계층적 클러스터링과 문장 적합성 피드백을 이용한 상담사례 검색 시스템

김 승 일, 곽 희 규, 김 수 형
전남대학교 전산학과

Counseling Case Retrieval System Using Hierarchical Clustering and Sentence Relevance Feedback

Seung-Il Kim, Hee-Kue Kwag, Soo-Hyung Kim
Dept. of Computer Science, Chonnam National University

요 약

본 논문에서는 카운셀링을 원하는 사용자가 카운셀러와 전자메일을 통해 상담을 원할 때 사용자의 상담 내용에 근거하여 유사한 사례를 검색해 주는 시스템을 제안한다. 제안방법은 문서의 계층적 클러스터링과 용어 적합성 피드백을 상담 사례 검색 시스템에 적용시켜, 상담사례에 나타나는 단어의 출현 빈도와 유사도를 통해 트리 구조를 형성하고, 이 트리 구조를 통한 하향 탐색을 수행한다. 하향 탐색을 하는 도중 노드의 매칭함수의 값이 서로 유사하여 노드 선택이 어려울 경우, 사용자에게 질의를 통해 용어를 제시하고, 사용자의 피드백을 통해 질의로 입력된 사연 내용의 가중치를 개선하여 내용에 가장 부합되는 문서를 탐색한다.

1. 서론

정보통신의 발달로 상담을 원하는 사용자가 카운셀러와 직접 만나지 않고 통신을 통해 전자 메일로 상담하는 서비스가 제공되고 있다. 이 서비스를 통해서 사용자와 카운셀러간의 상담 내용이 저장된다. 이러한 정보는 상담을 원하는 다른 사용자에게 유용한 사례로 제시되기에 이를 조직적으로 저장하고 검색할 수 있는 시스템이 필요하다. 정보 검색 시스템 (Information Retrieval System)은 사용자가 필요로 하는 정보를 수집하여 내용을 분석한 뒤, 찾기 쉬운 형태로 저장해 두었다가 정보의 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템이다. 그러나 이러한 정보 검색 시스템은 100%의 정확도와 재현율은 나타내지 못한다. 즉 사용자는 검색을 통해 찾고자 하는 적합한 문서를 얻을 수 있지만 적합한 모든 문서를 얻을 수는 없다. 이것은 시스템이 갖고 있는 문서의 색인이 사용자가 쓰는 용어와 다를 수 있고, 사용자는 시스템이 가지고 있는 문서 구성에 대해 알 수 없으므로 검색하기에 가장 적합한 질의를 작성할 수 없다는데 근거한다. 이러한 문제점을 해결하기 위해서 시스템의 도움을 받아 질의를 수정하는 사용자 피드백 방법이 이용되고 있다[1].

제안 방법에서는 사용자에게 카운셀러의 답장을 제공하기 전 사연과 가장 유사한 사례들을 보여주는 상담 사례 검색 시스템을 제시한다. 제안 방법에서는 문서들을 색인어에 의한 용어 벡터로 표현하였고, 검색 효율을 높이고 사례들에 대한 구조적 정보를 얻기 위해 계층적 클러스터링 방법을 사용한다. 본 논문의 2장에서는 계층적 클러스터링과 용어 적합성 피드백, 3장에서는 시스템 설계 및 구현에 대한 내용을 기술한다.

2. 계층적 클러스터링과 용어 적합성 피드백

2.1 계층적 클러스터링

문서 클러스터링의 목적은 컴퓨터에 의한 파일 검색을 효율적으로 하는데 있으므로 일종의 파일 조직 기법으로 볼 수 있다. 클러스터링 결과에 의해 서로 연관된 문서들이 하나의 클러스터로 재조직되며, 따라서 파일 전체를 탐색하는 대신 정보 요구 주제와 관련된 클러스터를 탐색함으로써 탐색시간의 절약과 검색 효율의 향상을 기대할 수 있다[2]. 또한, 정보에 대한 시각적 표현을 가능하게 하여 연관된 문서들을 확장시킬 수 있다[3]. 문서 클러스터링 기법들은 일반적으로 문서에 부여된 색인어나 또는 기계적으로 추출된 키워드를 문서 내용의 식별요소로 삼아 클러스터를 형성한다. 형성된 클러스터는 클러스터를 대표하는 용어 집합을 갖게 되며, 탐색시 정보 요구와 각 클러스터 용어 집합이 대조되어 정보 요구에 가장 유사한 클러스터가 선택되는 것이다.

