ISSN 2383-6318(Print) / ISSN 2383-6326(Online)
KIISE Transactions on Computing Practices, Vol. 20, No. 12, pp. 676-681, 2014. 12
http://dx.doi.org/10.5626/KTCP.2014.20.12.676

고품질 슬라이드 선별을 위한 지식구조 기반 분류 기법

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

정 원 철 † 김 성 찬 † 이 문 용 ††
(Wonchul Jung) (Seongchan Kim) (Mun Y. Yi)

요 약 본 연구는 내용적으로 고품질인 슬라이드를 구분하고 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.

키워드: 지식구조, 정보품질, 슬라이드, 분류

Abstract In order to discern and classify high-quality slides, our research proposes a classification method that utilizes a knowledge structure containing information on the presentation slides. After analyzing whether our knowledge structure captures the content's quality information, we developed a classification method based on the knowledge structure produced from the analysis results. With the proposed method, we compared results classified by quality of presentation slides. Through this comparison, we verified that the slides in the high quality group could be classified and were able to retrieve high quality slides. The results show that, by utilizing the cognitive model of a knowledge structure, our method can increase the effectiveness of classification when search or recommendation is conducted mainly with high-quality slides.

Keywords: knowledge structure, information quality, presentation slides, classification

· 본 연구는 2014년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. NRF-2011-0024560)
· 이 논문은 2014 한국컴퓨터종합학술대회에서 '고품질 슬라이드 선별을 위한 지식구조 기반 분류 기법'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 한국과학기술원 지식서비스공학과
wonchul.jung@kaist.ac.kr
sckim@kaist.ac.kr
†† 종신회원 : 한국과학기술원 지식서비스공학과 교수
(KAIST)
munyi@kaist.ac.kr
(Corresponding author임)

논문접수 : 2014년 9월 4일
(Received 4 September 2014)
논문수정 : 2014년 10월 15일
(Revised 15 October 2014)
심사완료 : 2014년 10월 22일
(Accepted 22 October 2014)

정보과학회 컴퓨팅의 실제 논문지 제20권 제12호(2014. 12)

1. 서 론

오늘날 PowerPoint와 같은 컴퓨터 프로그램으로 만들어진 발표용 슬라이드(Presentation Slide)는 폭발적으로 늘어나는 추세이다[1]. 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 온라인 상에는 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]에서 학습자가 슬라이드의 품질을 측정할 때 고려한 상위 11개의 요소들 중에 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 연관관계를 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.

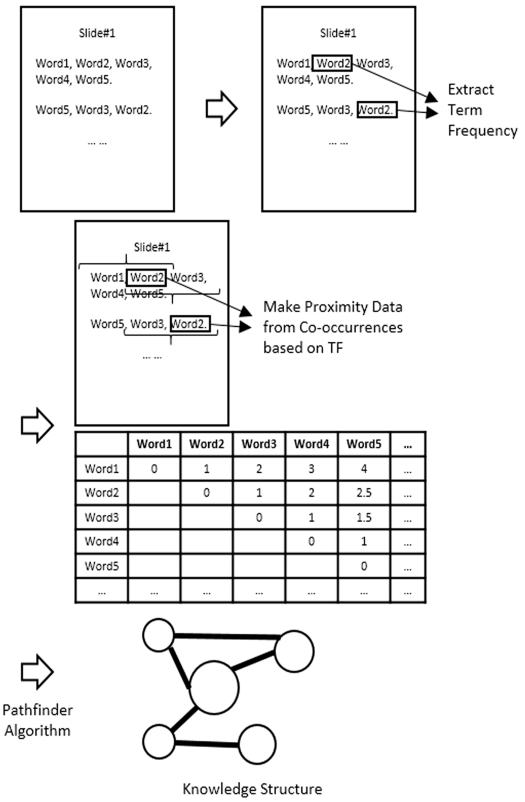
본 연구에서는 1) 슬라이드에서 추출한 지식구조가 슬라이드의 내용적 품질(Quality)을 잘 반영하는지 통계적 방법론으로 검증하고, 2) 슬라이드에서 지식구조를 활용한 문맥적 품질 분류 방법을 제안하고, 그에 따른 효과를 입증하며, 3) 제안한 분류 기법의 정밀도(Precision)를 측정하고자 한다.

2. 제안하는 방법

지식구조를 이용한 고품질 슬라이드 분류방법의 순서는 크게 1) 슬라이드로부터 지식구조 추출, 2) 지식구조를 활용한 슬라이드 분류과정으로 나눌 수 있다.

