http://dx.doi.org/10.3745/KIPSTB.2012.19B.1.053

문서 군집화의 정확률 향상을 위한 범용어 수집과 문서 재분류 알고리즘

신 준 철[†] · 옥 철 영[††] · 이 응 봉[†††]

요 약

정보검색에서 많은 검색 결과 문서들을 효율적으로 다루기 위해 군집화 기술을 사용하고 있지만, 대체로 군집화의 정확률은 일부 영역에서만 요구 사항을 만족시키고 있다. 본 논문에서는 검색 결과 문서들의 군집화 정확률을 향상시키기 위한 두 가지 방법을 제안한다.

첫째는 군집화 과정에서 흔히 쓰이지만 낮은 가중치를 가진 범용어를 정의하고, 검색 결과들을 비교하여 범용어를 자동 수집하고 그의 가중치를 계산하는 방법을 제안한다. 실험 결과 불용어에 비해 범용어를 사용했을 때 군집화 오류의 34%가 개선되었다.

둘째는 집단평균연결 방식의 군집화 알고리즘으로 일차 군집들을 생성 후, 문서와 군집 간의 유사도를 측정하여 가장 유사도가 높은 군집으로 문서를 재분류하는 알고리즘을 제안한다. 네이버 지식인 카테고리를 이용한 군집 결과의 비교 실험을 통해 일차 군집보다 재분류된 군집의 정확률이 1.81% 향상되는 것을 확인하였다.

키워드 : 웹 검색 결과 군집화, 분류, 재분류, 불용어, 범용어, 군집빈도, 점진적 군집화

Gathering Common-word and Document Reclassification to improve Accuracy of Document Clustering

Joon-Choul Shin[†] · Cheol-Young Ock[††] · Eung-Bong Lee[†††]

ABSTRACT

Clustering technology is used to deal efficiently with many searched documents in information retrieval system. But the accuracy of the clustering is satisfied to the requirement of only some domains. This paper proposes two methods to increase accuracy of the clustering.

We define a common-word, that is frequently used but has low weight during clustering. We propose the method that automatically gathers the common-word and calculates its weight from the searched documents. From the experiments, the clustering error rates using the common-word is reduced to 34% compared with clustering using a stop-word.

After generating first clusters using average link clustering from the searched documents, we propose the algorithm that reevaluates the similarity between document and clusters and reclassifies the document into more similar clusters. From the experiments using Naver JiSikIn category, the accuracy of reclassified clusters is increased to 1.81% compared with first clusters without reclassification.

Keywords : Web Searching Results Clustering, Classification, Reclassification, Stop-Word, Common-Word, Cluster Frequency, Incremental Clustering

1. 서 론

수많은 자료를 효율적으로 다루기 위해 정보처리기술이 발달하고 있으며 대표적으로 인터넷 검색 서비스가 있다. 검색 결과가 너무 많아 한 번에 보기가 어려울 경우, 현재의 검색시스템은 검색 결과의 유사도나 인기도 순으로 자료를 나열하여 사용자의 검색을 돕는다. 또한 일부 검색시스템들은 군집화 기술을 사용하여 유사한 자료들을 묶어서 보여줌으로써 사용자의 편의성을 높이고 있다[1, 2]. 이 군집화 기술의 정확률이 향상된다면 더 편리하게 검색에 이용될 수 있을 것이며, 더 많은 영역에서 군집화를 사용할 수 있을 것이다.

일반적으로 검색 결과 문서들에 대한 군집화의 경우, 검색 결과의 문서 크기가 작고 수가 적기 때문에 계층적 군집

※ 이 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2010-32A-H00006).
† 정 회 원 : 울산대학교 컴퓨터정보통신공학과 박사과정
†† 종신회원 : 울산대학교 컴퓨터정보통신공학과 교수(교신저자)
††† 정 회 원 : 충남대학교 문헌정보학과 교수
논문접수 : 2011년 8월 5일
수 정 일 : 1차 2011년 11월 30일
심사완료 : 2011년 12월 7일

화가 적합하며, 그 중에서도 특히 속도와 정확률 면에서 집단평균연결(Average-Link) 군집화가 가장 적합한 것으로 알려져 있다[12].

문서 군집화 과정에서 두 문서의 유사도나, 문서와 군집 사이의 유사도를 구하기 위해 두 개체가 공통적으로 가지는 단어를 이용한다. 검색 결과를 문서 군집화하는 측면에서 볼 때 여러 문서에서 반복적으로 나타나는 '불용어(stop-word)'들을 유사도 계산에서 제거할 필요가 있다. 그러나 색인화 과정에서 불용어를 제거하는 검색시스템과는 달리 군집화에서는 이런 불용어에 낮은 가중치를 부여하여 유사도 계산에 사용하는 것이 유용할 수 있다[4]. 본 논문에서는 군집화 과정에서 흔히 쓰이면서 가치가 낮은 단어를 '범용어(common-word)'로 정의하고, 여러 검색 결과들을 비교하여 범용어를 자동으로 수집하는 새로운 방법을 제안한다.

계층적 군집화 과정에서는 문서와 문서 사이의 유사도나 아직 완성되지 않은 군집과의 유사도를 통해 군집을 구성해 나간다. 이 과정에서 특정 군집에 문서를 포함시키는 것이 유사도 계산에 의한 최상의 결정이더라도, 군집화가 계속 진행되면서 군집들의 성격이 변한다면, 형성된 군집 내의 문서들은 오류일 가능성이 있다. 이러한 오류들은 계층적 군집에서 상위 군집 내의 문서들에서 발견된다. 또한 군집화 과정에서 미분류 문서를 남기기도 한다.

이러한 오류 및 미분류 문서들은 최초 형성된 군집을 대상으로 군집과 문서 간의 유사도를 재계산하여 가장 유사한 군집으로 재분류함으로써 해결할 수 있다. 본 논문에서는 재분류 가중치로 단어의 역군집빈도(ICF, Inverse Cluster Frequency)[5, 6]를 사용하여 재분류 유사도 함수를 제안한다. 이 방법은 군집화 알고리즘과는 완전히 독립되며, 문서(단말 노드)간의 유사도를 구할 수 있다면 적용가능하다.