클러스터링 기법은 접근 방법에 있어서 두 가지 부류로 구분되는데, 문서간의 유사도를 측정하여 유사도 행렬을 작성하고 이로부터 계층적 클러스터를 형성하는 기법과 임의로 선택된 초기의 클러스터로부터 문서를 클러스터로 재배치하는 작업을 반복하여 최종의 클러스터를 형성하는 방법을 들 수 있다. 후자의 경우, 클러스터링 시간은 빠르나 대부분 검색 효율이 떨어지고 문서의 입력순서에 따라 클러스터링 결과가 변화하는 문제점을 갖는다.

본 논문에서는 클러스터링 방법으로 유사도 측정에 의한 계

층적 클러스터링 기법을 사용한다. 각 클러스터들의 대표벡터는 포함된 용어들의 가중치 평균으로서, 용어들의 빈도 수를 가중치로 설정하였다. 문서 i에 포함된 용어 j의 가중치 계산은 다음과 같다.

$$W_{ij} = TF_{ij} * [\log_2(n) - \log_2(IF_j) + 1]$$

W_{ij} : 문서 i에 포함된 용어 j의 가중치
n : 전체 문서의 수
TF_{ij} : 문서 i에서 용어 j의 빈도
IF_j : 용어 j가 존재하는 문서 수

가중치 W_{ij}는 문서 i에 포함된 용어 j의 빈도 수가 높을수록 높은 값을 가지고, 용어 j가 포함된 문서 수가 많을수록 문서간의 분별력이 떨어지므로 낮은 값을 가진다. 또한, 클러스터링을 위한 문서와 문서간의 유사도는 비교적 널리 사용되는 코사인 계수를 사용하였다.

$$SIM(D_i, D_j) = \frac{\sum_{k=1}^{n}(D_{i,k} \times D_{j,k})}{\sqrt{\sum_{k=1}^{n}(D_i, k)^2 \times \sum_{k=1}^{n}(D_j, k)^2}}$$

코사인 계수에서 D_i와 D_j는 문서를 의미하고, 문서와 클러스터간에도 이 공식을 사용한다.

2.2 적합성 피드백

적합성 피드백은 문서 적합성 피드백과 용어 적합성 피드백으로 나누어지는데, 문서 적합성 피드백은 사용자가 검색된 문서중 자신에게 적합하다고 생각되는 문서를 선택하며, 시스템은 이 문서에서 용어를 추출하여 질의를 개선한다. 이와 비교하여 용어 적합성 피드백은 사용자에게 용어를 보여주고, 선택된 용어를 질의에 추가하거나 가중치를 개선하여 질의를 확장하는 방법이다. 기존 연구에서 용어 적합성 피드백 방법으로 검색된 상위 5개 문서에서 용어를 추출하고 가중치를 계산하여 사용자에게 용어를 보여주는 방안을 제시하기도 하였다[6].

용어 적합성 피드백은 시스템이 의미 있는 용어를 사용자에게 제시하는 방법이 중요한데, 본 논문에서는 계층적 클러스터링을 통해 시스템이 소장하고 있는 문서의 구조적 정보를 얻어내고, 질의로 들어온 사용자의 사연에 대해 계층적 클러스터 탐색을 수행한다. 탐색 중, 매칭함수 값이 유사한 노드가 있다면 탐색을 멈추고 질의를 생성한다.

3. 시스템 설계 및 구현

3.1 상담 사례 검색

상담 사례 베이스는 형태소분석에 의해 추출된 색인어와 빈도수 그리고 빈도수에 따른 가중치로 구성되어 있다. 이들은 평균 연결 클러스터링 방법(average linkage method) 에 의해 계층적 구조를 형성한다. 그리고, 상담사례베이스에서 질의어 생성시 쓰일 용어 가중치 사전을 구축한다. 용어 가중치 사전은 색인어를 가중치 순으로 정렬한 것이다. 시스템의 검색 과정은 다음과 같이 진행된다. 먼저 사용자가 사연을 보내면, 인터페이스를 통해 색인어를 추출하여 사연을 벡터화시키고, 이를 질의 벡터로 사용하여 상담사례 검색 엔진의 계층적 구조를 탐색한다.

