그래프 중심성 분석에 의한 CQI 보고서 핵심어 추출 시스템의 설계 및 개발†

테이퍼악떠라*, 임종범**, 이종혁***, 길준민****,† †
*대구가톨릭대학교 컴퓨터정보통신공학과
**한국산업기술대학교 게임공학부
***대구가톨릭대학교 빅데이터공학과
****대구가톨릭대학교 IT공학부
e-mail: pheaktra97@gmail.com, jblim@kpu.ac.kr, {jonghyuk, jmgil}@cu.ac.kr

Design and Implementation of Keywords Extraction System from CQI Reports by the Analysis of Graph Centrality†

They Pheaktra*, JongBeom Lim**, JongHyuk Lee***, Joon-Min Gil****,† †
*Dept. of Computer and Info. & Comm. Eng., Daegu Catholic University
**Dept. of Game & Multimedia Eng., Korea Polytechnic University
***Dept. of Big Data Eng., Daegu Catholic University
****School of Information Technology Eng., Daegu Catholic University

요 약

최근 대학교는 CQI(Continuous Quality Improvement) 등의 방대한 교육 관련 데이터를 수집하고 있고 이를 분석하여 교육 및 경영에 활용하고 있다. 핵심어는 텍스트의 내용을 간결하게 표현할 수 있는 단어이다. 그래서 CQI 보고서의 의미를 파악하기 위해서는 먼저 핵심어 추출이 필요하다. CQI 보고서에서 핵심어를 추출하면 이후 정보 검색, 인덱싱, 분류, 클러스터링, 필터링 등과 같은 많은 응용 작업을 용이하게 수행할 수 있다. 따라서 방대한 양의 CQI 보고서로부터 핵심어 추출을 자동화한다면 이후 요약 및 의미 파악에 많은 도움이 될 것이다. 이 논문에서는 CQI 보고서 요약을 위해 자동적으로 핵심어를 추출하는 방법을 제안한다.

1. 서론

CQI(Continuous Quality Improvement)[1-2] 보고서는 지속적 수업 질 개선을 위해 강의에 대한 학생 평가, 강의에 대한 교수 자체평가, 강의개선 계획사항, 강의개선 결과 등의 강의평가 보고서로서 대표적인 교육 관련 텍스트 데이터이다. 최근 대부분 대학교에서는 교양 및 전공과목을 포함한 모든 과목에 대해서 CQI 제도를 도입하고 있으며 이를 통해 강의 개선에 많은 노력과 투자를 하고 있다. 그러나 한 과목 단위가 아닌 학과 또는 단과대학 단위로 사람이 직접 CQI 보고서를 요약 및 분석하기에는 보고서의 양이 방대하여 핵심어 추출과 의미 파악에 오랜 시간이 소요되는 등 여러 가지 어려움에 직면한다.

본 논문에서는 CQI 보고서 데이터를 분석하기 위해 자동 핵심어 추출 방법을 제안한다. 수집된 CQI 보고서 데이터는 분석이 용이하도록 전처리 과정을 먼저 수행한다. 본 논문에서 사용한 전처리 방법은 CQI 보고서 데이터에서 분석에 불필요한 동사, 조사, 기호, 숫자, 한 글자 등을 제거하고 명사를 추출하는 것이다. 그런 다음, Word Count 알고리즘을 이용하여 단어 개수를 계산하여 상위 N개 단어를 추출하며, 추출된 단어는 핵심어 추출(keyword extraction)[3-5]을 통해 단어와 단어 사이의 관계를 그래프로 표현한다. 본 연구에서는 핵심어 추출 기법으로 KE1(Keyword Extraction 1)과 KE2(Keyword Extraction 2)를 정의하여 사용한다[6]. KE1 알고리즘은 이웃한 단어 간의 관계를 찾는 방법이다. 예를 들면, 어떤 단어와 이 단어의 이웃 단어 사이에 어떤 관계가 있는지를 계산한다. KE2 알고리즘은 슬라이딩 윈도우(sliding window)를 이용하여 단어 관계를 좀 더 광범위하게 찾는 방법이다. 즉, KE1처럼 이웃 단어 사이의 관계뿐만 아니라 문장 안에서 조금 더 멀리 있는 단어 간의 관계를 찾는 방법이다.