본 논문의 2장에서는 관련 연구를 살펴보고, 3장에서 범용어를 수집하고 가중치를 계산하는 방법을, 4장에서는 범용어를 사용하기 위해 선택한 군집화 알고리즘을 살펴본다. 그리고 5장에서 군집화 이후의 재분류 가중치 및 유사도 함수를 이용한 알고리즘을 설명하고, 6장에서 범용어와 재분류를 적용한 실험 결과를 분석한다. 마지막으로 7장에서 결론 및 향후 연구 방안을 논의한다.

2. 관련 연구

문서 군집화에 대한 연구는 다양한 목적과 관점에 따라 여러 가지 방법으로 진행되어 왔다. 특히 컴퓨터와 인터넷이 발전하고 웹 문서가 많아지면서 다양한 웹 문서 군집화가 연구되었다[7, 13, 16]. 문서 군집화에는 다양한 알고리즘이 존재하며, 크게는 계층적/비계층적 알고리즘으로 나뉘고 각자의 장단점이 알려져 있다[9]. 비록 알고리즘들의 방법은 다르지만 대부분 문서간의 공통된 단어를 찾으며 단어의 가중치를 계산하여 적용한다는 공통점이 있다.

빠르고 정확한 문서 군집화를 위해 가치가 없는 불용어에 대한 연구가 있었고[3, 10], 불용어가 군집화의 성능에 좋다는 실험 결과가 있다[13]. 불용어와는 반대로 일부 중요 단어(자질어)만을 수집하는 연구도 있었다[11]. 그리고 불용어에도 낮은 가중치를 부여하여 정확률을 향상시킨 연구가 있었다[4]. 본 논문에서도 불용어 대신에 낮은 가중치를 부여하는 단어(이하 범용어)를 사용하여 웹 검색 결과 군집화에 사용하는 방법을 제안한다.

검색 시스템에서 무작위로 선택한 단어로 검색한 결과 문서들을 분석하여 불용어를 수집하는 연구가 있었다[14]. 이 방법은 검색 시스템에서 사용하기 위한 것이며, 본 논문에서는 웹 검색 결과 군집화(WSRC, Web Searching Results Clustering)에 적합한 새로운 수집 방법을 제안한다. 본 논문이 제안하는 방법과의 가장 큰 차이점은 여러 개의 검색 결과 문서 집단에서 공통으로 나타나는 단어를 범용어로 수집한다는 것이다. 이러한 단어가 군집화에서도 서로 다른 군집 간에 동시에 존재할 확률이 높을 것이기 때문이다.

문서 군집화에서 단어의 가중치는 다양한 형태로 사용되며 정확률에 큰 영향을 줄 수 있다. 전통적으로 단어빈도(TF, Term Frequency)와 문서빈도(DF, Document Frequency)를 사용하고 있으며[5], 각각 상황과 목적에 맞추어 다양한 방법으로 응용하고 있다. 본 논문에서도 범용어 가중치를 계산하기 위해 WSRC 환경을 고려한 방법을 제안한다.

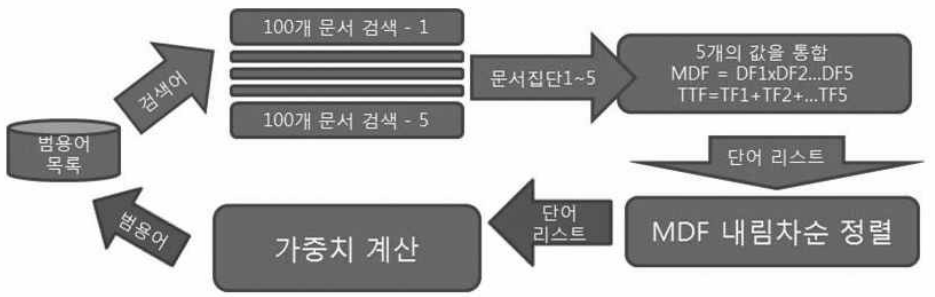
군집이 형성된 후에도, 사용자마다 차별되는 요구와 필요에 따라 군집 구조가 변하거나 문서가 재배치되는 동적 재분류 방법이 있다[8]. 또한 군집들의 개수나 구조의 변화 없이, 문서의 배치만을 바꾸어 더 높은 정확률을 위해 계속해서 반복 재분류하는 방법도 연구되었다[7].

대량의 문서를 빠르게 군집화하기 위한 점진적 군집화(Incremental Clustering)가 연구되었다[15]. 점진적 군집화는 미리 군집화된 정보를 기반으로 나머지 문서들을 선형적으로 처리하는 것으로 속도가 빠른 장점이 있다. 또한 추가적으로 발생하는 문서들을 처리할 수 있어 WSRC에서 계속되어 검색되는 문서들을 추가적으로 군집화하기에 적합하다[16].

본 논문에서는 군집화된 정보를 기반으로 하는 점진적 군집화의 특징을 응용한 새로운 재분류 알고리즘을 제안한다. 점진적 군집화의 대상이 이미 군집화된 문서가 된다는 점을 고려하여, 문서와 군집간의 유사도 함수를 제안하였으며 문서와 그 문서를 포함한 군집 간의 유사도에서 예외 처리를 추가하였다.

3. 범용어

대부분의 문서 군집화는 두 문서의 유사도나, 문서와 카테고리 사이의 유사도를 구하기 위해 두 개체가 공통적으로 가지는 단어를 이용한다. 그러나 만약 그 단어가 여러 주제의 문서들 사이에서 쉽게 나타나는 범용어라면 낮은 가중치를 줄 수 있다. 본 논문에서는 범용어를 수집하고 가중치를 계산하는 새로운 방법을 제안한다.



(그림 1) 범용어 자동 수집 알고리즘

3.1 범용어 수집과 가중치 계산법

범용어는 불용어와 마찬가지로 그 수가 많지 않기 때문에 목록을 사람이 작성할 수도 있으나 컴퓨터를 이용해 수집한다면 더 빠르고 정확하게 목록을 작성할 수 있다. 본 논문에서는 컴퓨터로 자동 수집한 뒤에 사람이 최종 편집하는 방법을 제안한다.

만약 범용어로 검색을 한다면, 검색 결과로 나타나는 문서들은 단지 범용어를 포함한다는 이유로 검색되어질 것이기 때문에 서로 연관이 없는 문서들이 검색될 것이다. 더욱 더 연관이 없는 문서들을 얻기 위해서 다른 범용어로 검색한 문서들을 포함할 수 있다. 여기서 하나의 범용어로 검색된 상위 D개의 문서들을 하나의 문서집단이라고 정의한다.

