고품질 슬라이드 선별을 위한 지식구조 기반 분류 기법

정 원 철○ 김 성 찬 이 문 용
한국과학기술원 지식서비스공학과
wonchul.jung@kaist.ac.kr, sckim@kaist.ac.kr, munyi@kaist.ac.kr

Proposing and Validating a Classification Method based on Knowledge Structure to Identify High-Quality Presentation Slides

Wonchul Jung○ Seongchan Kim Mun Y. Yi
Dept. of Knowledge Service Engineering, KAIST

요 약

본 연구는 내용적으로 고품질인 슬라이드를 분류하기 위해, 슬라이드의 지식정보를 내포하는 지식구조를 이용하는 분류 방법을 제안한다. 지식구조가 슬라이드의 내용적 품질정보를 내포하는지에 대해서 분석한 후, 그 결과로부터 지식구조를 이용한 분류 방법을 개발하였고, 슬라이드의 품질별로 분류한 결과를 비교하였다. 비교를 통해 고품질군에 속하는 슬라이드일수록 높은 품질의 슬라이드 위주로 분류할 수 있다는 점을 검증하였다. 이는 품질이 높은 슬라이드 위주로 검색하거나 추천하고자 할 때, 지식구조라는 인지적 모형을 활용하여 그 효과를 높일 수 있음을 보여준다.

1. 서 론

오늘날 PowerPoint처럼 컴퓨터로 만들어진 발표용 슬라이드(Presentation Slide)는 폭발적으로 늘어나는 추세이다. Simons[1]는 "전세계적으로 PowerPoint 기반의 발표 자료들이 하루에 2,000 ~ 3,000만개씩 만들어진다."라고 하였다. 또한, Cassady [2]는 학생들이 전통적인 칠판 방식의 수업보다 슬라이드 발표 방식을 선호한다는 것을 밝혀냈다.

이러한 현상을 반영하듯이 교육을 위해서 슬라이드를 제공해주는 SlideShare[1], CourseShare[2]같은 서비스도 생겨났다. 하지만 온라인상에 존재하는 수많은 슬라이드 중에 전문가가 만든 고품질의 슬라이드뿐만 아니라, 비전문가가 만든 낮은 품질의 슬라이드들도 많이 존재한다. 이와 같이 다양한 품질의 슬라이드가 존재하는 학습 환경에서, 학습자가 고품질의 자료를 선별할 수 있도록 도와주는 서비스 및 시스템 개발이 필요하다.

슬라이드의 품질을 측정하는 Kim et al.의 연구[3]에 따르면, 학습자가 슬라이드의 품질을 측정할 때 고려한 상위 11개의 요소들 중에 예제의 유무, 슬라이드의 길이와 같은 내용적 측면(Informativeness)이 6개에 해당한다. 이는 슬라이드에서 품질을 측정할 때 내용적 요소가 중요하다는 것을 보여준다. 본 연구에서는 내용적 측면을 분석할 수 있는 도구로 지식구조를 활용하고, 지식구조가 품질을 효과적으로 반영하는지 검증하고자 한다. 지식구조는 학습자가 어떤 문서나 매체 등을 통해 학습할 때 생성되는 핵심 개념들과 그들의 연관 관계를 조직적으로 나타낸 모형이다[4]. 본 연구에서는 지식구조를 무향 그래프(undirected graph)로 나타냈다. 지식구조를 나타내는 그래프에서 원(node)은 슬라이드에서 중요한 핵심 개념이며, 원의 크기는 자주 발생한 빈도이다. 선(link)은 핵심 개념 간의 관계를 나타내며, 선이 굵고 짙을수록 핵심 개념 간의 관계가 높다.

본 연구에서는 1) 슬라이드에서 추출한 지식구조가 슬라이드의 내용적 품질(Quality)을 잘 반영하는지 통계적 방법론으로 검증하고, 2) 슬라이드에서 지식구조를 활용한 문맥적 품질 분류 방법을 제안하고, 그에 따른 효과를 입증하려한다.

2. 제안하는 방법

2.1. 전체 수행 과정

본 연구에서 제안하는 지식구조를 이용한 고품질 슬라이드 분류방법의 순서는 다음과 같으며, 그림 1과 같이 나타낼 수 있다.

