(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2020-0067341
(43) 공개일자 2020년06월12일

(51) 국제특허분류(Int. Cl.)
G06F 40/20 (2020.01)
(52) CPC특허분류
G06F 40/284 (2020.01)
G06F 40/216 (2020.01)
(21) 출원번호 10-2018-0154113
(22) 출원일자 2018년12월04일
심사청구일자 2018년12월04일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
이원규
서울특별시 성북구 종암로9길 71, 208동 1202호 (종암동, 종암2차아이파크)
우호성
경기도 김포시 걸포1로 40, 102동 903호 (걸포동, 한강파크뷰 우방아이유쉘)
김자미
서울특별시 용산구 한남대로46길 18, 501호 (한남동, 주노빌)
(74) 대리인
김홍석

전체 청구항 수 : 총 8 항

(54) 발명의 명칭 **컴퓨터과학 교육과정 상의 전문용어 추출 방법**

(57) 요 약

본 발명은 전문용어 추출 방법에 관한 것으로서, 보다 상세하게는 컴퓨터과학 표준 커리큘럼에 기반한 지식영역과 지식 세부영역과 단계별 주제들이 계층적 구조를 이루는 전문용어 계층화 사전에 전문용어를 등록하여 각 전문용어들에 계층적 클러스터링 구조 내의 위치를 부여한 후, 분석하고자 하는 문헌에서 추출된 전문용어를 전문용어 계층화 사전에서 검색하여 검색된 전문용어의 지식영역 상의 위치와 수준을 파악할 수 있게 함으로써, 컴퓨터과학 교육에 대한 글로벌 트렌드를 간편하고 객관적으로 파악하여 국내 컴퓨터과학 교육과정의 개선방향 수립에 기여할 수 있게 한 컴퓨터과학 교육과정 상의 전문용어 추출 방법에 관한 것이다.

대 표 도

(52) CPC특허분류
G06F 40/258 (2020.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 2016R1A2B4014471
부처명 한국연구재단
연구관리전문기관 한국연구재단
연구사업명 이공분야기초연구사업,중견연구자지원사업,중견연구(총연구비1.5억초과~3억이하)체계적인 정보교육과정 기반의 교육평가 및 교원양성체제 지원시스템 개발
연구과제명 체계적인 정보교육과정 기반의 교육평가 및 교원양성체제 지원시스템 개발
기 여 율 1/1
주관기관 고려대학교 산학협력단
연구기간 2018.04.01 ~ 2019.02.28

명 세 서

청구범위

청구항 1

컴퓨터과학 교육과정에 기반한 다수의 지식영역을 계층적 클러스터링 구조로 재구성하여 전문용어 계층화 사전을 구축함으로써, 대상문헌에서 추출되는 전문용어의 지식영역상 위치를 계층적으로 분석할 수 있게 한 전문용어 계층화 사전 구축단계;

전문용어를 추출하고자 하는 대상문헌 상의 문장을 분석하기 쉬운 형태로 가공하여 상기 전문용어 계층화 사전의 기본형으로 정규화시키는 대상문헌 전처리단계;

전처리가 완료된 문장의 단어들을 인접한 단어들과 묶어 의미 파악이 가능한 구문의 형태로 구조화하는 데이터 구조화단계;

구문의 형태로 구조화가 이루어진 데이터들을 상기 전문용어 계층화 사전에 등록되어 있는 데이터와 비교하여 대상문헌 상에 존재하는 전문용어를 추출하고, 해당 전문용어의 상기 전문용어 계층화 사전상 위치를 파악하여 지식영역 상의 위치와 수준을 파악하는 구조화 데이터 분석단계; 및

데이터 분석결과 대상문헌에서 추출된 전문용어와 그 전문용어의 지식영역상 위치를 그래프 또는 텍스트 형태로 제시하고 저장하는 결과요약단계;를 포함하는 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

청구항 2

제1항에 있어서,

상기 전문용어 계층화 사전 구축단계는, 컴퓨터과학 교육과정에 포함되어 있는 다수의 지식영역과, 각 지식영역에 포함되어 있는 지식 세부영역과, 각 지식 세부영역 항목에서 취급하는 내용의 수준을 단계별 주제로 나누고, 각 전문용어들의 사전내 위치를 지시영역, 지식 세부영역 및 단계별 주제에 따라 부여하여 저장함으로써 계층적 클러스터링 구조의 전문용어 계층화 사전을 구축하도록 구성되는 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

청구항 3

제2항에 있어서,

상기 전문용어 계층화 사전 구축단계는, 상기 각 지식 세부영역에서 취급하는 주제들을 입문 수준의 기본 개념을 다루고 있는 단계1(Tier1)과, 전공자 수준의 개념을 다루고 있는 단계2(Tier2) 및 전공자 수준보다 심화된 내용을 다루는 선택단계(Elective)로 이루어진 3단계로 구분하고, 각 단계에서 사용되는 전문용어와 주제의 내용을 등록하여 저장하도록 구성되는 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

청구항 4

제2항에 있어서,

상기 대상문헌 전처리단계는,

전문용어를 추출하여 분석하고자 하는 대상문헌에 포함되어 있는 단어들을 토큰으로 분할하는 토큰화 과정;

분할된 토큰에 포함되어 있는 특수문자들을 추출대상에서 제거하는 특수문자 제거과정;

의미에 영향을 주지 않는 하나의 음절로 표시된 알파벳을 제거하는 1 음절 제거과정;

대상문헌에 포함되어 있는 모든 단어를 영문 소문자로 변환하는 소문자 정규화 과정; 및

소문자로 정규화된 단어들의 형태를 자연언어 처리하여 정규화하는 자연언어 처리과정;을 포함하는 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

청구항 5

제4항에 있어서,

상기 자연언어 처리과정은 품사 정보가 보존된 형태의 기본형으로 단어를 변환시켜 정규화하는 표제어 추출(Lemmatization) 처리에 의해 이루어지는 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

청구항 6

제4항에 있어서,

상기 데이터 구조화단계는, 전처리가 완료된 문장의 단어들을 1어절 단위나, 연속적인 2어절의 묶음 또는 연속적인 3어절의 묶음으로 구문을 생성하여, 상기 전문용어 계층화 사전에 등록되어 있는 전문용어와의 일치 여부를 판단하기 위한 구문의 형태를 생성하도록 구성되는 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

청구항 7

제6항에 있어서,

상기 데이터 구조화단계는, 상기 전문용어 계층화 사전에 등록되어 있는 전문용어와의 일치 여부를 판단하기 위한 구문을 생성하는 어절의 개수를 증감시켜 새로이 추가되는 전문용어에 대한 추출의 정확도를 향상시킬 수 있게 한 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