2.1 슬라이드로부터 지식구조 추출

슬라이드로부터 지식구조를 추출하는 방법은 Kim & Yi의 연구[5]에서 제시한 방법론을 바탕으로 슬라이드에 맞게 응용하였으며, 그림 1과 같은 과정으로 추출된다.



	Word1	Word2	Word3	Word4	Word5	...
Word1	0	1	2	3	4	...
Word2		0	1	2	2.5	...
Word3			0	1	1.5	...
Word4				0	1	...
Word5					0	...
...	...	...	...	...	...	...

그림 1 한 슬라이드에서 지식구조 추출 과정 순서도
Fig. 1 Flowchart of Extraction of the Knowledge Structure

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.

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

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:관

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

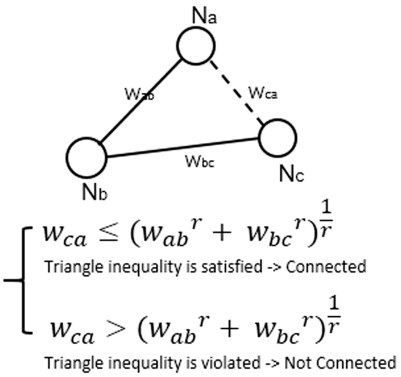


그림 2 Triangle Inequality의 원리
Fig. 2 Theorem of Triangle Inequality

련 있음, 7:관련 없음)로 변환한 후, 패스파인더 알고리즘[6]을 적용하여 각 핵심 개념 간 최단거리를 연결해주는 지식구조를 생성한다.

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

본 연구에서 지식구조를 무향 그래프로 표현하기 위해 사용한 패스파인더 알고리즘[6]은 그림 2와 같은 Triangle Inequality와 Supremum Distance를 사용한다.

Supremum Distance는 식 (3)의 Minkowski Distance에서 r값이 무한대로 발산할 때의 Distance이다. 식 (4)는 Supremum Distance를 정의한다.

$$w(P) = (w_1^r + w_2^r + ... + w_k^r)^{\frac{1}{r}} = (\sum_{i=1}^{k} w_i^r)^{\frac{1}{r}} \quad (3)$$

$$r \to \infty : w(P) = \lim_{r \to \infty} (w_1^r + w_2^r + ... + w_k^r)^{\frac{1}{r}} = Max(w_1, w_2, ..., w_k) \quad (4)$$

본 연구에서 사용된 Triangle Inequality의 원리는 과정 2)에서 설명한 방식으로, 그림 1의 매트릭스와 같이 Proximity Data를 생성하고, 생성된 다차원 공간에서 각 핵심개념 간의 최단거리를 계산할 때 사용되었다.

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

2.2 지식구조를 활용한 슬라이드 분류

그림 3과 같이, 모든 슬라이드들의 지식구조 정보를 추출하여 저장한 후, 다차원 지식구조 벡터 공간을 이용하여 고품질의 슬라이드를 선별한다. 본 연구에서는 슬라이드의 지식구조에서 하나의 핵심 개념을 분류 질의어(Query) 기준으로 사용하고, 핵심 개념과 연결된 다른 핵심 개념들과 각 연관관계들로부터 분류 질의어를 확장하며, 자세한 순서는 다음과 같다.

1) 하나의 지식구조에서 하나의 핵심 개념을 선택한다.

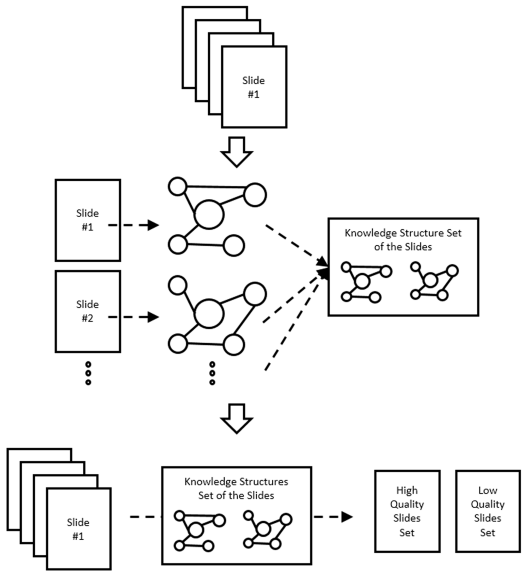


그림 3 분류 기법 전체 수행과정 순서도
Fig. 3 Flowchart of the Classification Method

2) 선택된 핵심 개념과 연결된 다른 핵심 개념들 및 각 연관관계들을 추출한다. 식 (5)는 연관관계를 추출하는 공식이며, 역수를 취하는 이유는 후의 과정에서 다차원 공간에 대입할 때 연관관계가 높을수록 가중치를 주기 위함이다.

$$L_{ij} = \frac{1}{D_{ij}} \times 7, (1 \le L \le 7) \quad (5)$$

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, 그림 4와 같이 다차원 벡터 공간을 나타내는 2차원 배열로 저장하며, 각 값은 지식구조에서의 연관관계 값이다.