모든 단어의 관계를 쉽게 표현하기 위해 그래프 도구(graph-tool) 라이브러리[7]를 이용하여 그래프로 표현한다. 또한 중심성(centrality) 기법을 통해 단어와 다른 단어 사이에 관계 정도들 표현하기 위해 다음과 같이 세 가지 기법을 이용한다: 근접 중심성(Closeness Centrality), 페이지랭크 중심성(Pagerank Centrality), 그리고 매개 중심성(Betweenness Centrality)[8]. 모든 중심성의 계산 결과는

† 이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2016R1D1A3B03933370).
† † 교신저자

상황에 따라 우수 여부가 달라지기 때문에 본 연구는 세 가지 기법의 중심성 결과를 정규화(normalization)한 후 가중치를 부여한 방법을 사용한다.

2. 연구 모델

이 절에서는 핵심어 추출(keyword extraction) 기법을 이용하여 단어와 단어 사이에 관계를 추출하여 단어 간의 관계를 그래프 도구 라이브러리[14]를 이용하여 표현하고, 그래프에 기반하여 정점의 중심성을 유도함으로써 CQI 보고서에서 핵심어를 찾아주는 본 연구의 전반적인 연구 모델에서 단계별 설명은 다음과 같다(그림 1).

[단계 1] CQI 보고서 수집 정보를 설정하여 CQI 보고서 분석에 필요 없는 단어인 불용어(동사, 조사, 기호, 숫자, 한 글자 등)를 배제하고 명사만으로 구성된 데이터 세트를 구성하는 전처리 과정을 수행한다.

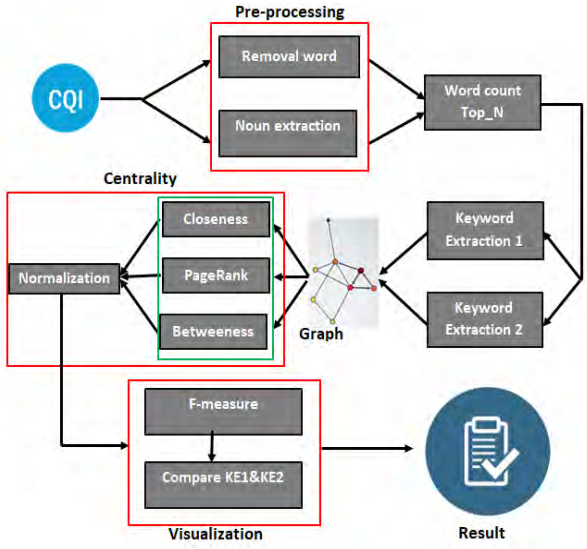
[단계 2] 수집된 데이터 세트에서 단어들의 개수를 카운트함으로써 CQI 보고서 내에서 어떤 단어가 제일 많은지를 계산하고 상위 N개(TopN)의 단어를 추출한다.

[단계 3] 단어 간의 관계를 얻어내기 위해 핵심어 추출 기법을 활용한다. 본 연구에서는 KE1과 KE2이라 명명된 기법을 이용한다. 이 두 기법에 대한 자세한 설명은 3장에서 살펴본다.

[단계 4] 단어 간의 관계를 쉽게 파악하고 주요 단어를 추출할 수 있도록 그래프를 구성한다.

[단계 5] 그래프에서 주요 단어를 찾기 위해 3가지 중심성 기법(근접 중심성, 페이지랭킹 중심성, 매개 중심성)을 적용하며, 각 중심성 기법의 결과에 대해서 정규화를 적용하고 가중치 평균값을 구한다.

[단계 6] F-score를 적용한 성능평가를 통해 CQI 보고서에서 핵심어를 추출하고 시각화를 통해 결과를 확인한다. 비잔틴 실패가 아닌 프로세스 P_i는 자신의 초기 값을 다른 노드에게 전달한다.