하향 탐색을 하는 도중 매칭 함수 값이 비슷한 노드를 만나면 3.2 에 기술된 질의 생성 규칙에 의해 질의어를 생성한다. 시스템은 질의어 생성규칙에 의해 사용자에게 관련 용어를 제시하고, 사용자는 관련용어를 선택하여 질의 벡터를 개선한다. 그리고 설정한 임계치 이상인 노드를 만나면 탐색을 중지하고 그 클러스터 안의 문서를 탐색한다. 본 논문에서 제시하는 상담사례 검색 시스템의 구성도는 그림 1과 같다.

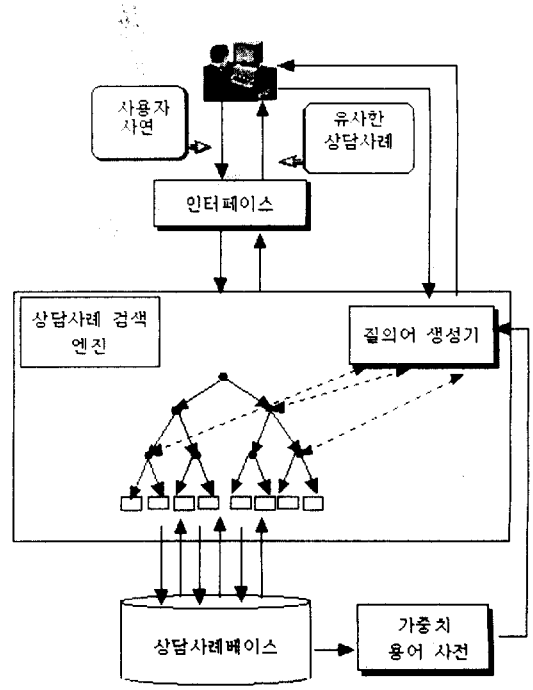

그림1. 상담 사례 검색 시스템 구성도

3.2 질의 문서 벡터의 가중치 개선

노드 선택이 어려울 경우의 사용자에게 질의를 생성하게 되는데, 질의 생성 함수는 다음과 같이 표현한다.

$$f_Q(Clust_1, Clust_2) = \left| \frac{D(Clust_1) - D(Clust_2)}{D(Clust_1) + D(Clust_2)} \right| < T$$

단, $D(Clust_1)$: $Clust_1$에 대한 결정함수
$D(Clust_2)$: $Clust_2$에 대한 결정함수
T : 질의 생성 임계치(Threshold)
f_Q : 질의 생성 함수
$Clust$: 클러스터

질의 생성함수에 따라 임계치 이하일 때 질의 용어를 제시하게 되는데, 질의 용어 제시는 각 클러스터간의 차집합 용어 중 가중치 사전에 존재하며, 가중치가 높은 것부터 각 클러스터 별로 다섯 개씩 제시한다. 이처럼 제시된 용어에 대하여 사용자는 관련정도를 시스템에 돌려준다. 그림 2 는 계층적 클러스터링을 이용한 용어 적합성 피드백 과정이다.

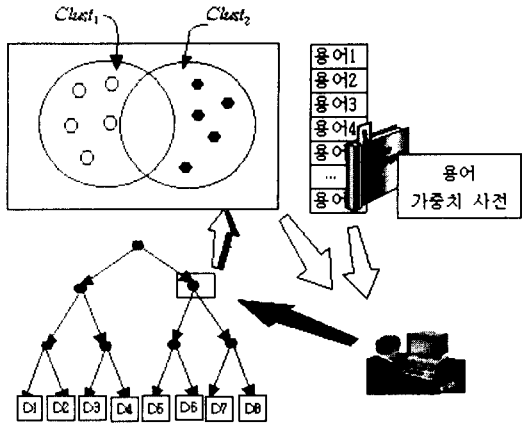

그림 2. 용어 적합성 피드백

용어 적합성 피드백 과정에 질의어의 개선은 다음 식으로 하였다.