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

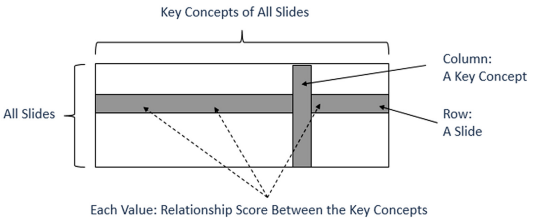


그림 4 지식구조 연관관계로 이루어진 다차원 벡터 공간
Fig. 4 Multidimensional Vector Space Consisting of Each Relationships from the Knowledge Structure

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. 2명 이상 같은 품질로 응답을 한 슬라이드들만 추렸고, 카파 값은 0.67이였으며, 총 884개의 슬라이드를 얻어냈다.

검수자들로부터 얻어낸 884개의 슬라이드 중 품질이 낮음인 슬라이드는 197개, 보통인 슬라이드는 427개, 높음인 슬라이드는 260개였으며, 슬라이드의 품질이 낮음은 0, 보통은 1, 높음은 2로 기록하였다.

884개의 슬라이드들로부터 각각의 지식구조를 생성하였고, 지식구조가 슬라이드의 품질정보를 내포하는지 알아보기 위해, 품질에 따라 분류한 후, 지식구조가 품질에 따라 다른 특성을 보이는지 관찰하였다. 그림 5와 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.

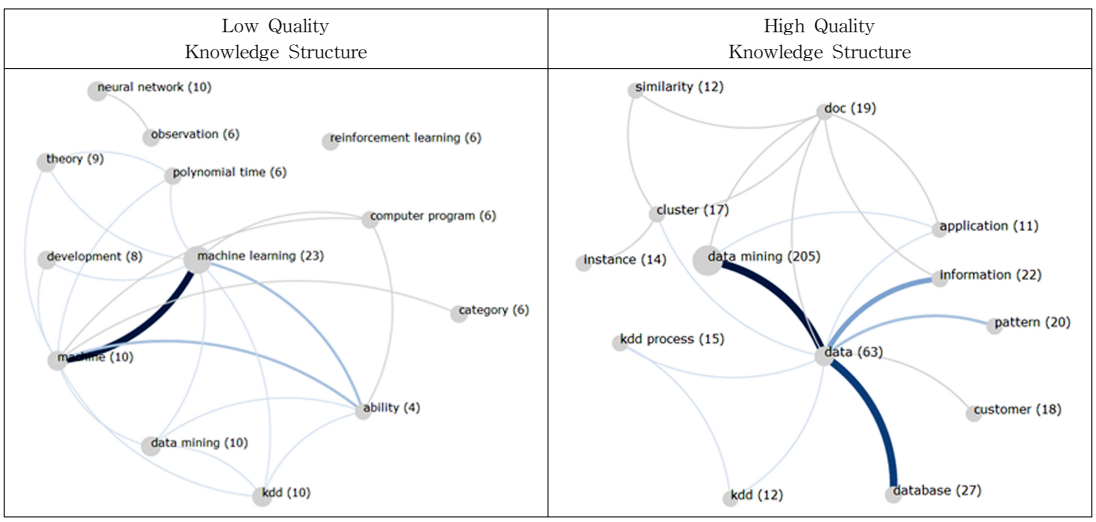
이를 통계적으로 검증하기 위해, 품질 낮음 집단(A, Low+Fair)과 높음 집단(B, High)으로 설정하고, t-검정(이분산 가정 두 집단)을 시행하였다. 두 집단 간 핵심 개념들의 발생빈도를 비교하기 위해 상위 5순위 핵심 개념까지의 발생빈도 평균을 비교해보았다. 또한, 두 집단 간 핵심 개념 간의 연관관계를 비교하기 위해 모든 연관관계의 평균을 비교해보았다.

표 1에서 보여주듯, 두 집단 간 핵심개념 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.