1) 슬라이드로부터 지식구조 추출
2) 지식구조를 활용한 슬라이드 분류

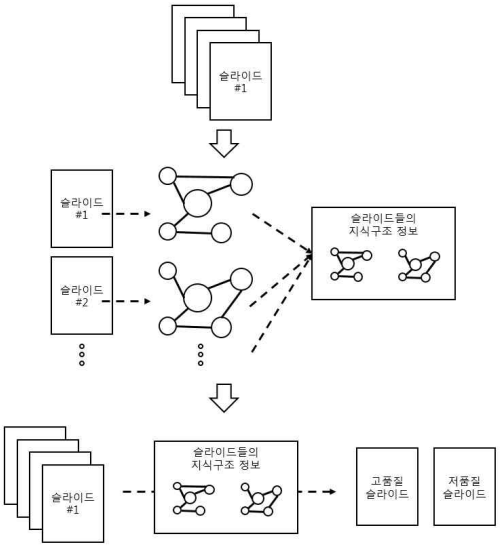


그림 1 분류 기법 전체 수행과정 순서도

2.2. 슬라이드로부터 지식구조 추출

슬라이드로부터 지식구조를 추출하는 방법은 Kim & Yi의 단일문서에서 지식구조 추출법에 관한 연구[5]에서 제시한 방법론

* 이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2013-053852).

1) http://slideshare.net
2) http://courseshare.org

을 기본적으로 적용하고, 슬라이드에 맞게 응용하였다.

1) 핵심 개념 추출: Term Frequency(TF)를 이용해 슬라이드의 핵심 개념을 빈도수로 13순위까지 추출한다. 또한, 핵심 개념이 단위 슬라이드의 제목일 경우 빈도수에 3배의 가중치를 줬으며, 소제목일 경우에는 2배의 가중치를 주었다.

2) 핵심 개념 간 연관관계 추출: 식 (1)은 슬라이드의 각 단위 슬라이드 간 공기정보를 이용하여 핵심 개념 간 유사도(Unit-Slide co-occurrences Similarity: US)를 구하는 공식이다.

$$US_{ij} = \frac{\sum_{1}^{N_u} n(W_i \cap W_j)}{Max(C_u)}, (0 \le US \le 1) \qquad (1)$$

위의 식에서 N_u는 슬라이드에 나타난 순서에 따른 단위 슬라이드 번호가 된다. 핵심 개념 W_i와 핵심 개념 W_j의 유사도는 각 단위 슬라이드에서 함께 나타난 횟수의 총합을 슬라이드에서 나타난 각 단위 슬라이드 공기정보의 최대값으로 나누어 정규화 시킨다.

3) 패스파인더 알고리즘 적용: 아래 기술된 식 (2)를 사용하여 각 핵심 개념 간 연관관계를 7점 스케일(1:관련 있음, 7:관련 없음)로 변환한 후, 패스파인더 알고리즘[6]을 적용하여 각 핵심 개념 간 최단거리를 연결해주는 지식구조를 생성한다.

$$D_{ij} = 7 - US_{ij} \times 6, (1 \le D \le 7) \qquad (2)$$

4) 슬라이드들의 지식구조 정보 저장: 과정 1) ~ 3)을 모든 슬라이드들에 적용하여, 지식구조 정보 집단을 저장한다.

2.3. 지식구조를 활용한 슬라이드 분류

본 연구에서는 슬라이드의 지식구조에서 하나의 핵심 개념을 분류 질의어(Query) 기준으로 사용하고, 핵심 개념과 연결된, 다른 핵심 개념들과 각 연관관계들로부터 분류 질의어를 확장한다. 슬라이드들의 지식구조 정보가 갖춰진 환경에서 지식구조를 사용한 품질 분류 방법의 자세한 순서는 다음과 같다.

1) 하나의 지식구조에서 하나의 핵심 개념을 선택한다.