$$QD_{new} = QD_{old} + \beta R_{word} - \gamma NR_{word}$$

단, QD_{new} : 개선된 질의 문서 벡터
QD_{old} : 예전의 질의 문서 벡터
R_{word} : 선택되어진 관련된 용어
NR_{word} : 선택되어진 관련 없는 용어
β, γ : 가중치

가중치는 관련 있는 용어를 관련 정도에 따라 상, 중, 하로 구분하고, β의 값을 각각 1, 1/2, 1/4로 주었고, γ는 1/4로 하였다.

4. 실험 및 평가

실험에 사용한 문서 리스트는 통신상에서 추출한 상담사례 텍스트 파일 300개와 질의할 사연 20개로 구성되어 있다. 평가 방법은 코사인 유사도를 이용한 매칭함수에 의한 방법과 계층적 클러스터링과 용어 적합성 피드백을 결합한 방법을 비교 평가하였다. 평가에 사용되는 평균정확율과 평균재현율은 다음과 같이 기술된다.

$$\text{평균재현율} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{검색된 적합 문서수}_i}{\text{적합 문서 총수}_i}$$

$$\text{평균정확율} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{검색된 적합 문서수}_i}{\text{검색 문서 총수}_i}$$

여기에서, n은 질의 문서의 총수를 의미한다.

표 1은 4가지 검색 기준치에 따른 검색 결과를 보여준다. 검색 기준치는 문서와 문서 사이의 코사인 유사도를 의미한다. 결과에서, 정확율이 떨어진 요인은 용어 적합성 피드백 과정에서, 사용자가 용어 관련정도를 평이하게 기록하였기 때문이다. 실험 결과에서 매칭함수에 의한 일반적인 검색과 비교하여 정확율은 약간 떨어지지만 재현율이 향상됨을 보였다.

표1. 4가지 검색 기준치에 따른 정확율과 재현율

방법 / 검색 기준치	매칭함수에 의한 방법		계층적 클러스터링과 용어 적합성 피드백	
	평균정확율	평균재현율	평균정확율	평균재현율
0.2	0.833	0.109	0.778	0.152
0.15	0.545	0.130	0.407	0.239
0.1	0.344	0.239	0.230	0.391
0.05	0.187	0.501	0.138	0.630

5. 결론 및 향후 연구 방향

본 논문에서 제안한 계층적 클러스터링과 용어 적합성 피드백을 이용한 질의 벡터의 개선을 통해 재현율을 향상시켰으나 정확율은 떨어졌다. 앞으로, 정확율을 떨어뜨린 요인을 분석하여 이를 개선하는 방향에 대해 연구할 것이다. 또한, 용어 제시 방법에 있어서 관련 있는 단어를 보여 주고 사용자가 용어를 체크하는 방식은 사용자에게 너무 딱딱함을 준다. 향후, 이러한 딱딱함을 완화시키기 위해서 제시할 용어로 이루어진 문장을 구성하는 방안을 연구할 것이다.

참고 문헌

[1] 박세진, "한국어 정보 검색 시스템에서 적합성 피드백에 관한 연구", 부산대학교 석사 학위 논문, 1998.

[2] 정영미, "정보검색론", 구미 무역(주) 출판부, pp. 354, 1993.

[3] J. Davies, R. Weeks, and M. Revett, "Using Clustering in a WWW Information Agent", 18th BCS IR Colloquium, Manchester, UK, 1996.

[4] 김호성, 고희정, "용어 빈도수를 이용한 영문 문헌정보의 점진적 개념적 집단화", 한국정보과학회 논문지, Vol. 19, No. 2, pp. 12-22, 1992.

[5] G. Kowalski, "Information Retrieval systems", Kluwer Academic publishers, pp. 282, 1997.

[6] 윤보현, 백대호, 김상범, 한경수, 임해창, "대화적 질의 확장을 통해 의미적 용어불일치를 완화하는 정보 검색 방안", 한국정보과학회 봄 학술발표논문집, Vol. 26, No. 1, pp. 345-347, 1999.