KIPS Trans. Softw. and Data Eng.
Vol.7, No.12 pp.469~476 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2018.7.12.469

A Study on Similarity Calculation Method Between Research Infrastructure

Kim Yong Joo† · Kim Young Chan††

ABSTRACT

In order to jointly utilize research infrastructure and to build efficient construction, which are essential in science and technology research and development process. Although various classification methods have been introduced for efficient utilization of registered information, functions that can be directly utilized such as similar research infrastructure search is not yet been implemented due to limitations of collection information. In this study, we analyzed the similar search technique so far, presented the methodology for the calculation of similarity of research infrastructure, and analyzed the learning result. Study suggested that a technique can be use to extract meaningful keywords from information and analyze the similarity between the research infrastructure.

Keywords : Research Infrastructure, Information-Similarity Algorithm, Word Vector, LSA

국가연구시설장비의 유사도 판단기법에 관한 연구

김 용 주† · 김 영 찬††

요 약

연구개발과정에서의 필수요소인 연구장비의 공동활용 및 효율적인 구축을 위해 한국에서는 국가예산으로 구축된 장비정보를 필수적으로 등록하도록 하고 있다. 등록정보의 다양한 활용(중복성 검토, 성능예측, 대체장비추천)을 위해 본 연구에서는 현재 유사장비검색기법에 대해 분석하고 유사도 산출 방법을 제시하였다. 이를 통해 자연어 상태인 장비정보에서 키워드를 추출하여 LSA 기법을 적용하면 키워드간의 유사도산출 및 장비정보 간 유사도 분석이 가능함을 확인하였으며 향후 연구장비분류정보를 접목하여 적용할 경우 의미있는 유사도 산출 및 이를 활용한 다양한 서비스가 가능 할 것으로 예측된다.

키워드 : 연구장비, 정보유사도 알고리즘, 단어백터, LSA

1. 서 론

1.1 연구목적

한국의 국가연구개발 사업에서 연구시설·장비 구축에 대한 예산투자 비중이 꾸준히 3% 이상[1]을 차지하고 있으며, 이의 효율적인 도입 및 활용이 국가적인 이슈가 됨에 따라 투자 효율성 제고에 대한 관심이 높아지고 있다. 이에 관련 법령에서는 국가연구개발사업으로 구축한 연구시설장비정보를 30일이내에 ZEUS(Zone of Equipment Utilization Service, http://www.zeus.go.kr) 서비스에 등록하도록 규정하고 있으며 다양한 분야에 활용되고 있다. 특히 다른 모델이지만 유사한 성능을 발휘하는 장비정보의 검색은 연구장비의 중복성검토, 대체장비의 추천, 성능 예측등은 그 활용도가 매우 높음에도 불구하고 등록자가 입력한 분류정보의 정확도 부족 및 다양한 모델정보로 인해 유사도 구현에 한계가 있었다. 이에 이전에도 문헌정보·경제학 관점의 유사도 측정방식에 대한 연구가 있었으나 전문가의 정성적 판단을 보조하는 수단으로 밖에 활용되지 못하고, 정보의 의미에 기반 한 유사도 측정이 불가능 하였다.

본 연구에서는 공공목적의 데이터베이스인 연구시설·장비 정보간의 컴퓨터공학 관점의 유사도측정방법을 제시하고 그 성능에 대해 정량적으로 분석하고자 한다.

1.2 연구방법 및 범위

본 연구에서는 연구시설·장비 정보의 특성을 분석하여, 연구시설·장비에 특화된 유사도 측정 알고리즘을 제안하고자 한다.

※ 이 논문은 국가연구시설장비 선진화 지원사업(2018)에 의해 연구되었음.
† 정 회 원 : 국가연구시설장비진흥센터 연구원
†† 비 회 원 : 한밭대학교 컴퓨터공학과 교수
Manuscript Received : May 14, 2018
Accepted : August 9, 2018
* Corresponding Author : Kim Young Chan(yckim@hanbat.ac.kr)

본 연구의 범위는 국가연구개발사업으로 구축한 연구시설·장비정보(98,000여건)를 대상으로 하며, 국가연구시설장비표준분류에 의해 분류된 정보를 유사도 산출방법의 결과와 비교하여 알고리즘의 성능을 측정하고자 한다.

2. 관련 연구

연구시설·장비의 성능정보 표준화 및 DB화를 통해 유사장비여부를 판정하는 방법을 정연대(2012)[2]는 제시하였다. 이는 가장 기본적인 유사 연구시설·장비 산출 방법이나 성능정보 추출에 많은 비용이 소요되는 문제점이 있다. NFEC(2013)[3]은 연구시설·장비를 LSI(Latent Semantic Indexing) 기반 색인 작업 후 장비 중복유사도를 측정하는 알고리즘을 제시하였다. 이는 동일모델에 속한 장비를 유사장비로 그룹화 한 후 설치장소(광역단위)와 좌표정보를 가중치로 적용한 방식으로 다른 모델이지만 동일한 성능을 발휘하는 장비간의 직접적인 유사도를 산출할 수 없는 문제가 있다. Jeong(2013)[4]은 국가연구개발과제의 중복도 산출을 위해 키워드 중심의 백터공간 측정기법을 제시하였다. 이는 과제정보가 보유한 텍스트정보에서 키워드를 추출하여 키워드의 빈도를 측정하는 기법으로 장비정보와 같이 보유 키워드가 한정적일 경우 유사도측정이 불가능한 문제가 있다. Sun(2014)[5], Hoang (2016)[6], Mahmood[7] 등은 사용자의 정보관심도(클릭스트림, 정보추천)에 의해 유사도를 측정하는 방식을 제안하였으나, 사용자의 관심도에 관한 정보가 없는 경우 정보의 유사도를 측정할 수 없는 한계가 있다. Kim(2015)[8], Neethukrishnan (2017)[9], Mahmood (2013)[10] 등은 보유정보의 RDF, 온톨로지등을 도출하고 연관된 엔티티를 기반으로 유사도 높은 정보를 수학적으로 표현하는 방법에 대해 제시하였으나 이는 장비간의 유사도를 정량적으로 산출할 수 없는 문제가 있다.