청구항 8

제4항에 있어서,

상기 구조화 데이터 분석단계는, 구문 단위로 구조화된 데이터가 상기 전문용어 계층화 사전에서 검색되면, 해당 데이터가 검색된 위치의 인덱스를 증가시켜, 어느 지식영역에 해당하는지, 어느 지식 세부영역에 해당하는지, 그리고 어느 주제에 관한 것인지를 체크할 수 있게 한 것을 특징으로 하는 컴퓨터과학 교육과정 상의 전문용어 추출 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 전문용어 추출 방법에 관한 것으로서, 보다 상세하게는 컴퓨터과학 표준 커리큘럼에 기반한 지식영역과 지식 세부영역과 단계별 주제들이 계층적 구조를 이루는 전문용어 계층화 사전에 전문용어를 등록하여 각 전문용어들에 계층적 클러스터링 구조 내의 위치를 부여한 후, 분석하고자 하는 문헌에서 추출된 전문용어를 전문용어 계층화 사전에서 검색하여 검색된 전문용어의 지식영역 상의 위치와 수준을 파악할 수 있게 함으로써, 컴퓨터과학 교육에 대한 글로벌 트렌드를 간편하고 객관적으로 파악하여 국내 컴퓨터과학 교육과정의 개선방향 수립에 기여할 수 있게 한 컴퓨터과학 교육과정 상의 전문용어 추출 방법에 관한 것이다.

배 경 기 술

[0002] 최근 급변하는 기술로 인한 사회 변화에 대처하기 위해 초중등 컴퓨터과학교육을 강화함으로써, 국가 경쟁력에 대한 기반을 마련할 수 있게 하려는 시도들이 여러 나라에서 진행되고 있다.

[0003] 그에 따라, 2000년대 초반부터 초중등 교육에서 컴퓨터과학(CS :Computer Science)의 개념과 원리를 가르치기 위한 교육과정의 개편이 시작되었으며, 2013년을 기점으로 인도, 영국, 일본, 미국 그리고 한국에 이르기까지 기술변화를 반영하기 위한 개편이 진행되고 있다.

[0004] 이처럼 초중등 교육에서 이루어지는 컴퓨터과학에 대한 교육 강화는 IT분야의 전문 인력 양성에 집중하는 것이 아니라, 미래의 구성원들이 갖추어야 할 기본 역량 배양의 관점으로 접근하고 있는바, 미래적 삶의 가치를 실현하기 위해 어떠한 역량을 균형있게 갖추어야 하는지에 대한 내용을 토대로 교육과정을 구축해야 한다.

[0005] 이러한 컴퓨터과학 분야의 교육과정은 각 나라마다 상이하게 구성되곤 하는데, 이는 동일한 학문분야임에도 불구하고 역사성, 지역성이 다르며 문화적 맥락의 차이가 있기 때문이라 할 수 있다. 따라서 교육과정에 대한 글

로벌 트렌드를 이해하여 교육과정을 구성하기 위해서는 각 국의 교육과정을 분석할 필요가 있게 된다.

[0006] 그러나, 일반적으로 교육과정에 적용되는 프레임도 학문 사조에 따라 다르며, 그 내용이 되는 지식의 추출에 대해서는 합의된 결과가 미흡한 실정인바, 특정한 프레임이 있다고 하더라도 지식의 추출과정에서 소수의 전문가나 교수자의 주관적 요소가 작용할 여지가 높아지는 문제점이 있었다. 그에 따라 보다 객관적인 지식의 추출을 위해서는 표준화된 전문용어에 기반할 필요가 있게 된다.

[0007] 이를 위하여 국내뿐만 아니라 세계 각국의 컴퓨터과학 교육에 반영되고 있는 글로벌 트렌드를 파악하는데 도움을 줄 수 있도록, 컴퓨터과학 분야의 교육과정에 포함되어 있는 전문용어를 객관적으로 추출할 수 있는 새로운 수단에 대한 필요성은 점차 높아지고 있다 할 것이다.

선행기술문헌

특허문헌

[0008] (특허문헌 0001) 대한민국 등록특허 제10-1061391호

(특허문헌 0002) 대한민국 공개특허 제10-2016-0040083호

발명의 내용

해결하려는 과제

[0009] 본 발명은, 컴퓨터과학 표준 커리큘럼에 기반한 지식영역과 지식 세부영역과 단계별 주제들이 계층적 구조를 이루는 전문용어 계층화 사전에 전문용어를 등록하여 각 전문용어들에 계층적 클러스터링 구조 내의 위치를 부여한 후, 분석하고자 하는 문헌에서 추출된 전문용어를 전문용어 계층화 사전에서 검색하여 검색된 전문용어의 지식영역 상의 위치와 수준을 파악할 수 있게 함으로써, 컴퓨터과학 교육에 대한 글로벌 트렌드를 간편하고 객관적으로 파악하여 국내 컴퓨터과학 교육과정의 개선방향 수립에 기여할 수 있게 한 컴퓨터과학 교육과정 상의 전문용어 추출 방법을 제공하는 것을 과제로 한다.

과제의 해결 수단

[0010] 상기 과제를 해결하기 위한 컴퓨터과학 교육과정 상의 전문용어 추출 방법은,

[0011] 컴퓨터과학 교육과정에 기반한 다수의 지식영역을 계층적 클러스터링 구조로 재구성하여 전문용어 계층화 사전을 구축함으로써, 대상문헌에서 추출되는 전문용어의 지식영역상 위치를 계층적으로 분석할 수 있게 한 전문용어 계층화 사전 구축단계; 전문용어를 추출하고자 하는 대상문헌 상의 문장을 분석하기 쉬운 형태로 가공하여 상기 전문용어 계층화 사전의 기본형으로 정규화시키는 대상문헌 전처리단계; 전처리가 완료된 문장의 단어들을 인접한 단어들과 묶어 의미 파악이 가능한 구문의 형태로 구조화하는 데이터 구조화단계; 구문의 형태로 구조화가 이루어진 데이터들을 상기 전문용어 계층화 사전에 등록되어 있는 데이터와 비교하여 대상문헌 상에 존재하는 전문용어를 추출하고, 해당 전문용어의 상기 전문용어 계층화 사전상 위치를 파악하여 지식영역 상의 위치와 수준을 파악하는 구조화 데이터 분석단계; 및 데이터 분석결과 대상문헌에서 추출된 전문용어와 그 전문용어의 지식영역상 위치를 그래프 또는 텍스트 형태로 제시하고 저장하는 결과요약단계;를 포함하는 것을 특징으로 한다.

[0012] 이때, 상기 전문용어 계층화 사전 구축단계는, 컴퓨터과학 교육과정에 포함되어 있는 다수의 지식영역과, 각 지식영역에 포함되어 있는 지식 세부영역과, 각 지식 세부영역 항목에서 취급하는 내용의 수준을 단계별 주제로 나누고, 각 전문용어들의 사전내 위치를 지시영역, 지식 세부영역 및 단계별 주제에 따라 부여하여 저장함으로써 계층적 클러스터링 구조의 전문용어 계층화 사전을 구축하도록 구성되는 것을 특징으로 한다.