만약 두개의 문서집단에서 모두 존재하는 단어라면 범용성이 있다고 볼 수 있다. 이것을 확장하여 여러 개의 문서집단에서 동시에 나타난다면 범용성이 더 높다고 할 수 있다. 이때 문서집단의 개수와 검색에 사용된 범용어의 개수를 W라고 정의한다.

자동 수집 알고리즘은 다음과 같다. 처음에 W개의 범용어를 사람이 직접 정하여 검색하면 W개의 문서집단이 수집된다. 여기서 수집된 단어가 범용성이 있는지 계산하기 위해 문서집단별로 DF를 구한다. 즉, 한 단어는 W개의 DF를 가지게 된다. 만약 한 단어의 모든 DF가 1이상이라면 범용어로 수집한다. (그림 1)에 이 과정이 나타나며 D는 100, W는 5를 적용한 상태다. 따라서 하나의 문서 집단은 100개의 문서를 포함하고, 총 5개의 문서 집단이 존재하며, 한 단어는 5개의 DF를 가진다. 한 단어가 범용어가 되기 위해서는 5개의 집단에서 모두 등장해야한다. 1개 이상의 집단에서 등장하지 않고 다른 집단에서 등장하는 단어는 집단들을 구분 지을 수 있다는 의미이므로 군집화에 중요한 것으로 간주하여 범용어로 수집하지 않는다.

가중치를 계산하기 위해 W개의 DF를 모두 곱한 값을 문서빈도총곱(MDF, Multiplication of DFs)이라고 하고, TF의 총합은 단어빈도총합(TTF, Total of TFs)이라고 정의한다. 일반 단어의 가중치를 1로 정의하고 범용어의 가중치는 1보다 작고 0보다 크게 계산한다. MDF가 높은 단어는 범용성이 높다는 것으로 가중치가 0에 가깝고, MDF가 0에 가까울수록 가중치는 1에 가까워야 한다. MDF가 0인 경우는 가중치가 1이 되며 이는 범용어로 수집할 필요가 없음을 뜻한다. 따라서 하나 이상의 DF가 0인 경우 범용어로 수집하지 않는다. 이렇게 가중치를 계산하는 과정은 다음 <의사코드>와 같다. <의사코드>에서 word.weight가 가중치다. 만약에 이전에 수집한 범용어가 또 수집된다면, 가중치는 그 둘의 중간 값을 사용한다.

최초 W개의 범용어로 다른 범용어를 수집한 뒤에, 반복수집하기 위하여 현재까지 수집한 범용어 중에서 무작위로 W개의 범용어를 새로 정하여 전 과정을 반복한다. 계속 반복하면 매번 새로 수집되는 범용어의 수가 줄어들게 되며, 결국 새로 수집되는 범용어가 없게 되면 중단한다.

이렇게 수집한 범용어 목록을 사람이 직접 보고 판단하여 다시 수집 알고리즘을 반복하거나 편집을 한다.

〈의사코드〉 범용어 가중치 계산 방법

```
Function WeightCommonWord( WordList )
    WordList.SortByMDF //MDF내림차순 정렬
    TotalTTF = 0//모든 단어의 TTF 총합
    for word In WordList
        TotalTTF = TotalTTF + word.TTF
    End for
    CurrentTTF = 0 //지나간 상위 범용어들의 TTF 합
    for word In WordList
        if( word.MDF == 0 ) break
        CurrentTTF = CurrentTTF + word.TTF
        word.weight=(CurrentTTF / TotalTTF)
        if( word.weight < Minimum ) //최소값
            word.weight = Minimum
        End for
    return WordList
End Function
```

3.2 수집된 범용어와 가중치

본 논문에서는 몇 차례의 실험을 거쳐 경험적으로 W를 5로하고 D를 100으로 설정하였다. 검색 도메인은 네이버 지식인이며 최초의 범용어로 {그, 나는, 또, 그래서, 지식인}을 선택하였다. 여기서 '지식인'은 이 도메인에서 자주 나타나는 단어다.

범용어를 수집할 때 문서나 어절에서 단어를 추출하는 방법은 군집화에서 사용하는 방법과 동일해야 한다. 본 논문

에서 실험으로 사용하는 군집화 알고리즘은 간단하게 어절에서 조사/어미를 제거하고 남은 어절에서 첫 1글자, 1~2글자, 1~3글자를 각각 수집하기 때문에 범용어 수집에서도 같은 방법을 사용하였다. 이런 수집 방법에 따른 결과의 일부가 <표 1>에 나타나 있다.

<표 1> 수집된 범용어의 가중치별 예

가중치	단어
0.05≦w≦0.10	있는, 그리, 하는, 제가, 없, 알려
0.10<w≦0.15	지금, 부탁드, 합니, 들어, 하고, 다른
0.15<w≦0.20	좋은, 없어, 이렇, 것이, 대한, 있으
0.20<w≦0.25	있었, 먼저, 좋을, 지식, 그거, 등급
0.25<w≦0.30	예저, ㅡㅡ, 여러분, 좋습, 외국, 레이
0.30<w≦0.35	법, 분명, 뒤, 아무튼, 옷, 답답, 수업
0.35<w≦0.40	권

<표 1>에서 가중치가 낮은 단어들은 직관적으로 범용성이 높음을 알 수 있으며, 높아질수록 비교적 가치 있는 단어임을 쉽게 알 수 있다. 'ㅡㅡ'처럼 별 의미 없이 남용되는 이모티콘도 수집되는 것으로 보아 흔한 은어나 인터넷 채팅 용어, 오타 등도 수집된다는 것을 알 수 있다. {지식, 등급, 수업} 등의 단어는 네이버 지식인 도메인의 특성으로 볼 수 있다. 이 외에 '레이'같은 단어는 범용어로 사용하기에는 부적절할 수 있기 때문에 최후 편집에서 제거하거나 가중치를 높일 수 있다.

<표 1>에서 가중치는 0.05에서 0.4까지 존재한다. 최소 0.05가 나온 이유는 <의사코드>에서 Minimum을 사용하였기 때문이며, 최대 가중치가 1이하로 나온 이유는 MDF가 0인 단어가 존재하기 때문이다.

4. 웹 검색 결과 군집화

이 장에서는 본 논문에서 범용어와 재분류를 실험 연구하기 위해 사용한 WSRC를 간략히 설명한다. WSRC란 검색 결과에서 정보를 효과적으로 찾기 위해 결과들을 군집화하는 것이다.