2) 선택된 핵심 개념과 연결된 다른 핵심 개념들 및 각 연관관계들을 추출한다. 식 (3)은 연관관계를 추출하는 공식이며, 역수를 취하는 이유는 후의 과정에서 다차원 공간에 대입할 때 연관관계가 높을수록 가중치를 주기 위함이다.

$$L_{ij} = \frac{1}{D_{ij}} \times 7, (1 \le L \le 7) \qquad (3)$$

3) 선택된 핵심 개념을 포함하는 다른 지식구조들에서도 과정 2)와 같은 방식으로 값들을 추출한다.

4) 과정 2) ~ 3)에서 추출된 각 값들을 하나의 슬라이드를 대표하는 핵심 개념 집단 슬라이드로 가정하여, 표 1과 같이 거대한 2차원 배열로 저장한다. 2차원 배열의 행은 각 슬라이드를 나타내며, 열은 모든 슬라이드들의 핵심 개념들을 나타난다.

표 1 지식구조들에서 추출한 핵심 개념 집단 2차원 배열 예

	cluster	distance	pattern	...
슬라이드 #1	6	3	0	...
슬라이드 #2	1	0	0	...
슬라이드 #3	0	0	0	...
...	...	...	...	...

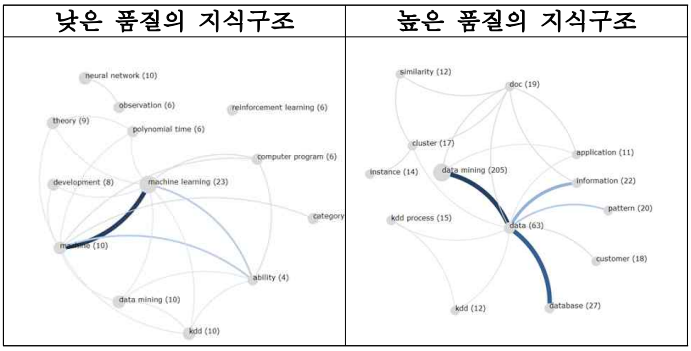
5) 2차원 배열의 열 값(추출된 모든 핵심 개념)으로 이루어진 다차원 공간에 2차원 배열의 행 값(하나의 슬라이드로 가정된 값)들을 대입한다.

6) 다차원 공간에서 각 지식구조에서 추출된 값들의 유사도를 계산하여 분류한다. 유사도를 계산하는 방법으로는 Cosine Similarity를 사용했다.

3. 실험 및 결과

3명의 검수자(Annotator)들에게 1,000개의 슬라이드 품질을 낮음(Low), 보통(Fair), 높음(High)으로 측정하게 하였다. 슬라이드들의 내용은 검수자들의 전공인 데이터마이닝, 컨텐츠 네트워킹에 관한 슬라이드였다. 2명이상 같은 품질로 응답을 한 슬라이드들만 추렸고, 카파 값은 0.67이였으며, 총 843개의 품질이 측정된 슬라이드를 얻어냈다. 검수자들로부터 얻어낸 843개의 슬라이드 중 품질이 낮음인 슬라이드는 197개, 보통인 슬라이드는 386개, 높음인 슬라이드는 260개였으며, 슬라이드의 품질이 낮음은 0, 보통은 1, 높음은 2로 기록하였다.

843개의 슬라이드들로부터 각각의 지식구조를 생성하였고, 지식구조가 슬라이드의 품질정보를 내포하는지 알아보기 위해, 품질에 따라 분류한 후, 지식구조가 품질에 따라 다른 특성을 보이는지 관찰하였다. 그림 2와 같이 품질이 낮은 슬라이드는 핵심 개념들의 발생빈도가 낮다는 것을 핵심 개념 옆에 나타난 발생빈도 숫자 값을 보고 알 수 있었다. 또한, 품질이 낮은 슬라이드는 핵심 개념 간의 연관관계를 나타내는 선들이 얇다는 것 관찰할 수 있었다. 관찰결과, 품질이 낮을수록 지식구조에서 핵심 개념의 발생빈도가 낮고, 핵심 개념 간의 연관관계 값이 작다는 것을 알 수 있었다.