[0013] 또한, 상기 전문용어 계층화 사전 구축단계는, 상기 각 지식 세부영역에서 취급하는 주제들을 입문 수준의 기본 개념을 다루고 있는 단계1(Tier1)과, 전공자 수준의 개념을 다루고 있는 단계2(Tier2) 및 전공자 수준보다 심화된 내용을 다루는 선택단계(Elective)로 이루어진 3단계로 구분하고, 각 단계에서 사용되는 전문용어와 주제의 내용을 등록하여 저장하도록 구성되는 것을 특징으로 한다.

[0014] 또한, 상기 대상문헌 전처리단계는,

[0015] 전문용어를 추출하여 분석하고자 하는 대상문헌에 포함되어 있는 단어들을 토큰으로 분할하는 토큰화 과정; 분할된 토큰에 포함되어 있는 특수문자들을 추출대상에서 제거하는 특수문자 제거과정; 의미에 영향을 주지 않는 하나의 음절로 표시된 알파벳을 제거하는 1 음절 제거과정; 대상문헌에 포함되어 있는 모든 단어를 영문 소문자로 변환하는 소문자 정규화 과정; 및 소문자로 정규화된 단어들의 형태를 자연언어 처리하여 정규화하는 자연언어 처리과정;을 포함하는 것을 특징으로 한다.

[0016] 이때, 상기 자연언어 처리과정은 품사 정보가 보존된 형태의 기본형으로 단어를 변환시켜 정규화하는 표제어 추출(Lemmatization) 처리에 의해 이루어지는 것을 특징으로 한다.

[0017] 또한, 상기 데이터 구조화단계는, 전처리가 완료된 문장의 단어들을 1어절 단위나, 연속적인 2어절의 묶음 또는 연속적인 3어절의 묶음으로 구문을 생성하여, 상기 전문용어 계층화 사전에 등록되어 있는 전문용어와의 일치 여부를 판단하기 위한 구문의 형태를 생성하도록 구성되는 것을 특징으로 한다.

[0018] 또한, 상기 데이터 구조화단계는, 상기 전문용어 계층화 사전에 등록되어 있는 전문용어와의 일치 여부를 판단하기 위한 구문을 생성하는 어절의 개수를 증감시켜 새로이 추가되는 전문용어에 대한 추출의 정확도를 향상시킬 수 있게 한 것을 특징으로 한다.

[0019] 또한, 상기 구조화 데이터 분석단계는, 구문 단위로 구조화된 데이터가 상기 전문용어 계층화 사전에서 검색되면, 해당 데이터가 검색된 위치의 인덱스를 증가시켜, 어느 지식영역에 해당하는지, 어느 지식 세부영역에 해당하는지, 그리고 어느 주제에 관한 것인지를 체크할 수 있게 한 것을 특징으로 한다.

발명의 효과

[0020] 본 발명은 국내외에서 컴퓨터과학 교육에 사용되는 문헌으로부터 추출된 전문용어를 전문용어 계층화 사전에서 검색하여 검색된 전문용어의 지식영역 상의 위치와 수준을 파악할 수 있게 함으로써, 컴퓨터과학 교육에 대한 글로벌 트렌드를 간편하고 객관적으로 파악하여 국내 컴퓨터과학 교육과정의 개선방향 수립에 기여할 수 있는 효과가 있다.

도면의 간단한 설명

[0021] 도 1은 본 발명에 따른 컴퓨터과학 교육과정 상의 전문용어 추출 방법의 구성도.

도 2는 본 발명에 따라 전문용어 추출과 분석이 이루어지는 전체 시스템을 나타내는 블록 구성도.

도 3은 본 발명에 따라 CS2013에 기반하여 계층적 클러스터링 구조로 재구성된 컴퓨터과학 교육과정 상의 전문용어 계층화 사전의 예시도.

도 4는 본 발명에 따라 분석하고자 하는 교육과정 문서를 분석하기 쉬운 형태로 가공하여 전처리하는 것을 나타내는 예시도.

도 5는 본 발명에 따라 전처리된 내용을 구문의 형태로 구조화하는 것을 나태는 예시도.

도 6은 본 발명에 따른 전문용서 계층화 사전에 근거하여 구조화된 데이터를 매칭시켜 분석하는 알고리즘을 나타내는 예시도.

도 7은 본 발명에 따라 교유과정 문서에서 추출된 전문용어의 추출 결과를 나타내는 예시도.

발명을 실시하기 위한 구체적인 내용

[0022] 이하에서는 본 발명의 구체적인 실시예를 도면을 참조하여 상세히 설명하도록 한다.

[0023] 도 1은 본 발명에 따른 컴퓨터과학 교육과정 상의 전문용어 추출 방법의 구성도이고, 도 2는 본 발명에 따라 전문용어 추출과 분석이 이루어지는 전체 시스템을 나타내는 블록 구성도이다.

[0024] 도 1 및 도 2를 참조하면, 본 발명에 따른 컴퓨터과학 교육과정 상의 전문용어 추출 방법은, 컴퓨터과학 교육과정에 기반한 다수의 지식영역을 계층적 클러스터링 구조로 재구성하여 전문용어 계층화 사전을 구축함으로써 대상문헌에서 추출되는 전문용어의 지식영역상 위치를 계층적으로 분석할 수 있게 한 전문용어 계층화 사전 구축단계(S100)와, 전문용어를 추출하고자 하는 대상문헌 상의 문장을 분석하기 쉬운 형태로 가공하여 상기 전문용어 계층화 사전의 기본형으로 정규화시키는 대상문헌 전처리단계(S200)와, 전처리가 완료된 문장의 주제어들을 인접한 주제어들과 묶어 의미 파악이 가능한 구문의 형태로 구조화하는 데이터 구조화단계(S300)와, 구문의 형

태로 구조화가 이루어진 데이터들을 상기 전문용어 계층화 사전에 등록되어 있는 데이터와 비교하여 대상문헌상에 존재하는 전문용어를 추출하고 해당 전문용어의 상기 전문용어 계층화 사전상 위치를 파악하여 지식영역상의 위치와 수준을 파악하는 구조화 데이터 분석단계(S400)와, 데이터 분석결과 대상문헌에서 추출된 전문용어와 그 전문용어의 지식영역상 위치를 그래프 또는 텍스트 형태로 제시하고 저장하는 결과요약단계(S500)를 포함하여 구성된다.

[0025] 본 발명은 국내뿐만 아니라 해외 여러 나라에서 컴퓨터과학 교육에 사용되는 문헌들에서 사용되고 있는 전문용어들을 추출하여 그 전문용어들이 컴퓨터과학 교육과정상의 어느 지식영역에 관한 것들인지를 분석함으로써, 해외 여러 나라에서 이루어지는 컴퓨터과학 교육의 글로벌 트렌드를 간편하고 객관적으로 파악할 수 있게 하여 국내 컴퓨터과학 교육과정의 개선방향을 수립함에 도움을 주기 위한 것이다.

[0026] 이를 위해, 컴퓨터과학 교육에 사용되는 문헌들에서 추출된 전문용어들의 개별적인 의미보다는 그 전문용어가 컴퓨터과학 교육과정을 이루고 있는 다수의 지식영역들 중 어느 영역에 대한 교육을 위해 어느 정도의 수준으로 사용되고 있는지를 파악함이 중요하게 된다.