WSRC는 서버보다는 클라이언트에서 수행하기에 적합하며[9], 현실적으로 검색 결과를 다운받는 시간을 고려하면 상위 일부분만을 군집화할 수 있다. 그리고 시간을 더 효율적으로 활용하기 위해 검색 결과로 나타나는 링크와 짧은 설명을 포함하는 토막(Snippet)만을 다루는 것이 좋다. 이런 이유로 WSRC에서는 군집화할 문서의 크기가 작고 수가 적기 때문에 계층적 군집화가 적합하다[12]. 그 중에서도 특히 속도와 정확률 면에서 집단평균연결(Average-Link) 군집화가 가장 적합하다.

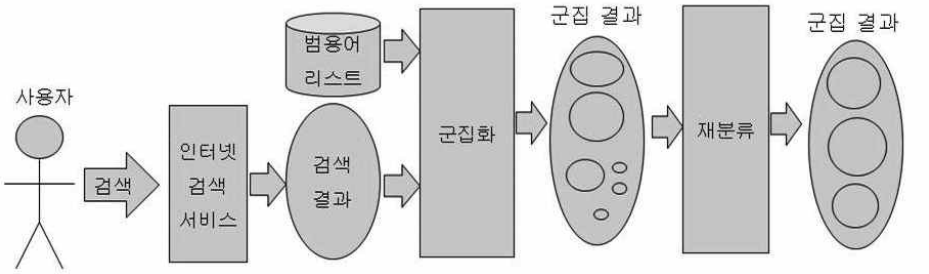
$$sim(a,b) = \sum_{i=1}^{n} w_{ai} \times w_{bi}$$ <수식 1> 내적 유사도

연결 군집화를 위해서는 두 개체 간 유사도를 구하는 유사도 함수가 필요하다. 일반적으로 문서 유사도 함수는 두 문서가 가지는 공통된 단어가 많을수록 그리고 가중치가 높을수록 유사도를 높게 계산하며, 두 문서의 크기로 정규화하는 부분도 포함한다. 그러나 본 논문은 토막의 크기가 대부분 일정하기 때문에 정규화를 하지 않는 내적 유사도<수식 1>를 사용한다.

단어의 가중치를 계산하기 위해서 본 논문에서 제시하는 범용어만을 사용하며, 범용어가 아닌 단어의 가중치는 1이다. <수식 1>에서 a와 b는 문서이며 i는 단어, n은 단어의 개수이다. 그리고 w_{ai}는 단어 i의 범용어 가중치와 문서 a 안에서 단어 i의 빈도수를 곱한 값이다.

연결 군집화는 모든 문서가 하나의 군집이 될 때까지 결합을 진행하거나 적정 시점에서 멈추기 위한 조건에 따라 중단될 수 있다. 본 논문에서는 총 3개의 중단 조건을 사용한다. 첫 번째 조건은 가장 유사한 쌍의 유사도가 일정 유사도 이하인 경우다. 두 번째는 총 군집의 수가 2개만 남은 경우다.

세 번째 중단 조건은 설명을 위해 몇 가지 정의가 필요하다. 일정 크기 이하의 작은 군집은 충분히 결합되지 않은 것으로 간주하여 미분류 군집이라고 하고, 이 군집에 포함된 문서는 미분류 문서라고 정의한다. 반대로 미분류가 아닌 경우 각각 분류 군집, 분류 문서라고 정의한다. 세 번째 중단 조건은 미분류 문서가 일정량 이하이고 분류 군집이 2개이며, 그 2개가 가장 유사한 쌍인 경우다. 따라서 분류 군집의 개수가 1개가 되지 않도록 중단한다.



(그림 2) 전체 시스템 흐름

5. 재분류

문서 군집화 알고리즘은 문서를 잘못 분류하기도 하고, 상황에 따라 미분류 문서를 남기기도 한다. 본 논문에서는 이런 오류들을 찾아 올바른 군집으로 재분류하는 알고리즘을 제안한다. 재분류 알고리즘은 (그림 2)와 같이 범용어를 사용한 군집화 이후에 적용된다. 재분류를 통해 군집 결과에서 아주 작은 군집(미분류 문서)들이 적합한 다른 군집으로 분류되고 군집들이 전체적으로 조금씩 변하게 될 것이다.

5.1 재분류 가중치

군집화가 끝난 후에는 단어를 포함하는 군집의 수를 이용해 가중치를 계산할 수 있다. 만약 어떤 단어가 다른 단어에 비하여 더 많은 군집에서 나타난다면 가중치가 낮아야 하고, 반대로 적은 군집에서 나타난다면 높아야 한다. 이것은 의미 그대로 역군집빈도(ICF)를 뜻하며, 본 논문에서는 재분류 가중치를 계산하기 위해 <수식 2>를 사용하였다.

$$r_i(t+1) = r_i(t) / CF_i(t)$$ <수식 2> 재분류 가중치

$r_i(t)$는 재분류 횟수t에서 단어i의 가중치이다. 최초의 재분류에서 t는 1이며 $r_i(0)$은 군집화에서 사용한 가중치를 그대로 사용할 수 있다. 본 논문에서는 $r_i(0)$으로 범용어 가중치를 사용하였다. 재분류가 반복될수록 t가 증가하며 가중치는 계속 누적되어 높고 낮은 가중치들의 격차가 뚜렷해진다.

5.2 재분류 유사도 함수

어떤 문서가 올바른 군집에 분류되었다면 그 문서와 그 문서를 포함하는 군집과의 유사도가 다른 군집과의 유사도보다 높을 것이다. 만약 아니라면 잘못 분류된 것으로 간주할 수 있고, 가장 유사도가 높은 군집으로 재분류할 수 있다. 미분류 문서도 이와 비슷하게 가장 유사도가 높은 군집으로 분류할 수 있다.

문서와 군집의 유사도를 구하기 위해서, 집단평균연결의 방법을 사용할 수 있다. 그러나 군집화와는 달리 재분류에서는 오직 하나의 문서와 크기가 다양한 군집 사이의 유사도만 계산한다는 차이점이 있다. 하나의 커다란 군집 안의 모든 문서가 비록 공통된 주제를 가질 수는 있어도, 모두 다 공통된 단어를 가지기는 힘들다. 오히려 군집이 작을수록 모든 문서가 공통된 단어를 가지기 쉽다. 재분류 대상인 문서와 아주 유사한 문서들을 포함하는 군집이라도 다른 문서 또한 많이 포함된다면 유사도의 평균값은 낮을 것이다. 따라서 본 논문에서는 재분류라는 상황에 맞추어 평균을 구하는 과정에 조절이 가능한 <수식 3>을 제안한다.