(그림 1) 제안하는 핵심어 추출 시스템의 전체 흐름도

3. 핵심어 추출 및 핵심어 간 관계 생성

3.1 전처리

전처리 단계는 동사, 조사, 기호, 숫자, 한 글자 등과 같은 불용어를 제거하고 의미적 판단이 명확한 명사를 추출하는 전처리 단계이다. 이를 통해 데이터양을 감소시켜 분석 시간을 단축시킬 수 있다. 실제로 본 연구 수행을 통해 획득된 전처리 결과에 따르면 데이터양이 약 50%로 감소되었다. (그림 2)는 2017년도 1학기 A 단과대학의 개선결과 항목을 전처리한 결과를 보여준다.

시각적 자료 활용 발표 보충설명 학생들 막연 불안감 성공함 한국문화관련 영상 유익 정보 제공 계속함 수업 시간 충분한 피드백 보충자료 활용 보충설명 강화 각자 작품 발표 국문학작품 이해 이행함 인원 발표시간 조정 예상 발표 학생들 마이너스 요소 가상 발표 융통성 운용 영화 시간 영화 수업 시간 강의용 중요 부분들 정리해서 교수 중요부분 정리 강의지원시스템 영상활용 설계 관련 발표 의견 공유 수업 계속 각자 시간 학생들 문화 공유 이해 활용 수업 학생들 동영상 자료 활용 원전 풀이 자세 유인물 배포 수업 도움 활용 수업 학생들 접근 강의 진행함 이해 교육함 강의 편안 분위기 조성 발표 부담 완화 단계 논술쓰기 진행 자발적 발표 유도 학생들 취업 진로 관심 취업 구체적 방향 학습 계획 수립 교재 내용 이해 재구성 유인물 제공 해당 지식 이해 적용 발표 지식 활용 도구 차후 학습 내용 이론 바탕 시하 중요 내용 강조 안내 학습자들 자발적 토론 발표 유도 실생활 어문규범 실제 제시 흥미 문제풀이 기출문제 제시 지식 적용

(그림 2) CQI 보고서 전처리 결과의 예

3.2 단어 추출 및 관계 생성

본 단계는 전 단계의 결과로부터 단어들을 추출하고 단어 간 관계를 표현하는 그래프의 생성 단계이다. 이를 위해 각 단어에 대한 빈도를 계산하고 내림차순으로 정렬하여 상위 N개의 단어를 정점(node)으로 설정한다. 그리고 KE1(Keyword Extraction 1)과 KE2(Keyword Extraction 2)에 의해 생성된 단어 간 관계를 그래프의 간선(edge)으로 설정한다. 이를 수식으로 표현하면 수식 (1)과 같다.

$$G=(V,\ E) \tag{1}$$

여기서, G는 그래프, V는 상위 N개 단어로 구성된 정점 집합(node set), E는 정점 간 관계로 구성된 간선 집합(edge set)을 나타낸다. 간선 집합 E는 KE1 또는 KE2 방법을 통해 구성된다.

KE1 방법은 두 단어가 서로 이웃 관계에 있을 경우 단어 간 관계가 있음을 표현하는 방법으로 수식 (2)와 같다.

$$W_{i,j}=[W_i,\ W_j] \tag{2}$$

여기서, $W_{i,j}$는 하나의 문장에서 i번째 키워드(W_i)와 j번째 키워드(W_j) 간의 관계 집합($i=1,2,...,n-1$, $j=i+1$)을 나타낸다. 그리고 n은 마지막 단어 인덱스를 나타낸다.

KE2 방법은 슬라이딩 윈도우(sliding window)를 이용하여 단어 간 이웃 관계 범위를 슬라이드 내로 확장한 것으로 수식 (3)과 같다.

$$W_i=[W_{i-(ws-1)}, W_{i-(ws-2)},..., W_{i-1}, W_i] \tag{3}$$

여기서, W_i는 단어 i와 관계있는 단어 집합($i=1,2,...,n$)을 나타낸다. 그리고 n과 ws는 각각 마지막 단어의 인덱스와 슬라이딩 윈도우의 크기를 나타낸다.

3.3 중심성 계산

이 단계는 그래프 분석을 통해 그래프에서 중요 정점을 발견하는 단계로 본 논문에서는 근접 중심성(closeness centrality), 페이지랭크 중심성(page-rank centrality), 매개 중심성(betweenness centrality)의 세 가지 방법을 도입한다.