[0027] 일반적으로 전문용어는 전문화된 지식에 대한 언어적 표현으로서 신속하고 명확한 의미의 전달을 위해 사용되는 것으로서, 특정 분야에서 사용되는 용어인 만큼 해당 분야의 교육과정을 구성함에도 사용될 수 있다. 그리고, 교육과정의 구성에서 학습자의 발달단계를 고려한 연계성 증진의 측면을 고려하면 전문용어는 교육내용을 선정함에도 영향을 미치게 된다. 또한, 선정된 교육내용의 수준을 어떻게 규정할 것인지에 대한 부분 역시 전문용어와 밀접한 관련을 갖게 된다.

[0028] 동일한 전문용어를 교육내용으로 취급하였다 하더라도 학습내용의 성격에 따라 학습자들의 교육 경험은 달라질 것이기 때문에, 교육내용의 수준, 범위의 측면에서 교육내용으로 취급해야 하는 전문용어의 선정이 중요한 의미를 갖게 된다.

[0029] 따라서, 교육과정에서 제시되고 있는 전문용어를 분석하는 것은 교육과정의고유한 교육적 가치를 조명하고 시대적 패러다임을 파악함에 도움을 주게 되는바, 이처럼 전문용어에 기반하여 여러 나라의 교육과정을 분석할 수 있다면 시대적 흐름에 적합한 교육과정을 구현함에도 도움을 줄 수 있을 것이다.

[0030] 상기 전문용어 계층화 사전 구축단계(S100)는, 컴퓨터과학 교육과정에 사용되는 문헌에 포함되어 있는 전문용어가 컴퓨터과학 교육의 어느 지식영역에 관한 것이며, 어떠한 수준으로 취급되고 있는지 여부를 판단할 수 있는 전문용어 계층화 사전을 구축하기 위한 것으로서, 컴퓨터과학 교육과정에 포함되어 있는 다수의 지식영역들을 계층적 클러스터링 구조로 재구성하고, 각 전문용어들에 계층적 클러스터링 구조내 위치를 부여함으로써 전문용어의 지식영영상의 위치를 계층화한 사전을 구축하도록 구성된다.

[0031] 종래에도 다양한 분야의 전문용어 사전이 사용되고 있었던 만큼 이미 개발된 전문용어 관련 오픈 라이브러리나 온톨리지가 존재함은 물론이다. 그러나, 이러한 종래의 전문용어 사전들의 경우 해당 용어의 기술적 의미와 학문적 의미를 찾아보기에는 유용하지만, 학문적 관점과 초중등 학교와의 연계성을 고려하여 교육과정을 구축함에 도움을 주기에는 한계가 있었다.

[0032] 그에 따라, 상기 전문용어 계층화 사전 구축단계(S100)에서는, 컴퓨터과학과 관련하여 각 초증등 학교급에서 이루어지는 교육과정에 포함될 수 있는 전문용어에 대한 고려가 반영될 수 있도록, 컴퓨터과학 분야의 교육과정 표준으로 제시되는 지식영역들이 계층적인 구조를 갖도록 사전을 구축하였다. 이를 위해 상기 전문용어 계층화 사전 구축단계(S100)에서는 현시점에서 컴퓨터과학(CS) 분야의 전문용어를 가장 잘 반영하고 있는 컴퓨터과학 표준 커리큘럼(이하, ‘컴퓨터과학 교육과정’과 동일한 의미로 사용한다.) 인 CS2013을 토대로 컴퓨터과학 교육분야의 계층적 구조를 구축하는 것이 바람직하다. 이후 새로운 전문용어들이 추가 반영된 컴퓨터과학 표준 커리큘럼이 제시된다면, 그에 따라 전문용어 계층화 사전의 계층적 구조를 갱신할 수 있음은 물론이다.

[0033] 이러한 상기 전문용어 계층화 사전 구축단계(S100)는, 먼저 컴퓨터과학 교육과정에 포함되어 있는 다수의 지식영역과, 각 지식영역에 포함되어 있는 지식 세부영역과, 각 지식 세부영역 항목에서 취급하는 내용의 수준을 나눈 단계별 주제로 나누고, 각 전문용어들의 사전내 위치를 지시영역, 지식 세부영역 및 단계별 주제에 따라 부여하여 저장함으로써 계층적 클러스터링 구조의 전문용어 계층화 사전을 구축하도록 구성된다.

[0034] 이와 같이 계층적 클러스터링 구조의 전문용어 계층화 사전을 구축함으로써, 분석하고자 하는 대상인 문헌에 포함되어 있는 전문용어를 추출하여 전문용어 계층화 사전에 매치시키는 것만으로도, 그 분석 결과가 컴퓨터과학의 어느 분야에 대해서 어떤 수준으로 사용되고 있는지를 보다 명확하게 파악할 수 있게 된다.

[0035] 이를 위하여, 상기 전문용어 계층화 사전 구축단계(S100)에서는, 도 3에 도시된 바와 같이, 먼저 컴퓨터과학 표준 커리큘럼인 CS2013의 지식체계를 18개의 지식영역과, 각 지식영역에 포함되어 있는 지식 세부영역과, 각 지식 세부영역에서 취급하는 주제들로 구분한다.

[0036] 이때, 컴퓨터과학 표준 커리큘럼인 CS2013은 하기의 표 1 및 도 3에 나타난 바와 같이 18개의 지식영역으로 구성되어 있는바, 이러한 18개의 지식영역들이 모두 계층적 클러스터링 구조의 지식영역으로 포함될 수 있도록 지식영역의 항목들을 설정한다.

표 1

[0037]

지식영역(Knowledge Area)
알고리즘과 복잡도(AL : Algorithm and Complexity)
구조와 조직(AR : Architecture and Organization)
컴퓨팅 과학(CN : Computational Science)
이산 구조(DS : Discrete Structures)
그래픽과 시각화(GV : Graphics and Visualization)
인간-컴퓨터 상호작용(HCI : Human-Computer Interaction)
정보 보호와 보안(IAS : Information Assurance and Security)
정보관리(IM : Information Management)
지능형 시스템(IS : Intelligent Systems)
네트워크와 통신(NC : Networking and Communication)
운영체제(OS : Operating System)
플랫폼 기반 개발(PBD : Platform-based Development)
병렬 및 분산 컴퓨팅(PD : Parallel and Distributed Computing)
프로그래밍 언어(PL : Programming Language)
소프트웨어 개발 기초(SDF : Software Development Fundamentals)
소프트웨어 공학(SE : Software Engineering)
시스템 기초(SF : Systems Fundamentals)
사회적 이슈와 전문적 실현(SP : Social Issue and Professional Practice)

[0038] 그리고, 각 지식영역에는 CS2013에 나타나 있는 각 지식영역별 지식 세부영역(Detail of Knowledge Area)을 매치시켜 계층적 구조를 형성하도록 구성된다. 도 3에서는 이러한 지식 세부영역의 예로서, 그래픽과 시각화(GV) 지식영역에 Fundamental Concepts, Basic Rendering, Geometric Modeling, Advanced Rendering, Computer Animation, Visualization의 지식 세부영역들이 위치하도록 구성되는 것을 나타내었으며, 다른 지식영역들의 경우에도 이와 같이 각 지식영역마다 지식 세부영역들이 계층적 구조를 이루도록 구축된다.