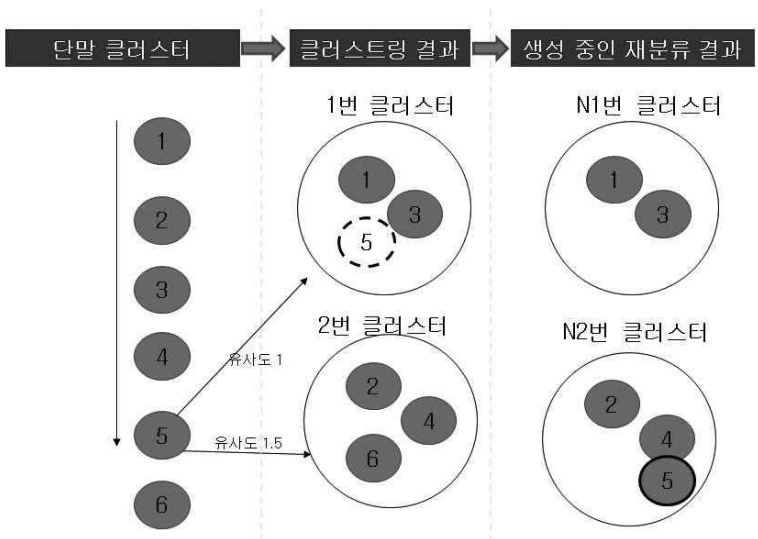
$$rsim(C,a) = \frac{\sum w_{Ci} \times w_{ai}}{|C|^k}$$

<수식 3> 재분류 유사도 함수

<수식 3>에서 a는 문서, C는 군집, |C|는 군집이 포함하는 문서의 수, i는 단어를 의미한다. 그리고 w_{Ci}는 C에서 i의 빈도수와 재분류 가중치를 곱한 값이며, w_{ai}는 a에서 i의 빈도수와 재분류 가중치를 곱한 값이다. 상수 k는 수식을 재분류에 최적화하기 위한 값으로 실험을 통하여 구할 수 있다. 만약 단순한 평균값이 가장 높은 정확률을 보인다면 k는 1이 될 것이다.

5.3 재분류 과정

재분류는 각 문서와 군집간의 유사도를 측정하여 가장 유사한 군집으로 문서를 옮기는 것이다. 단, 문서는 미분류 군집으로는 옮기지 않으며, 자기 자신과의 유사도를 측정하지 않는다. 재분류 진행의 예가 (그림 3)에 나타나 있다.



(그림 3) 재분류 진행 예

(그림 3)에서 1~4번 문서가 재분류 되었고 5번 문서가 재분류 중인 모습이다. 중앙의 1, 2번 군집이 이전 군집화(또는 재분류)로 형성된 것이며, 우측의 N1, N2번 군집이 재분류로 새로 형성 중인 군집이다. 5번 문서는 본래 1번 군집에 포함되어 있으나 자기 자신과의 유사도를 측정하지 않기 위하여 일시적으로 제거되어 점선으로 표시되었다. 결국 5번 문서는 2번 군집과의 유사도가 더 높기 때문에 재분류 결과 N2번 군집으로 분류되었다.

이 알고리즘은 결과에 변화가 없을 때까지 반복이 가능하다. 그러나 현실적으로 시간을 고려해야하기 때문에 고정된 횟수만큼 반복할 것을 제안한다.

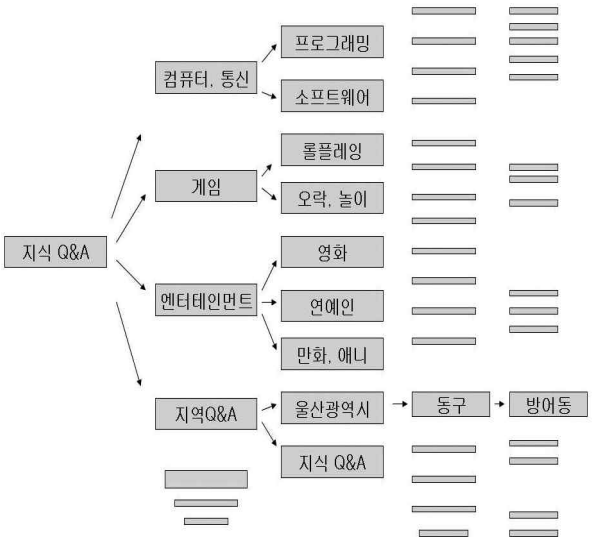
6. 실험 및 평가

실험을 위해서 객관적으로 미리 분류된 정답이 필요하다. 본 논문은 네이버 지식인의 글들을 검색하고, 지식인 카테고리 번호를 통해 정답을 확인하는 실험을 하였다. 비록 군집과 카테고리는 1:1 대응하는 개념은 아니지만 실험 결과의 객관성을 위해 이를 사용하였다.

6.1 네이버 지식인 카테고리

네이버 지식인의 글들은 카테고리 번호를 가진다. 비록 군집화에서 분류하는 것이 카테고리로 분류하는 것과 동일한 기준을 가지지는 않지만 충분히 비교 가능하다. 만약 두 문서의 내용이 크게 달라 다른 군집으로 분류될 정도라면 카테고리도 다를 것이기 때문이다. 반대로 카테고리가 같은 두 문서가 다른 군집에 분류되었다면 둘 중 하나는 오답일 것이다.

이 카테고리는 트리 구조를 가지며, 최대 깊이는 4이며 대체로 3까지 이어진다((그림 4) 참조). 카테고리 연결에서 특정 카테고리를 선택하면 모든 자식 카테고리도 군집으로 연결된 것으로 간주한다.



(그림 4) 네이버 카테고리 구조

카테고리는 사람이 결정하며, 비슷한 글이라도 다양한 관점에서 다른 카테고리를 선택할 수 있다. 따라서 군집 하나가 여러 개의 카테고리를 가질 수 있다. 정확한 실험을 위해 군집-카테고리 연결 정보는 사람이 직접 작성한다. <표 2>는 검색어 '비'에 대한 군집-카테고리 연결의 일부를 나타내고 있다.