추가로, 고품질의 슬라이드에서 지식구조 기반 분류를 하면 높은 품질의 슬라이드들을 분류할 수 있는지에 대해 검증해보았다. 실험은 앞의 실험들과 동일한 884개의 슬라이드 환경에서 시행되었다. 각 슬라이드의 지식구조에서 높은 순위로 발생한 핵심 개념 순서로 2.3절에서 제안한 분류 방법을 적용하여 분류된 슬라이드들의 값들을 표 2와 같이 기록하였다.

고품질의 슬라이드에서 분류할수록, 분류된 슬라이드들이 높은 품질인지 알아보기 위한 실험을 위해, 가장 먼저 표 2와 같이 기록된 값에서, 핵심 개념 순위별로 분류하였다. 이는 동일한 발생빈도의 핵심 개념 기준을 가지고 각 분류결과를 비교해보기 위함이었다.

표 1 품질에 따른 지식구조의 핵심 개념 빈도 및 핵심 개념 간 연관관계 비교



*Node: Key Concept. The number is a term frequency of the key concept.
*Link: Relationship between key concepts. A thicker and darker line indicates that the relationship is stronger.

그림 5 품질에 따른 지식구조
Fig. 5 Knowledge Structure - Low vs. High Quality

Table 1 Comparison of Knowledge Structure Key Concepts' Frequencies and their Relationships between Two Quality Groups

Key Concepts' Frequency	Low+Fair (A) Keyword1~5	High (B) Keyword1~5
Mean	117.45833	175.91154
Variance	12653.82812	48088.30489
Observations	624	260
Degree of Freedom	317	
t Stat	-4.08022	
P(T<=t) one-tail	**0.00003**	

Relationships between the Concepts	Low+Fair (A)	High (B)
Mean	4.22155	4.36818
Variance	1.00784	0.69998
Observations	624	260
Degree of Freedom	577	
t Stat	-2.23413	
P(T<=t) one-tail	**0.01293**	

표 2 지식구조 기반 분류를 통해 분류된 값 기록 예
Table 2 the Example of the Results Classified by the Proposed Method

ID	Quality	Criterion of Classification	Ranking of Key Concept	Number of Result	Average of Quality
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: Slide ID
*Quality: Quality of the Slide(0:Low, 2:High)
*Criterion of Classification: Classification Query mentioned at Section 2.3
*Ranking of Key Concept(n): Ranking of the Frequency
*Number of Result: Number of the Classified Results
*Average of Quality: Average of Quality of Classified Results(0:Low, 2:High)

핵심 개념 순위를 5순위까지 분류한 후, 다시 분류의 기준이 되는 슬라이드의 품질 낮음(A), 보통(B), 높음(C) 세 집단으로 분류한 후, 표 3과 같이 기록하였다. 표 3과 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.

표 3에서 나온 결과들의 유의성을 알아보기 위해, 표 3 지식구조에서 중요도가 N번째인 핵심 개념을 기준으로 분류된 슬라이드들의 품질평균의 평균 및 분산

Table 3 Mean and Variance of Quality Average of the Classification Based on the N-th Ranked Key Concept of the Knowledge Structure

Ranking of Key Concept	Slide Quality	Mean of Quality Average	Variance of the Quality Average	Obser-vations
1	Low	0.84146	0.20735	7
	Fair	1.05981	0.13742	26
	High	1.19911	0.18718	19
2	Low	0.95526	0.36022	76
	Fair	1.05046	0.21394	167
	High	1.38020	0.31681	107
3	Low	0.87251	0.33469	62
	Fair	1.06671	0.21687	157
	High	1.40023	0.28230	94
4	Low	0.71449	0.35070	59
	Fair	1.06494	0.24719	151
	High	1.43973	0.32875	93
5	Low	0.81587	0.28196	60
	Fair	1.07507	0.22362	128
	High	1.44338	0.32380	89

본 수가 적은 결과를 제외하고, t-검정(이분산 가정 두 집단)을 시행하였다. 표 4와 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.

마지막으로, 제안한 분류기법의 정밀도를 동일한 884개의 슬라이드 환경에서 확인해 보았다. 고품질의 슬라이드에서 분류된 슬라이드들이 고품질인지 아닌지에 대해 평균 정밀도를 측정해보았다. 앞선 실험에서 표본 수가 적은 결과를 제외한, 2순위에서 5순위 핵심 개념 기준까지를 대상으로 실험하였으며, 그림 6과 같은 결과를 확인할 수 있었다.