[0039] 또한, 상기 컴퓨터과학 표준 커리큘럼인 CS2013에서 제시되는 내용 수준을 3단계로 구분한 후, 상기 각 지식 세부영역에서 취급하는 주제들을 3단계로 구분하여 각각 매치시키게 된다. 이때, 주제들은 입문 수준의 기본 개념을 다루고 있는 단계1(Tier1)과, 전공자 수준의 개념을 다루고 있는 단계2(Tier2) 및 전공자 수준보다 심화된 내용을 다루는 선택단계(Elective)로 구분하여, 각 단계에서 사용되는 전문용어(Technical Terms)와, 주제의 내용(Original Topic)을 매치시켜 저장하게 된다.

[0040] 그에 따라, 추후 상기 구조화 데이터 분석단계(S400)에서 구문의 형태로 구조화가 이루어진 데이터들을 Tier1, Tier2, 또는 Elective 단계에 포함되어 있는 전문용어(Technical Terms)들과 비교하여, 일치하는 전문용어가 위치하는 지식영역과 지식 세부영역과 단계를 해당 전문용어의 지식영역상 위치와 수준으로 파악할 수 있게 된다.

[0041] 상기 대상문헌 전처리단계(S200)는, 상기 전문용어 계층화 사전을 기반으로 하여 분석하고자 하는 대상문헌에 포함되어 있는 전문용어를 추출하여, 어느 지식영역의 교육을 위해 어떠한 수준으로 사용되고 있는지 여부를 판단하기 위한 것으로서, 대상문헌상의 문장을 분석하기 쉬운 형태로 가공하고 정규화시킬 수 있도록 구성된다.

[0042] 이러한 대상문헌 전처리단계(S200)는, 도 1에 도시된 바와 같이, 전문용어를 추출하여 분석하고자 하는 대상문헌에 포함되어 있는 단어들을 토큰으로 분할하는 토큰화 과정(S210)과, 용어간 매칭율과 정밀도를 높이기 위해 분할된 토큰에 포함되어 있는 특수문자들을 추출대상에서 제거하는 특수문자 제거과정(S220)과, 의미에 영향을 주지 않는 하나의 음절로 표시된 알파벳을 제거하는 1 음절 제거과정(S230)과, 대상문헌에 포함되어 있는 모든 단어를 영문 소문자로 변환하는 소문자 정규화 과정(S240)과, 소문자로 정규화된 단어들의 형태를 자연언어 처리하여 정규화하는 자연언어 처리과정(S250)을 포함하여 구성된다.

[0043] 그에 따라, 도 4에 도시된 바와 같이, 먼저 전문용어를 추출하여 분석하고자 하는 대상문헌의 문장(주제어)을 음절단위의 토큰으로 분할하게 된다. 이때, 도 4에서는 토큰화를 통해 분할된 단어들을 '/' 으로 구분하여 나타내었다. 이후, 분할된 토큰에 포함되어 있는 특수문자인 ',' 와 '.' 를 제거하여 영문 알파벳만이 분석의 대상이 될 수 있게 한다. 그리고, 하나의 음절로만 표시된 알파벳인 'a' 를 제거하여 의미를 갖는 단어들만이 추출대상으로 포함될 수 있게 한다. 이후, 남아 있는 모든 단어들을 소문자로 변환하여 문장 내의 위치에 따른 대소문자 차이를 제거한 후, 자연언어 처리과정(S250)에 의해 상기 전문용어 계층화 사전에서 전문용어를 저장한 것과 동일한 형태로 정규화시켜 용어간 매칭률과 검색의 정밀도를 높이게 한다.

[0044] 일반적으로 자연언어는 일상적으로 사용하는 언어로서, 컴퓨터에서 프로그램을 작성하는데 사용하는 언어나 기계어 등과는 대치되는 개념이다. 이러한 자연언어를 대상으로 하여 컴퓨터로 정보를 추출하거나 텍스트를 요약하거나, 기계번역하거나 새로운 지식을 추출하는 등의 작업하기 위해서는 일정한 처리를 하여 정규화시켜야 한다. 이처럼 자연언어로 구성된 문서를 처리하기 위해서는 다양한 방법들이 이루어지고 있으나, 그 중 영어 문서 처리의 정규화에 주로 사용되는 방식으로 표제어 추출(Lemmatization) 처리와 어간 추출(Stemming) 처리가 이용된다.

[0045] 이때, Lemmatization은 다양한 형태로 활용된(infected) 단어의 표제어(lemma)를 찾을 때 사용하는 기술로서, 사전에서 단어의 뜻을 찾을 때 사용하는 기본형으로 단어를 변환시켜 단어의 원형을 추출하는 방식으로 처리가 이루어진다. 따라서 정확한 단어가 필요한 기계 번역 등에 주로 사용되는 방식이지만, 변환 시간이 오래 걸리게 되는 단점이 있게 된다.

[0046] 또한, Stemming은 어형이 변형된 단어로부터 접사 등을 제거하고 그 단어의 어간을 분리해내는 기술로서, 어근과 차이가 있더라도 관련 있는 단어들이 일정하게 동일한 어간으로 맵핑될 수 있게 한다. 이러한 Stemming은 색인의 크기를 줄여주기 때문에 시간적인 효율성을 필요로 하는 검색엔진에 주로 사용되는 방식이지만, 단어로부터 분리된 어간은 의미가 불명확해질 수도 있는 단점이 있게 된다.

[0047] 이러한 Lemmatization 처리와 Stemming 처리시에 변환되는 단어의 예를 하기의 표 2에 나타내었다.

표 2

[0048]

	Word	Stemming	Lemmatization
1	Innovation	Innovat	Innovation
2	Innovations	Innovat	Innovation
3	Innovate	Innovat	Innovate
4	Innovates	Innovat	Innovate
5	Innovative	Innovat	Innovative

[0049] 상기 표 2에 나타난 바와 같이 Stemming 처리는 단어 자체만을 고려하여 축약형으로 변환해주는 것이므로 어간을 추출할 수 있는 매우 편리한 방법이지만, 용어의 의미가 희석되어 정확도나 정밀도에 영향을 미칠 수 있게 된다. 그에 반해 Lemmatization은 품사 정보가 보존된 형태의 기본형으로 단어를 변환하게 되므로, 변환시간은 오래 걸리게 되지만 보다 높은 정확도를 얻을 수 있게 된다.

[0050] 그에 따라, 상기 자연언어 처리과정(S250)에서는 Lemmatization을 적용하여 처리함으로써 보다 높은 정확도를 갖고 정밀한 분석이 이루어질 수 있게 하는 것이 바람직하다.