〈표 2〉 네이버 카테고리와 군집 연결 일부

군집	카테고리	문서 수
소나기	엔터테인, 예술>꿈, 해몽	14
	가정,생활>주택,인테리어	9
	교육,학문>인문,사회과학>문학	10
	교육,학문>인문,사회과학>우리말	14
	교육, 학문>자연, 공학>지구과학	18
	교육, 학문>자연, 공학>환경	5
	쥬니버>과학	10
	이 외 ...	
가수	엔터테인, 예술>연예인	74
	쥬니버>연예, TV	19
비율	교육, 학문>자연, 공학>수학, 통계	12
	쥬니버>수학	60

6.2 실험 방법

<표 2>처럼 군집-카테고리 연결 정보가 사람에 의해 구축되면, 다음 작업은 실제 군집과의 연결이다. 예를 들어 '비'로 검색하고 군집화하면 실제로 3개의 군집이 형성되지만 어느 것이 '소나기' 군집인지 구분해야 한다.

먼저 '소나기'에 속해야할 모든 카테고리의 문서로부터 단어들을 수집한다. 그리고 이 단어들로 군집과의 내적 유사도 <수식 1>을 계산한다. 가장 유사도가 높은 군집이 '소나기' 군집이 된다. 이 방법은 군집화의 정확률이 일정 수준 이상이 되면 컴퓨터를 통해 자동화 가능하다.

문서는 그 문서를 포함하는 군집이 그 문서의 카테고리와 연결되지 않았다면 오답으로 처리한다. 따라서 미분류 문서도 모두 오답이다.

실험에 사용된 모든 군집화의 중단 조건은 동일하다. 중단 조건에서 사용되는 모든 상수(유사도 제한, 미분류 기준 등)도 동일하게 적용하였다.

6.3 검색어와 검색된 문서들

본 논문에서는 실험을 위해 중의적이거나, 고유 명사로서 의미가 여러 개인 검색어를 20개 선택하였다. 한국어 처리에서 중의성 해소가 힘든 단음절 단어 {비, 배, 차}를 포함하고 고유 명사 {빅뱅, 레오파드} 등을 포함한다.

20개의 단어로 각각 550개의 문서를 검색하였고, 몇몇 검색어는 550개 보다 적은 결과만이 검색되어 총 10,971개의 문서가 검색되었다. 그 중에서 8,092개의 문서가 군집-카테고리 연결 정보에 포함되어 정답세트을 구성한다. 나머지 문서는 카테고리 정보가 모호하거나 잘못된 경우다.

〈표 3〉 불용어와 범용어 비교 실험

임계점	0.250	0.275	0.300	0.325	0.350	0.375	0.400	범용어
정확률	97.7%	97.4%	96.5%	96.5%	96.7%	96.7%	96.7%	98.5%

〈표 4〉 네이버 카테고리 코드와 대응 실험 결과 (단위 %)

상황	재분류 전	재분류 횟수와 정답 문서량의 변화				
		1	2	3	4	5
평균 정확률(%)	96.68	98.34	98.42	98.49	98.49	98.49

6.4 결과와 분석

범용어와 불용어를 비교 실험하기 위해 범용어의 일부를 불용어로 바꾸고 나머지는 가중치 1로 적용하여 정확률을 측정하였다. 불용어를 구분하기 위해서 가중치 임계점을 사용하였으며 이것이 <표 3>에 나타나 있다. 실험에서 나머지 환경은 모두 동일하며 군집화 후 재분류를 5회씩 거친 것이며 재분류 유사도 상수 k는 0.4이다.

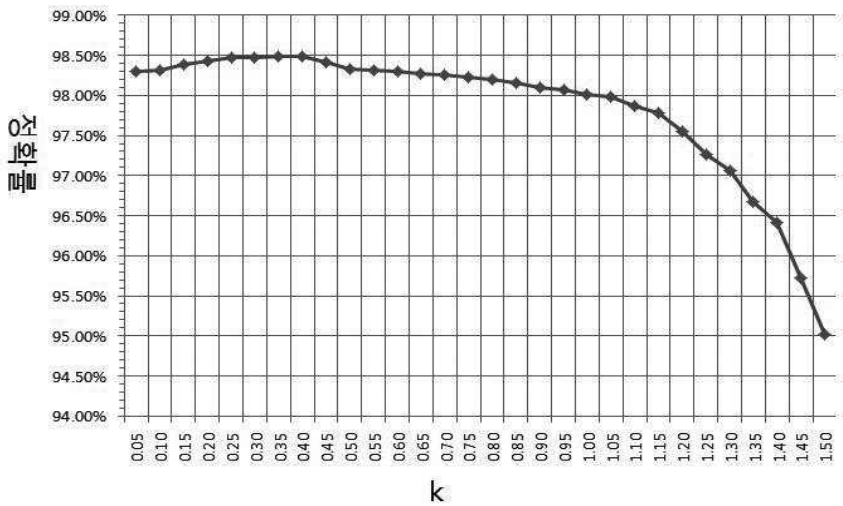
전체적으로 96% 이상의 높은 정확률이 나타났다. 이것은 실험 데이터인 네이버 지식인에서는 이용자들이 비슷한 내용을 반복적으로 올리기 때문에 군집화가 비교적 정확하게 되기 때문이다. 예를 들어 검색어 '비'의 결과에는 가수 '비'의 노래 '레이니즘'의 가사를 요청하는 글이 여러 번 등장한다. 이런 식으로 완전히 같은 내용의 문서가 다량 포함되어 군집화 정확률이 매우 높게 나타났다.

가중치 임계점 0.25보다 낮은 경우에는 자동화된 정확률 측정이 불가능할 정도로 군집화의 정확률이 낮았다. 이것은 즉 0.225~0.25 사이에는 반드시 불용어가 되어야할 단어가 많다는 것이다. 불용어 실험 정확률 중 가장 높은 것은 임계점 0.25인 경우로, 0.25보다 큰 단어에는 불용어로서 사용하기엔 부적합한 단어가 있음을 뜻한다.