고품질 슬라이드에서, 2순위 핵심 개념을 기준으로 한 분류 결과들의 평균 정밀도는 0.48, 3순위는 0.51, 4순위

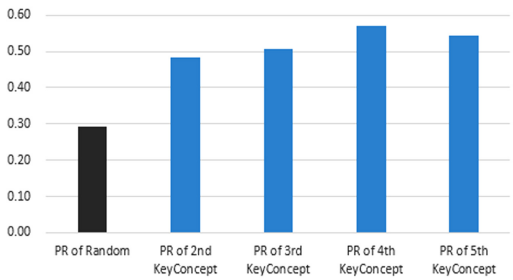


그림 6 평균 정밀도 비교
Fig. 6 Comparison of the Average Precision

표 4 핵심 개념 순위가 2~5번째로 높은 핵심 개념을 질의어 기준으로 분류한 결과들의 t-검정

Table 4 T-test Results from Classification Results Based on the Second to Fifth Ranked Key Concept

Second Key Concept				
	Low	Fair	Fair	High
Degree of Freedom	117		194	
t Stat	-1.22683		-5.06295	
P(T<=t) one-tail	0.11117		**0.00000**	

Third Key Concept				
	Low	Fair	Fair	High
Degree of Freedom	94		176	
t Stat	-2.35846		-5.03690	
P(T<=t) one-tail	**0.01021**		**0.00000**	

Fourth Key Concept				
	Low	Fair	Fair	High
Degree of Freedom	92		174	
t Stat	-4.02496		-5.21155	
P(T<=t) one-tail	**0.00006**		**0.00000**	

Fifth Key Concept				
	Low	Fair	Fair	High
Degree of Freedom	104		166	
t Stat	-3.22824		-5.01901	
P(T<=t) one-tail	**0.00083**		**0.00000**	

는 0.57, 5순위는 0.54로 측정되었다. 이는 임의로 슬라이드의 품질을 분류했을 시, 고품질의 슬라이드를 분류하는 경우의 정밀도인 0.29(260/(197+427+260))보다 모두 높게 측정되었으며, 최소 65.5%, 최대 96.6%의 향상율을 보이는 것을 확인하였다.

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.

References

[1] R. J. Craig, J. H. Amernic, "PowerPoint presentation technology and the dynamics of teaching," *Journal of Innovative Higher Education*, Vol. 31, No. 3, pp. 147-160, Aug. 2006.

[2] J. C. Cassady, "Student and instructor perceptions of the efficacy of computer-aided lectures in undergraduate university courses," *Journal of Educational Computing Research*, Vol. 19, No. 2, pp. 175-189, 1998.

[3] S. C. Kim, W. C. Jung, K. J. Han, J. G. Lee, M. Y. Yi, "Quality-based Automatic Classification for Presentation Slides," *Proc. of the European Conference on Information Retrieval 2014*, pp. 638-643, 2014.

[4] E. A. Day, W. Arthur Jr, D. Gettman, "Knowledge structures and the acquisition of a complex skill," *Journal of applied psychology*, Vol. 86, No. 5, pp. 1022-1033, Oct. 2001.

[5] H. W. Kim, M. Y. Yi, "Empirical Validation of an Automated Method of Knowledge Structure Creation from Single Documents," *Proc. of the 9th International Conference on ICT and Knowledge Engineering*, pp. 161-165, 2011.

[6] R. W. Schvaneveldt, F. T. Durso, D. W. Dearholt, "Network structures in proximity data," *The psychology of learning and motivation*, Vol. 24, pp. 249-284, Elsevier, Amsterdam, 1989.

정 원 철

2013년 한양대학교 정보시스템학과 학사. 2013년~현재 한국과학기술원 지식서비스공학과 석사과정. 관심분야는 정보검색, 텍스트 마이닝, 지식공학

김 성 찬

2004년 전북대학교 컴퓨터공학과 학사. 2010년 한국과학기술원 정보통신공학과 석사. 2010년~현재 한국과학기술원 지식서비스공학과 박사과정. 관심분야는 정보검색, Q&A시스템, 텍스트 마이닝

이 문 용

1998년 University of Maryland 정보시스템 박사학위. 1998년~2004년 University of South Carolina 조교수. 2005년~2009년 University of South Carolina 부교수 (tenured). 2009년~2013년 한국과학기술원 지식서비스공학과 부교수. 2013년~현재 한국과학기술원 지식서비스공학과 교수(tenured). 관심분야는 지식공학, 시맨틱 웹, 개인화, 사용자 경험