali Tishby, & Lillian Lee. 1993. "Distributional clustering of English words." *Proceedings of the 31st Annual Meeting of the ACL*, 183-190.

[21] Plas, L. van der, V. Pallotta, M. Rajman, & H. Ghorbel. 2004. "Automatic keyword extraction from spoken text." *Proceedings of the 4th International Conference on Language Resources and Evaluation 2004*, 2205-2208.

[22] Sneath, P. H. A., and R. R. Sokal. 1973. *Numerical Taxonomy*. SF: Freeman.

[23] Sparck Jones, K. 1971. *Automatic Keyword Classification for Information Retrieval*. London: Butterworth&Co.

[24] Sparck Jones, K. 1972. "Automatic indexing." *Journal of Documentation*, 30(4): 393-432.

[25] Strehl, Alexander, Joydeep Ghosh, & Raymond Mooney. 2000. "Impact of similarity measures on web-page clustering." *Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search(AAAI 2000)*, 58-64.

[26] Suzuki, Y., F. Fukumoto, Y. Sekiguchi. 1998. "Keyword extraction of radio news using term weighting with an encyclopedia and newspaper articles." *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 373-374.

[27] Tombros, Anastasios. 2002. *The Effects of Query-based Hierarchical Clustering of Documents for Information Retrieval*. Ph.D. diss., Cornell University.

[28] Turney, Peter D. 2000. "Learning algorithm for keyphrase extraction." *Information Retrieval*, 2(4): 303-36.

[29] Weeds, J. E. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph. D. diss., University of Sussex.

[30] White, H. D., & B. C. Griffith. 1981. "Author cocitation: a literature measure of intellectual structure." *Journal of the American Society for Information Science*, 32: 163-171.

[31] Witten, Ian H., Paynter, Gordon W., Frank, Eibe., Gutwin, Carl., & Nevill-Manning, Craig G. 1999. "KEA: practical automatic keyphrase extraction." *Proceedings of the 4th ACM Conference on Digital Library*, 254-255.

[32] Zobel, J., A. Moffat, R. Wilkinson, & R. Sacks-Davis. 1995. "Efficient Retrieval of Partial Documents." *Information Processing and Management*, 31(3): 36-377.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Su-Yeon Kim, & Young-Mee Chung. 2006. "An Experimental Study on Selecting Association Terms Using Text Mining Techniques." *Journal of the Korea Society for Information Management*, 23(3): 147-165.

[2] Eun-Gyoung Seo. 1984. "A Study on Automatic Keyword Classification." *Journal of the Korea Society for Information Management*, 1(1): 78-99.

[3] Sarah Yoo. 1999. *Jeongbohakyeonguwa Bunseokbangbeopron*. Seoul: Nanamchulpan.

[4] Sungjick Lee, & Han-joon Kim. 2009. "Keyword Extraction from News Corpus using Modified TF-IDF." *The Journal of Society for e-Business Studies*, 14(4): 59-73.

[5] Jae-Yun Lee. 2007. "Improving the Performance of Document Clustering with Distributional Similarities." *Journal of the Korea Society for Information Management*, 24(4): 267-283.

[6] Jooho Lee, & Harksoo Kim. 2009. "Keyword Extraction of Single Document using Dependency relation." *2009 Proceedings of KIISE*, 36(1): 293-296.

[7] Young-Mee Chung. 2005. *Jeongbogeomseakyeongu*. Seoul: kumimuyeok.

[8] Young-Mee Chung. 1993. *Jeongbogeomseakron*. Seoul: kumimuyeok.

[9] Seung-Hee Han, & Young-Mee Chung. 2004. "Automatic Generation of the Local Level Knowledge Structure of a Single Document Using Clustering Methods." *Journal of the Korea Society for Information Management*, 21(3): 251-267.