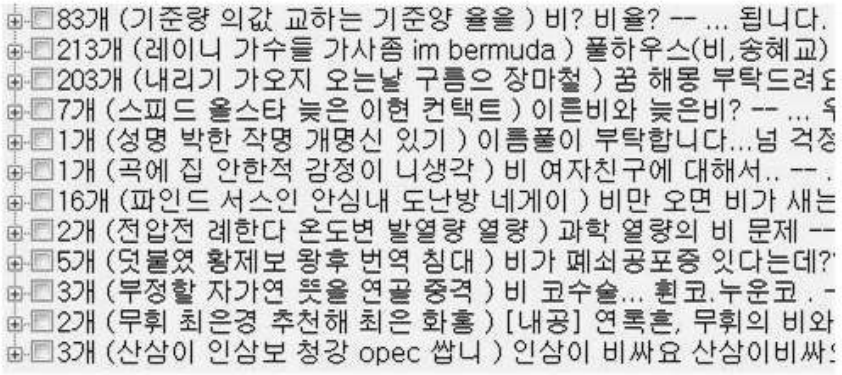
불용어 사용 중 가장 높은 정확률은 97.7%로 범용어를 사용했을 때의 정확률98.5%보다 낮다. 따라서 범용어가 불용어에 비해 정확률 면에서 더 우수한 것으로 분석된다. 비록 개선된 정도는 0.8% 이지만 오류율로 계산할 경우 2.3%에서 1.5%로 줄었고, 이것은 전체 오류 중 34%가 개선되었다는 의미다. 이 개선 비율은 군집화 정확률이 낮게 나오는 검색 결과에서도 크게 변하지 않았다. 예를 들어 군집화 정확률이 가장 낮았던 검색어 '배'의 경우 84.24%에서 95.35%로 11.11%가 증가하였다. 따라서 범용어 적용은 확실히 의미 있는 것으로 분석된다.

재분류 유사도에서 상수 k를 결정하기 위해 k를 바꾸어 가며 실험하였다. 실험에서는 <표 3>에 따라 정확률이 가장 높은 범용어를 사용하였으며, k는 0.05~1.5 까지 변화를 주었다. 실험의 결과가 (그림 5)에 나타나 있다.

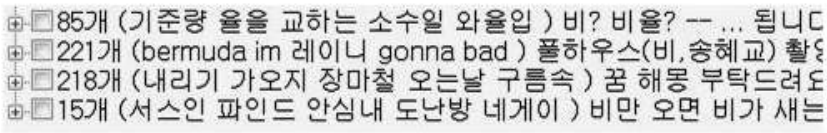
(그림 5)에서 정확률의 변화는 곡선을 그리고 있다. 집단 평균연결과 동일하게 평균값을 측정한 결과는 k가 1인 경우로 정확률 98.0%를 보이며, 가장 높은 정확률 98.5%를 보이는 k값은 0.4이다. 재분류 전 정확률은 96.68%로 <표 4>에 나타나 있으며 k값 1.35부터는 재분류가 정확률을 오히려 낮추는 결과를 보였다.



(그림 5) 재분류 유사도 상수 k의 실험 결과



(그림 6) 검색어 '비'의 군집화 결과



(그림 7) 검색어 '비'의 군집화 후 1차 재분류 결과

재분류의 전과 후에 대한 변화를 측정한 결과가 <표 4>에 나타나 있다. 범용어와 k값 0.4를 사용하여 정확률이 최대가 되도록 하였다. 실험 결과 재분류를 할수록 변화가 줄었으며, 4~5회에는 완전히 변화가 없었다. 정확률은 재분류 전에 96.68%였고 재분류 3회 후 98.49%로 향상되어 전체 오류 중 54.52%가 개선되었다.

검색어 '비'로 550개의 문서를 네이버 지식인에 요청하였고 실제로 539개의 문서가 검색되었다. 그 중에 267개가 <표 2>에 나타나 있는 연결 정보에 포함되어 있으며, 군집화 결과를 (그림 6)에 나타내었다. 각 줄은 군집을 나타내며, 처음에 나타나는 숫자는 문서의 수다. 괄호 안의 단어는 군집의 대표 단어이며, 괄호 우측의 내용은 군집이 포함하는 문서 중 대표를 선택한 것이다. 첫 번째 군집은 <표 2>에 나타나 있는 '비율'에 해당하고, 두 번째와 세 번째는 각각 '가수', '소나기' 군집이다. 그 외의 군집에 포함된 문서들은 군집-카테고리 연결 정보에 있을 경우 모두 오답으로 처리하였다.

(그림 7)은 재분류를 1회 거친 후의 상태다. 군집에 포함된 문서가 10개 이하인 경우 미분류로 취급되어 재분류를 통해 다른 군집으로 분류되었다. 예를 들어 6번째 군집 "비 여자친구에 대해서"는 2번 군집으로 올바르게 분류되었다. 그 외에 몇몇 잘못 분류된 문서도 올바르게 분류되어 정확률이 향상되었다. 그림에서 직접 보이지는 않지만 '소나기' 군집에 있던 문서 "비 5집 성공 가능성은?"문서가 1차 재분류 후 '가수' 군집으로 분류되었다.

7. 결론 및 향후 연구 방안

본 논문은 문서 군집화의 정확률을 향상시키기 위해서 범용어 수집과 재분류 알고리즘을 제안하였다. 두 방법을 실험하기 위해 네이버 지식인 검색 결과를 대상으로 "내적 유사도 함수"와 "집단 평균 연결 군집화"을 사용하였으며, 정확률을 객관적으로 측정하기 위해 네이버 지식인 카테고리를 사용하였다.

본 논문이 제안하는 군집화용 범용어를 수집하는 방법은 검색 시스템용 불용어 수집 방법[14]과 달리 여러 문서 집단에 공통으로 존재하는 단어를 찾는다. 이렇게 수집한 범용어를 비교 실험하기 위해 범용어로 수집된 단어들을 불용어로 사용하여 군집화하였다. 실험한 결과 중에서는 범용어에서 가중치가 일정 수준 이하인 단어만 불용어로 사용하는 것이 가장 정확률이 높았으나 범용어를 사용한 정확률 보다는 낮았다.

범용어 자동 수집 알고리즘에서 무작위로 선택하는 단어의 개수와, 한 번에 검색하는 문서의 개수는 경험적으로 선택했을 뿐이며, 수집되는 범용어와 계산되는 그 가중치는 무작위로 검색어를 선택하기 때문에 매번 변할 수 있다는 문제점이 있다. 또한 알고리즘의 중단 조건은 무작위로 선택되는 검색어에 의해 지나치게 빠르게 수집을 중단시킬 수 있다. 따라서 범용어 수집은 앞으로 더 연구해야할 부분이 많다.