근접 중심성은 그래프에서 두 정점 사이의 최단 경로 길이의 합을 역수로 정의한 것으로 수식 (4)와 같다.

$$C_i^C = \frac{1}{\sum_{i,j \in V} d_{ij}} \quad (4)$$

여기서, i와 j는 그래프에서의 정점(단어)을 나타내며, $d_{(i,j)}$는 정점 i와 정점 j간의 최단 경로를 나타낸다.

페이지랭크 중심성은 정점의 반복적인 관계 정도를 정의한 것을 의미하며 수식 (5)와 같다.

$$C_i^{PR} = \sum_{j \in \Gamma -(i)} \frac{PR(j)}{L(j)} \quad (5)$$

여기서, $PR(j)$는 정점 i에 연결된 모든 정점을 포함하는 j의 페이지랭크 값을 나타내며, $\Gamma -(i)$와 $L(j)$는 정점 i와 연결된 정점과 정점 j의 모든 연결을 각각 나타낸다.

매개 중심성은 그래프에서 정점 대 정점 사이의 연결이 가장 중심이 되는 정도를 의미하며 다음과 같이 정의된다.

$$C_B(v) = \sum_{i,j \in v} \left(\frac{\sigma_{ij}(v)}{\sigma_{ij}} \right) \quad (6)$$

여기서, v는 그래프에서 두 정점 간 연결 경로 상의 정점을 나타내며, $\sigma_{ij}(v)$는 정점 i에서 정점 j까지 정점 v를 경유하는 최단 경로의 개수를 나타낸다. 그리고 σ_{ij}는 정점 i에서 정점 j까지 최단 경로 개수를 나타낸다.

또한, 여러 상황에 따라 유연하게 중심성을 계산하기 위해 세 가지 중심성 방법의 결과값에 대해 수식 (7)을 이용하여 정규화한다.

$$y = 1/(\max - \min) \times (x - \min) \quad (7)$$

여기서, x와 y는 정규화 전의 값과 정규화 후의 값을 각각 나타낸다. 그리고 min과 max는 구간의 최소값과 최대값을 각각 나타낸다.

다음으로 세 가지 중심성 중에 어떤 중심성이 제안 시스템에 적합한지를 평가하기 위해 수식 (8)로 정의되는 가중치 기반의 비용 함수를 사용한다.

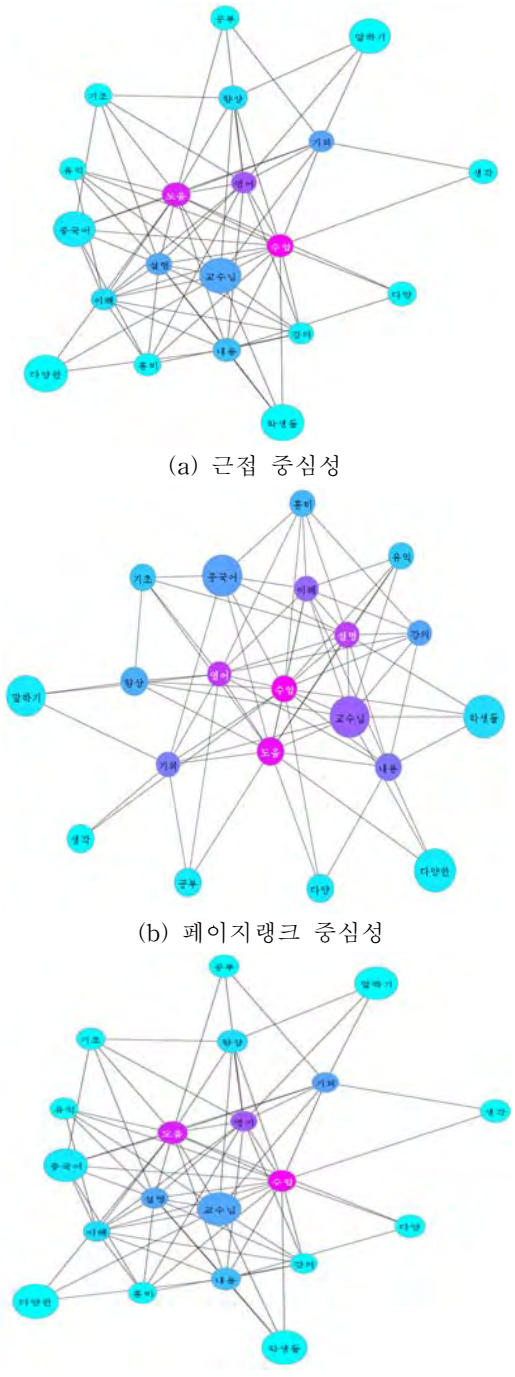
$$S = w_1 \cdot C + w_2 \cdot P + w_3 \cdot B \quad (8)$$