*원: 핵심 개념을 나타냄, 숫자는 발생빈도
*선: 핵심 개념들 간의 연관관계, 굵고 짙을수록 관계가 높음

그림 2 품질에 따른 지식구조

이를 통계적으로 검증하기 위해, 품질 낮음 집단(A)과 높음 집단(B)으로 설정하고, t-검정(이분산 가정 두 집단)을 시행하였다. 두 집단 간 핵심 개념들의 발생빈도를 비교하기 위해 상위 3순위 핵심 개념까지의 발생빈도 평균을 비교해보았다. 또한, 두 집단 간 핵심 개념 간의 연관관계를 비교하기 위해 모든 연관관계의 평균을 비교해보았다. 표 2와 같이 두 집단 간 핵심 개념 발생빈도 및 핵심 개념 간의 의미적 연관관계가 차이가 있다는 것을 알 수 있었다. 두 실험 모두 95% 신뢰수준에서 유의하다는 결론이 나왔다. 실험을 통해 핵심 개념과 각 핵심 개념 간의 공기정보를 바탕으로 만들어진 지식구조가 슬라이드의 내용적 품질정보를 내재한다는 것을 발견할 수 있었다.

표 2 품질에 따른 지식구조의
핵심 개념 빈도 및 핵심 개념 간 연관관계 비교

핵심 개념 빈도 비교	Low+Fair (A) Keyword1~3	High (B) Keyword1~3
평균	87.2487	130.9615
분산	8241.6236	32153.8904
관측수	583	260
자유도	320	
t 통계량	-3.7237	
P(T<=t) 단측검정	0.0001	

핵심 개념 간 연관관계 비교	Low+Fair (A)	High (B)
평균	4.2102	4.3607
분산	1.0159	0.7044
관측수	583	260
자유도	591	
t 통계량	-2.2563	
P(T<=t) 단측검정	0.0122	

앞선 실험들을 통해, 지식구조가 슬라이드의 내용적 품질 측면을 잘 반영한다는 것을 알게 되었다. 추가로, 고품질의 슬라이드에서 지식구조 기반 분류를 하면 높은 품질의 슬라이드들을 분류할 수 있는지에 대해 검증해보았다. 실험은 앞의 실험들과 동일한 843개의 슬라이드 환경에서 시행되었다. 각 슬라이드의 지식구조에서 높은 순위로 발생한 핵심 개념 순서로 2.3절에서 제안한 분류 방법을 적용하여 분류된 슬라이드들의 값들을 표 3과 같이 기록하였다.

표 3 지식구조 기반 분류를 통해 분류된 값 기록 예

ID	품질	핵심 개념 기준	핵심 개념 순위	분류수	품질 평균
1	0	pattern	1	7	0.71429
2	2	distance	3	3	1.33333
3	2	diaper	2	1	2.0
4	0	cluster	1	5	0.8
...	...	...	...	...	...

*ID: 분류 기준이 되는 슬라이드 ID
*품질: 슬라이드의 품질(0:낮음, 2:높음)
*핵심 개념 기준: 해당 핵심 개념을 기준으로 2.3절에서 제안한 방법 적용
*핵심 개념 순위(n): 핵심 개념 기준의 발생빈도 순위
*분류수: 분류 방법을 통해 분류된 슬라이드 개수
*품질평균: 핵심 개념 기준으로 분류된 슬라이드들의 평균품질(0:낮음, 2:높음)

고품질의 슬라이드에서 분류할수록, 분류된 슬라이드들이 높은 품질인지 알아보기 위한 실험을 위해, 가장 먼저 표 3와 같이 기록된 값에서, 핵심 개념 순위별로 분류하였다. 이는 동일한 발생빈도의 핵심 개념 기준을 가지고 각 분류결과를 비교해보기 위함이었다. 핵심 개념 순위를 3순위까지 분류한 후, 다시 분류의 기준이 되는 슬라이드의 품질 낮음(A), 보통(B), 높음(C) 세 집단으로 분류한 후, 표 4와 같이 기록하였다.