군집화가 끝난 후 ICF를 단어 가중치에 추가하고, 문서와 군집 사이의 유사도를 새로 계산하여 재분류한 결과 96.68%에서 98.49%로 정확률이 향상되어 전체 오류 중 54.52%가 개선되었다. 대부분의 미분류 문서가 재분류에 의해 가장 유사한 군집으로 분류된 것이 크게 작용한 것으로 분석된다. 퍼지 이론을 이용한 동적 재분류[8]와 실험 환경은 다르지만 본 논문이 제안하는 재분류는 개선 비율이 높고 안정적으로 작동하여 사용자의 의사 결정 없이 항

상 적용할 수 있다. 그리고 하나의 문서는 하나의 군집에만 포함되게 하며 알고리즘의 복잡도는 문서 수에 선형이라는 차이점도 있다.

본 논문이 제안하는 재분류 방법은 하나의 문서를 미리 분류된 기존 군집 중 어디에 포함시킬지를 결정하는 부분에서 점진적 군집화와 유사점이 있다. 점진적 군집화는 대량의 문서를 빠르게 처리 가능하고[15] 새로 추가되는 문서들도 빠르게 분류하는 장점[16]이 있기 때문에 점진적 군집화에 적용할 방법을 연구하고 비교해 볼 필요가 있다.

이번 실험 분석을 통해 재분류는 미분류가 발생하여 재현율이 떨어지는 군집화에 적합할 것으로 예상된다. 만약 재분류의 특성을 더 파악한다면 다른 군집화에 대해서도 이것을 적용하기에 적합한지 미리 알 수 있고 정확률 향상 정도를 예측할 수 있을 것이다. 따라서 재분류를 다양한 군집화에 적용하여 결과를 분석하는 실험 연구가 필요하다.

참 고 문 헌

[1] 네이버 뉴스 클러스터링, http://news.search.naver.com/newscluster/

[2] Carrot2 Clustering Engine. http://search.carrot2.org

[3] 김판구 외. 한국어 정보 검색을 위한 불용어의 구성 및 적용, 한국정보과학회 봄 학술발표논문집 제20권 제1호, pp.809-812, 1993.

[4] 권호경 외. 통계정보를 이용한 가중치 부여 불용어 사전의 구성, 한국정보과학회 봄 학술발표논문집 제23권 제1호(A), pp.903-906, 1996.

[5] 김영수 외. 등급에 따른 웹 유해 문서 분류 기술. 한국정보처리학회, 13C(7): pp.859-864, 2006.

[6] 정하용 외. 특허 분류를 위한 효과적인 자질 선택, 한국정보과학회 가을 학술발표 문집(II)제32 제2호, pp.670-672, 2005.

[7] 이문기 외. 웹 디렉토리 서비스를 위한 문서 클러스터링, 한국정보과학회 봄 학술발표논문집 제27권 제1호(B), pp.351-353, 2000.

[8] 박선 외. 비음수 행렬 분해와 동적 분류 체계를 사용한 자동 이메일 다원 분류, 한국정보과학회논문지, 37(5): pp.347-417, 2010.

[9] 황태호 외. 점진적 알고리즘을 이용한 웹 문서 클러스터링 시스템의 설계 및 구현, 한국정보과학회 가을 학술발표논문집, 26(2-1): pp.207-209, 1999.

[10] 주길홍 외. 효율적인 문서검색을 위한 레벨별 불용어 제거에 기반한 문서 클러스터링, 컴퓨터교육학회논문지, 1(3): pp.67-80, 2008.

[11] 윤보현 외. 자동 문서 클러스터링을 위한 디스크립터 추출 방안, 정보처리학회 춘계학술대회 논문집, pp.230-233, 2000.

[12] 윤보현 외. 검색결과의 브라우징을 위한 계층적 클러스터링, 한국정보과학회 봄 학술발표논문집, 27(1): pp.342-344, 2000.

[13] Mark Sinka, David Corne. A Large Benchmark Dataset for Web Document Clustering. Soft Computing Systems: Design, Management and Applications. Volume 87 of Frontiers in Artificial Intelligence and Applications. 2002.

[14] Lo, Rachel Tsz-Wai 외. Automatically building a stopword list for an information retrieval system, Journal of Digital Information Management, 3(1). 2005.

[15] Fazli Can, Edward A. Fox, Cory D. Snavely and Robert K. France. Incremental clustering for very large document databases: Initial MARIAN Experience, Information Sciences Vol.84, Issues1-2, pp.101-114, May, 1995.

[16] Oren Zamir and Oren Etzioni. Web document clustering: a feasibility demonstration. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp.46-54, 1998.

신 준 철

e-mail : ducksjc@hotmail.com

2007년 울산대학교 컴퓨터정보통신공학과 (학사)

2009년 울산대학교 컴퓨터정보통신공학과 (석사)

2011년 울산대학교 컴퓨터정보통신공학과 박사과정 수료

2011년~현 재 울산대학교 지능형컴퓨터 연구실 연구원

관심분야 : 한국어정보처리, 문서 군집화, 소프트웨어 공학

옥 철 영

e-mail : okcy@ulsan.ac.kr

1982년 서울대학교 컴퓨터공학과(학사)

1984년 서울대학교 컴퓨터공학과(석사)

1993년 서울대학교 컴퓨터공학과(박사)

1994년 러시아 TOMSK 공과대학 교환교수

1996년 영국 GLASGOW 대학교 객원교수

2007년~2008년 한국정보과학회 언어공학연구회 위원장

2007년 몽골국립대학교 IT대학 명예박사학위

2008년 국립국어원 객원연구원

1984년~현 재 울산대학교 컴퓨터정보통신공학부 교수

관심분야 : 한국어정보처리, 의미분별, 온톨로지, 지식베이스, 기계학습, 문서분류

이 응 봉

e-mail : eblee@cnu.ac.kr

1985년 성균관대학교 도서관학과(학사)

1992년 성균관대학교 문헌정보학과(석사)

1996년 성균관대학교 문헌정보학과 (정보학 박사)

2007년 미국 Univ. of Illinois at Urbana-Champaign 교환교수

2009년～현 재 한국문헌정보학회 부회장

2010년～현 재 충남대학교 사회과학연구소장

2010년～현 재 한국연구재단 자문위원

1996년～현 재 충남대학교 문헌정보학과 교수

관심분야 : 정보검색, 정보서비스, 데이터베이스 품질평가, 정보시스템 및 정보정책