여기서, C, P, B는 근접 중심성, 페이지랭킹 중심성, 매개 중심성을 각각 나타낸다. 그리고 w_1, w_2, w_3는 근접 중심성, 페이지랭킹 중심성, 매개 중심성의 가중치를 각각 나타낸다($w_1 + w_2 + w_3 = 1.0$).

4. 성능 실험

CQI 데이터 수집을 위해 D 대학교의 2015년부터 2017년까지 매해 2개 학기 동안 8개 단과대학에서 작성한 데이터를 활용하며, CQI 보고서는 총 6개의 영역인 좋은점, 어려운점, 개선건의사항, 개선계획사항, 개선결과, 교수자 체평가로 구성되어 있다. 이 데이터는 개인별 혹은 학과별이 아닌 단과대학별로 그룹화된 CQI 보고서이다. 따라서 개인이나 학과의 익명성이 보장되는 CQI 보고서를 활용하여 개인정보 문제를 최대한 회피하도록 하였다.

이 논문에서 제안한 CQI 보고서 요약 핵심어 추출 시스템은 파이썬 언어에 기반하여 구현되었으며, 핵심어 그래프의 구성과 중심성 분석을 위해 graph-tool 버전 27[8]이 사용되었다. 실험을 위한 계산 도구로 Core i7 프로세스와 메모리 32G바이트의 윈도우 10 기반 PC가 활용되었다.