[0051] 상기 데이터 구조화단계(S300)는, 전처리가 완료된 문장의 단어인 주제어들을 인접한 주제어들과 연속적으로 묶어, 상기 전문용어 계층화 사전에 등록되어 있는 전문용어와의 일치 여부를 판단하기 위한 구문의 형태를 생성하기 위한 것으로서, 도 5에 도시된 바와 같이, 1어절 단위나, 연속적인 2어절의 묶음 또는 연속적인 3어절의 묶음으로 구문을 생성하도록 구성된다.

[0052] 현재 CS2013에 대한 전체 내용 분석 결과, 3어절 이하의 전문용어들만이 포함되어 있는 것으로 파악되는바, 도 5에 도시된 바와 같이 3어절 이하의 구문만을 비교의 대상으로 설정하는 것이 바람직하게 된다. 그러나, 추후 4어절 또는 5어절의 전문용어들이 CS2013에 추가될 경우 데이터 구조화 단계에서 그러한 전문용어의 추출이 가능하도록 구문을 형성하는 어절의 개수를 증가시킬 수도 있음은 물론이다. 또한, 도 5에서는 3어절의 구문까지 포함하였지만, 분석 효율을 고려하여 2어절의 구문만으로 한정할 수도 있는 등, 전문용어 추출의 정확도 향상과 분석 효율을 고려하여 전문용어 추출을 위한 어절의 묶음을 증감할 수 있음은 물론이다.

[0053] 상기 구조화 데이터 분석단계(S400)는, 구문의 형태로 구조화가 이루어진 데이터들을 상기 전문용어 계층화 사전에 등록되어 있는 전문용어의 데이터와 비교함으로써, 대상문헌에 포함되어 있는 전문용어를 추출하고, 해당 전문용어가 전문용어 계층화 사전에 등록되어 위치를 확인하여 교육과정 중 어느 지식영역에 해당하는지 여부를 분석하도록 구성된다.

[0054] 이를 위해 도 6에 도시된 바와 같은 알고리즘에 의해 전문용어 계층화 사전에 기반하여, 대상문헌에 포함되어 있는 전문용어의 추출과 교육과정상의 위치 및 수준에 대한 분석이 이루어지게 된다.

[0055] 이러한 상기 구조화 데이터 분석단계(S400)에 의해 상기 데이터 구조화단계(S300)에서 구조화된 데이터들의 집합을 전문용어 계층화 사전에 등록되어 있는 단어들과 비교하게 된다. 즉, CS2013에 기반하여 구축된 전문용어 계층화 사전을 이루는 18개의 지식영역(KA), 136개의 각 지식 세부영역(KDA), 1113개의 단계별(Tier1, Tier2, Elective) 주제, 그리고 각 지식영역의 세부토픽과 비교하면서 분석이 이루어지게 된다.

[0056] 이러한 상기 구조화 데이터 분석단계(S400)에서, 구문 단위로 구조화된 데이터(SD)가 상기 전문용어 계층화 사전에서 검색되면, 해당 데이터가 검색된 위치의 인덱스를 증가시키도록 구성된다. 즉, 전문용어 계층화 사전에서 검색대상인 구문을 이루는 데이터가 발견되면, 18개의 지식영역 중 어느 영역에 해당하는지, 어느 지식 세부영역에 해당하는지, 그리고 어느 주제에 관한 것인지를 체크할 수 있게 된다.

[0057] 구조화된 모든 데이터들에 대하여 전문용어 계층화 사전의 모든 영역에 대한 검색이 완료되면, 검색된 데이터들의 리스트(ResultMT)와, 인덱스가 증가된 위치 인덱스 리스트(TC)를 결과를 획득하게 된다.

[0058] 상기 결과요약단계(S500)는, 상기 구조화 데이터 분석단계(S400)에서 대상문헌으로부터 생성된 구조화된 데이터들 중 상기 전문용어 계층화 사전에서 검색된 데이터들과 그 데이터들이 검색된 위치를 나타내는 정보를 도 7에 도시된 바와 같이, 그래프 또는 텍스트화 시켜 제시하고 저장하도록 구성된다.

[0059] 이때, 도 7의 <전체 요약>에 도시된 그래프에서는 대상문헌에 프로그래밍 언어(PL)와 관련된 전문용어가 가장 많이 검색되고, 그 다음으로는 정보관리(IM), 컴퓨팅 과학(CN)과 관련된 전문용어들이 많이 검색되는 것을 직관적으로 인식할 수 있게 표시됨을 알 수 있게 된다.

[0060] 또한, 검색된 전문용어에 대한 세부내용을 확인할 경우에는 도 7의 <세부 내용>에 나타난 바와 같이, 대상문헌에 'binary search'라는 전문용어가 검색이 되었고, 이 전문용어는 알고리즘과 복잡도(AL)에 관한 지식영역의 Basic Analysis 세부영역 중 Tier1 단계에서 다루어지는 단어임을 확인할 수 있게 된다. 그에 따라, 전체적인 분석결과를 나타내는 그래프뿐만 아니라, 대상문헌에서 실제로 사용된 전문용어와, 지식영역 상의 위치와 수준(지식영역, 지식 세부영역, 단계) 등을 간편하게 확인할 수 있게 된다.

[0061] 이와 같이, 본 발명에 따라 분석하고자 하는 대상문헌에 포함되어 있는 전문용어를 자동을 추출한 후, 그 전문용어가 CS2013에 기반하여 계층적 클러스터링 구조로 구축된 전문용어 계층화 사전에 있는지를 비교 확인하는 것만으로도, 대상문헌이 어떠한 지식영역에 대한 내용을 어떠한 수준으로 주로 포함하고 있는지에 대한 분석이 간편하게 이루어질 수 있게 되므로, 국내 각종 교과과정에서 사용되는 문헌들에 대한 검토뿐만 아니라, 해외 교과과정에 사용되는 문헌들에 대한 분석과 검토도 간편하게 수행할 수 있게 된다.

[0062] 이와 같이 본원발명에 의해 도출될 결과에 의해 컴퓨터과학 교육과정에 대한 글로벌 트렌드를 쉽고 간편하게 파악할 수 있음은 물론, 교과과정에 사용되는 전문용어에만 기초하게 되므로 보다 객관적인 분석이 가능하게 된다. 이처럼 본원발명에 의해 도출된 각국의 컴퓨터과학 교육과정에 대한 분석결과를 토대로 세계적 흐름을 고려한 교육과정을 구축함에 있어 많은 도움을 줄 수 있을 것이다.

[0063] 이상에서는 본 발명에 대한 기술사상을 첨부 도면과 함께 서술하였지만 이는 본 발명의 바람직한 실시예를 예시적으로 설명한 것이지 본 발명을 한정하는 것은 아니다. 또한 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 이라면 누구나 본 발명의 기술적 사상의 범주를 이탈하지 않는 범위 내에서 다양한 변형 및 모방이 가능함은 명백한 사실이다.