표 4 지식구조에서 중요도가 N번째인 핵심 개념을 기준으로 분류된 슬라이드들의 품질평균의 평균 및 분산

핵심개념 순위	슬라이드 품질	품질평균 평균	품질평균 분산	표본수
1	Low_Doc	0.8115	0.2503	6
	Fair_Doc	1.0842	0.1535	23
	High_Doc	1.1727	0.2159	19
2	Low_Doc	0.9478	0.3791	74
	Fair_Doc	1.0807	0.2267	138
	High_Doc	1.3982	0.3302	105
3	Low_Doc	0.8666	0.3434	62
	Fair_Doc	1.0772	0.2516	135
	High_Doc	1.3917	0.2981	93

표 4와 같은 결과에서, 분류 질의어가 최상위 핵심 개념을 기준으로 만들어졌을 경우, 분류 가능 슬라이드의 표본 수가 적은 것을 알 수 있었다. 이는 한 슬라이드의 최상위 핵심 개념은 다른 슬라이드들에서 발견되지 않았다는 것을 의미한다. 두 번째로 중요도가 높은 핵심 개념부터는 품질 분류의 표본 수가 통계적으로 유의했다. 표 4에서 나온 결과들의 유의성을 알아보기 위해, 표본 수가 적은 결과를 제외하고, t-검정(이분산 가정 두 집단)을 시행하였다. 표 5와 같이 각 품질 집단에서 분류된 슬라이드들의 품질이 서로 차이가 있다는 것을 95% 신뢰수준에서 검정할 수 있었다. 이는 고품질의 슬라이드에서 지식구조를 사용한 분류를 하면, 높은 품질의 슬라이드들로 분류할 수 있다는 것을 입증한다.

표 5 핵심 개념 순위가 두, 세 번째로 높은 핵심 개념을 기준으로 분류한 결과들의 t-검정

2순위 핵심 개념 기준 분류				
	Low	Fair	Fair	High
자유도	121		200	
t 통계량	-1.6152		-4.5888	
P(T<=t) 단측검정	0.0544		0.000004	
t 기각치 단측 검정	1.2886		1.6525	
3순위 핵심 개념 기준 분류				
	Low	Fair	Fair	High
자유도	104		187	
t 통계량	-2.4484		-4.4167	
P(T<=t) 단측검정	0.0080		0.000008	
t 기각치 단측 검정	1.2886		1.6525	

4. 결론 및 향후 연구

지식구조가 슬라이드의 품질을 내재하고 있는지에 대해서 알아본 결과, 지식구조는 품질 정보를 효과적으로 반영한다는 것을 검증할 수 있었다. 또한, 지식구조를 이용한 분류 방법을 사용하면 고품질의 슬라이드에서 분류할수록 높은 품질의 연관 슬라이드들을 찾아서 분류할 수 있다는 것도 입증할 수 있었다. 이는 검색 또는 추천에서 품질이 높은 연관 슬라이드를 찾아내는데 매우 유용하게 적용될 수 있다. 향후 연구로써는 지식구조가 슬라이드의 내용적 품질뿐 아니라, 시각적 품질, 가독성 등 다양한 부분의 품질정보도 반영하는지에 관한 연구 및 그러한 특성을 이용한 확장된 검색에 관한 연구가 필요하다.

참고문헌

[1] Simons T., “Does PowerPoint make you stupid?“, Presentations, 18 (3). Retrieved on November 21, 2005 from http://global.factiva.com/

[2] Cassady, J. C., “Student and instructor perceptions of the efficacy of computer-aided lectures in undergraduate university courses“, Journal of Educational Computing Research, 19:175-89, 1998

[3] Kim, S. C., Jung, W. C., Han. K. J., Lee, J. G., Yi, M. Y., “Quality-based Automatic Classification for Presentation Slides“, Proceedings of the European Conference on Information Retrieval (ECIR), 2014

[4] Day, E. A., Arthur, W. J., & Gettman, D., “Knowledge structures and the acquision of a complex skill“, Journal of applied psychology, vol.86, no.5, 2001

[5] Kim, H. W., Yi, M. Y., “Empirical Validation of an Automated Method of Knowledge Structure Creation from Single Documents“, IEEE ICT&KE, 2011

[6] Schvaneveldt, R. W., Durso, F. T., and Dearholt, D. W., “Network structures in proximity data“, The psychology of learning and motivation, vol. 25, pp.249-294, 1989