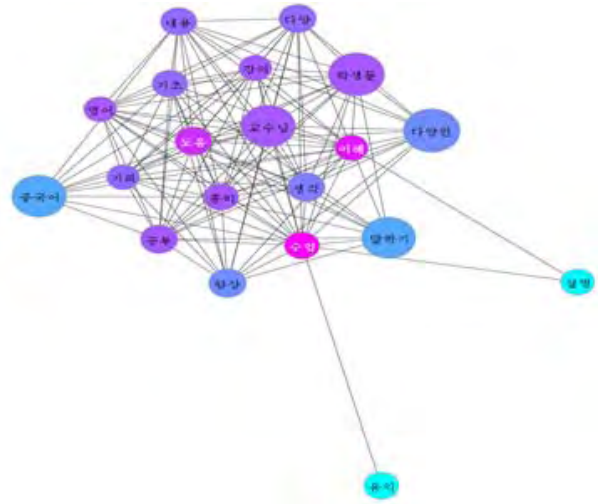


(a) 근접 중심성

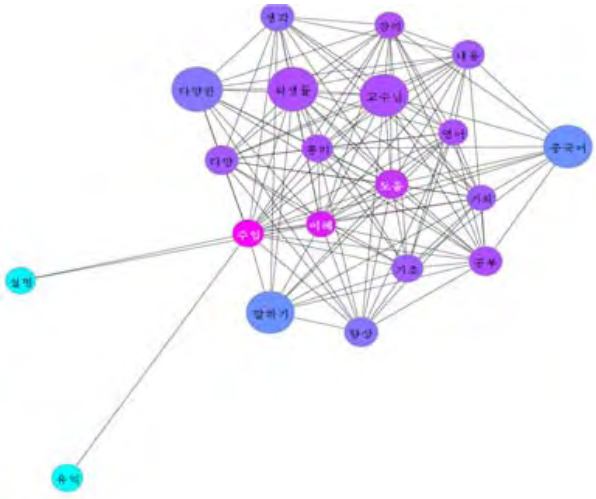
(b) 페이지랭크 중심성

(c) 매개 중심성

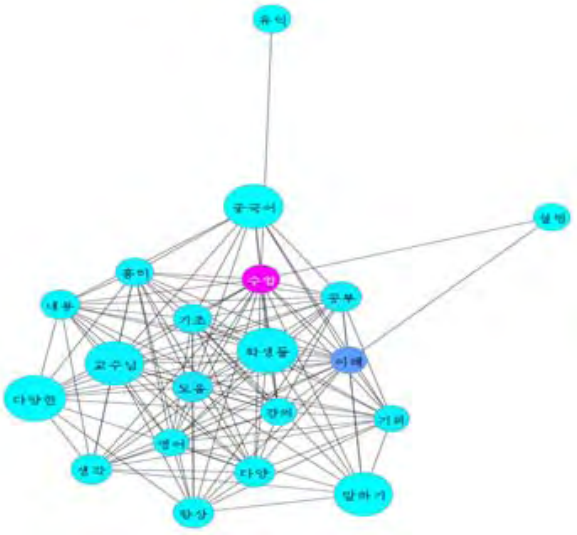
(그림 3) A 단과대학의 핵심어 그래프(KE1)



(a) 근접 중심성



(b) 페이지랭크 중심성



(c) 매개 중심성

(그림 4) A 단과대학의 핵심어 그래프(KE2)

(그림 3)과 (그림 4)는 각각 KE1과 KE2에 대해서 2017년도 1학기 A 단과대학의 중심성 기법별 핵심어 그래프를 보여준다. 이들 그림의 중심성 그래프를 비교하면, KE1이 단순한 형태의 연결성을 지닌 핵심어 그래프를 도출한 것에 비해 KE2가 좀 더 복잡한 형태의 연결성을 갖는 핵심어 그래프를 도출함을 알 수 있다.

5. 결론

이 논문에서는 CQI 보고서의 단어와 단어 사이에 관계를 분석하여 이를 토대로 핵심어를 추출하는 CQI 보고서를 위한 핵심어 추출 시스템을 개발하였다. 제안 시스템은 단어 간의 관계를 유도하기 위해 두 단어의 연속 발생만을 고려하여 추출한 방법(KE1)과 슬라이딩 윈도우를 이용하여 좀 더 복잡한 단어 관계를 고려하여 추출하는 방법(KE2)에 기반하여 추출된 정보에 의해 그래프를 구성하고 중심성 이론을 사용하여 그래프로부터 핵심어를 추출하도록 하였다. 근접 중심성, 페이지랭킹 중심성, 매개 중심성 기법을 통해 통해서 각 중심성 기법의 특징에 반영될 수 있는 핵심어를 추출하도록 하였다. 아울러, 추출된 상위 순위의 핵심어를 이용하여 방대한 양의 CQI 보고서로부터 요약문을 구성하였다.

참고문헌

[1] 김명랑, 윤우영, 김동환, 정진택, "'프로그램학습성과 및 평가' 실천을 위한 모형개발 및 전략에 대한 연구," 공학교육연구, 제10권, 제4호, pp. 29-42, 2007.

[2] 유인근, "공학교육인증 프로그램의 효과적인 운영방안에 관한 연구," 공학교육연구, 제10권, 제2호, pp. 62-72, 2007.

[3] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida, "KeyGraph: automatic indexing by co-occurrence graph based on building construction metaphor," Proceedings of IEEE International Forum on Research and Technology Advances in Digital Libraries, pp. 12-18, 1998.

[4] Girish Keshav Palshikar, "Keyword extraction from a single document using centrality measures," Proceedings of International Conference on Pattern Recognition and Machine Intelligence, pp. 503-510, 2007.

[5] Marina Litvak and Mark Last, "Graph-based keyword extraction for single-document summarization," Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, Association for Computational Linguistics, pp. 17-24, 2008.

[6] 유준현, 박순철, "완전그래프를 이용한 문서요약 연구," 한국산업정보학회 논문지, 제10권, 제2호, pp. 26-31, 2005.

[7] A. Nenkova, and K. McKeown, "Automatic summarization," Foundations and Trends in Information Retrieval, Vol. 5, No. 2-3, pp. 103-233, 2011.

[8] Graph-tool performance comparison, https://graph-tool.skewed.de/ performance, accessed 30 Dec. 2017.