부호의 설명

[0064] S100 : 전문용어 계층화 사전 구축단계

S200 : 대상문헌 전처리단계

S210 : 토큰화 과정 S220 : 특수문자 제거과정

S230 : 1음절 제거과정 S240 : 소문자 정규화 과정

S250 : 자연언어 처리과정

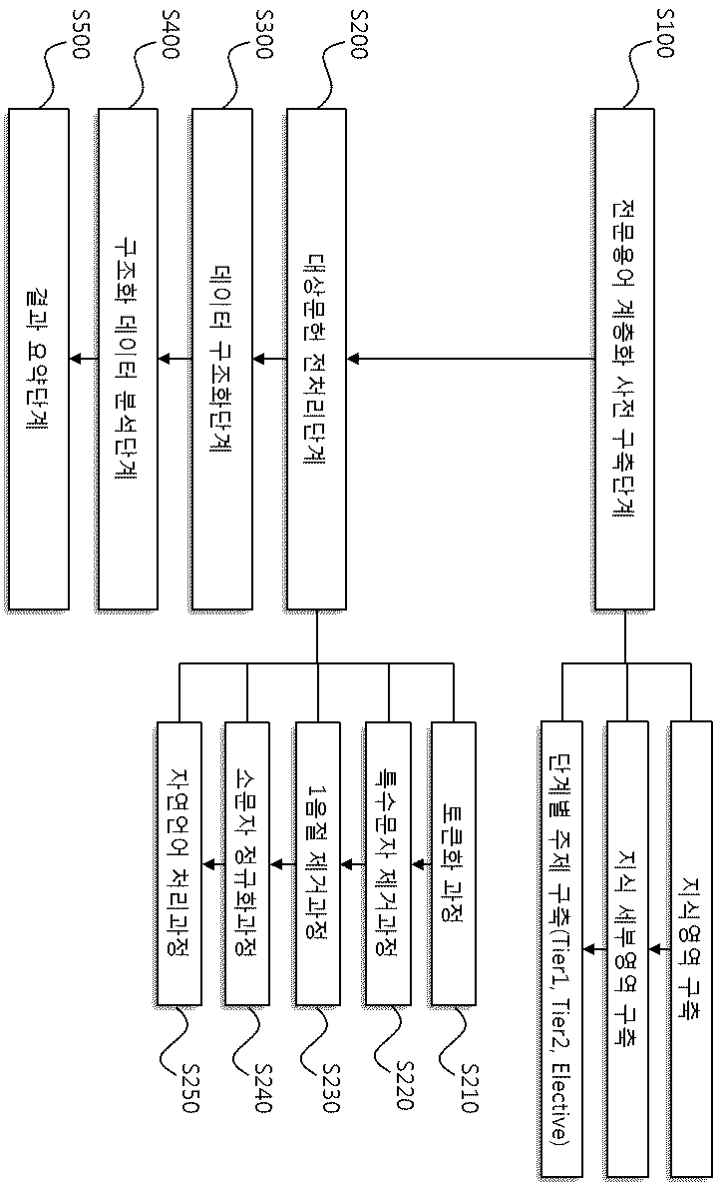
S300 : 데이터 구조화단계

S400 : 구조화 데이터 분석단계

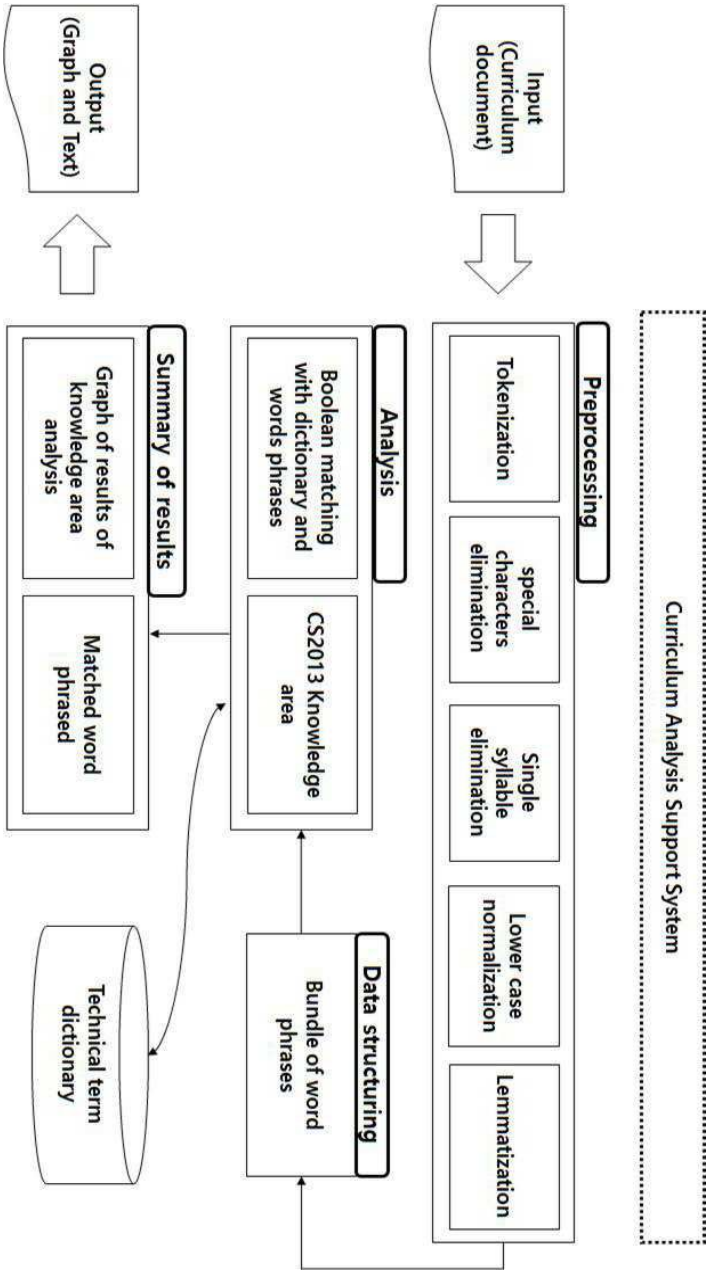
S500 : 결과요약단계

도면

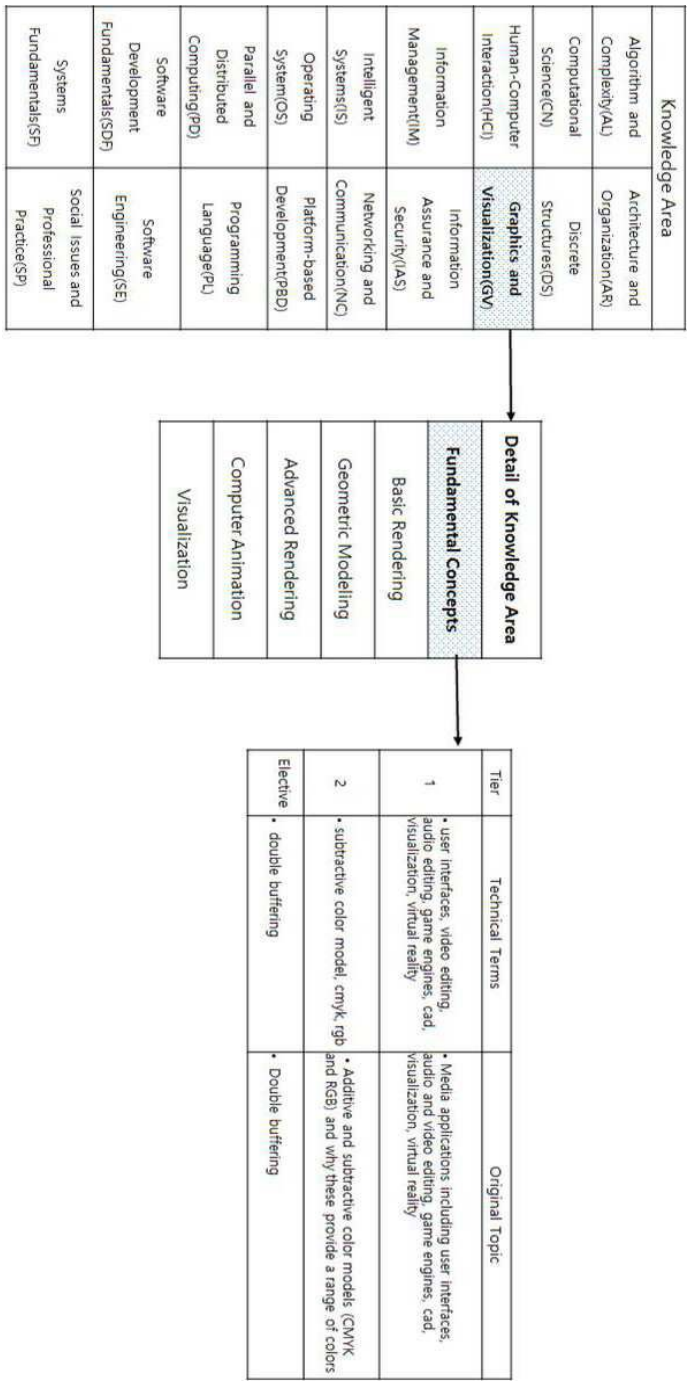
도면1



도면2


Input (Curriculum document)
Output (Graph and Text)
Curriculum Analysis Support System
Preprocessing
Tokenization
special characters elimination
Single syllable elimination
Lower case normalization
Lemmatization
Analysis
Boolean matching with dictionary and words phrases
CS2013 Knowledge area
Data structuring
Bundle of word phrases
Summary of results
Graph of results of knowledge area analysis
Matched word phrased
Technical term dictionary

도면3


Knowledge Area
Algorithm and Complexity(AL)
Architecture and Organization(AR)
Computational Science(CN)
Discrete Structures(DS)
Human-Computer Interaction(HCI)
Graphics and Visualization(GV)
Information Management(IM)
Information Assurance and Security(IAS)
Intelligent Systems(IS)
Networking and Communication(NC)
Operating System(OS)
Platform-based Development(PBD)
Parallel and Distributed Computing(PD)
Programming Language(PL)
Software Development Fundamentals(SDF)
Software Engineering(SE)
Systems Fundamentals(SF)
Social Issues and Professional Practice(SP)
Detail of Knowledge Area
Fundamental Concepts
Basic Rendering
Geometric Modeling
Advanced Rendering
Computer Animation
Visualization
Tier
Technical Terms
Original Topic
1
• user interfaces, video editing, audio editing, game engines, cad, visualization, virtual reality
• Media applications including user interfaces, audio and video editing, game engines, cad, visualization, virtual reality
2
• subtractive color model, cmyk, rgb
• Additive and subtractive color models (CMYK and RGB) and why these provide a range of colors
Elective
• double buffering
• Double buffering

도면4

• 주제어	"An example of a simple recurrence relation, such as Fibonacci numbers. "
• 토큰화	An / example / of / a / simple / recurrence / relation, / such / as / Fibonacci / numbers /.
• 특수문제 제거	An / example / of / a / simple / recurrence / relation / such / as / Fibonacci / numbers
• 하나의 음절 제거	An / example / of / simple / recurrence / relation / such / as / Fibonacci / numbers
• 소문자 정규화	an / example / of / simple / recurrence / relation / such / as / fibonacci / numbers
• Lemmatization	an / example / of / simple / recurrence / relation / such / as / fibonacci / number

도면5

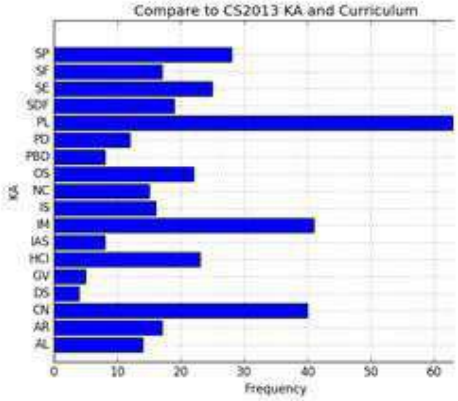
• 전처리 된 주제어
an / example / of / simple / recurrence / relation / such / as / fibonacci / number
• 1어절
an / example / of / simple / recurrence / relation / such / as / fibonacci / number
• 2어절
an example / example of /of simple / simple recurrence / recurrence relation / relation such / such as / as fibonacci / fibonacci number
• 3어절
an example of / example of simple / of simple recurrence / recurrence relation such / as fibonacci number

도면6

```
Algorithm : Analyze algorithm
    Input : SD /*the structured data list*/
    Output : ResultMT/*the matched Topic list */,
             TC/*the matched Topic count list in Detail of Knowledge area */
    Definition: CSDictionary /* the dictionary of CS2013 */,
                KA /* Knowledge area*/,
                DKA /* Detail of Knowledge area  */,
                TDKA /* topic in Detail of Knowledge area  */
1   begin
2       Initialize ResultMT = [], TC = []
3       for each i in len(KA)
4           for each j in len(DKA)
5               for each k in 3 // looping for Tier1, Tier2, Elective
6                   flag = False
7                   for each l in len(TDKA)
8                       for each m in len(SD)
9                           if(BooleanRetrievals(CSDictionary[i][j][k][l], SD[m]) == True)
10                              Save(ResultMT, SD[m]) // save sd[m] in ResultTT list
11                              flag = True
12                          end
13                      end
14                  end
15                  if flag == True
16                      TC[i][j][k] = TC[i][j][k] + 1
17                  end
18              end
19          end
20      end
21      return ResultMT, TC
22  end
```

도면7

<전체 요약>



<세부 내용>

구분	내용
지식 영역	AL
지식세부 영역	Basic Aanalysis
단계	Tier1
매칭 된 전문용어	binary search
CS2013의 원문 토픽	'Sequential and binary search algorithms.'
비율	19.05 %