기존 연구의 대부분은 정보관리자의 정제작업(자료입력원의 수작업)에 의해 수작업으로 연결된 정보(메타분류, 키워드분석, 온톨로지, 사용자 추천 등)를 사용하여 유사도를 측정하는 방식으로 이는 연관정보가 없는 경우 유사도 산출이 불가능한 문제가 있다. 공공영역의 데이터는 그 수집방식의 한계(정보가 의무적으로 입력하여 품질 저하)로 인해 정제작업이 필수적으로 국가연구시설장비 정보에 기존 연구결과 적용 시 많은 비용이 발생한다.

3. 유사도 측정방법 연구

3.1 연구시설·장비 정보의 특성 분석

국가연구개발사업으로 구축한 3천만원이상 연구시설·장비를 30일이내에 등록하도록 법령으로 규정하고 있다[11]. 이렇게 수집된 정보는 연구시설·장비의 취득, 활용, 관리, 공동활용, 연계등에 관련된 68개 정보항목으로 관리되고 있으며, 그 중 텍스트 정보항목 4개(연구시설·장비명, 설명, 구성 및 성능, 사용예)와 코드(CODE) 정보항목 6개(시설·장비표준분류, 제작사, 제작국가, 모델, 장비용도, 활용범위)는 간접적으로 유사도 측정에 활용할 수 있다.

기 수집된 98,000여점의 정보 분석결과 60,000여종의 모델로 분류되어 있으며(1종당 1.8점), 국가연구시설장비표준분류별 구축편차가 있어 이를 활용한 유사도 산출이 제한적인 상황이다.

Table 1은 국가연구시설장비 표준분류별 정보의 분포를 보여주고 있다.

또한 국가연구시설장비 정보는 다음과 같은 분류적 특성을 가지고 있다.

- 국가연구시설장비표준분류에 의해 정보를 분류하고 있다. 하지만 기능과 역할에 따라 같은 장비일지라도 다른 국가연구시설장비표준분류로 구분된다.
- 연구시설·장비는 제작사와 모델정보를 가지고 있으며 이를 통해 구분된다.

연구시설·장비정보는 명확하게 정의된 정보속성에 따라 값이 입력되어 있는 구조적 데이터로 볼 수 있으나, 다양한 장비분류 및 모델로 인해 정보군집도가 약해 이미지, 소리, 텍스트와 같은 반구조적 데이터[12]라고 볼 수 있다.

수집한 연구시설·장비 정보는 통상적인 명칭을 사용하지 않고, 연구주제를 장비명으로 활용(서버 → 분자해석을 위한 연산노드, 액체크로마토크래피→물성분 해석을 위한 분석시스템)하는 경우가 많아, 장비명을 통한 유사도 측정이 불가능하다. 하지만 장비가 보유한 다른 텍스트 정보(장비설명, 구성 및 성능, 사용예 등) 종합하여 분석할 경우, 장비의 특성을 유추해 볼 수 있다. 다시 말해 유사도를 가진 장비는 유사한 단어를 특정한 목적으로 사용한다는 알 수 있으며, 이는 단어간의 유사도측정 방법 적용 시 장비의 특성을 학습하여 유사장비를 구별할 수 있을 것으로 생각할 수 있다. 따라서 본 연구에서는 키워드분석을 통한 장비 정보 간의 유사도를 측정하는 방법을 제시한다.

3.2 유사 연구시설·장비 개념정의 및 한계

유사 연구시설·장비는 동일한 기능 및 연구목적을 가진 연구시설·장비로 정의할 수 있다.

Table 1. Construction Ratio of Standard Classification

Classification	Optical imaging	Compound pretreatment analysi	Machining test	Electrical and electronic	Data Processing	Measuring	Medical	Facility
Ratio	12.8%	23.1%	27.4%	13.8%	10.6%	9.1%	1.8%	1.6%

- 동일 연구시설·장비는 동일한 모델 및 연구목적을 가진 장비를 말한다. 반면 유사 연구·시설장비는 동일한 모델은 아니나 동일한 연구목적으로 활용되며 유사한 성능을 가진다.

현재 연구시설·장비의 유사도를 판단하려면, 유사 모델정보 또는 개별 장비의 성능정보(예, CPU성능, 분해능, 분석강도 등)를 가지고 있어야 한다. 수집된 정보는 유사 모델간 연관관계 및 구조화된 성능정보를 수집하지 않으며, 만약 현재의 정보를 가지고 동일·유사 연구시설·장비정보를 제공하려면 전문가에 의한 정보정제(Data cleansing)작업이 필수적이다.

3.3 연구시설·장비 유사도 산출 방법론

연구시설·장비 정보가 가진 텍스트 정보항목을 하나의 문서로 합치고 개별 단어-문서간의 빈도를 구한 후 다차원의 빈도배열을 2차원으로 축소한 후 유사도를 산출한다.

이에 대한 절차적 방법론은 다음과 같다.

1) 연구시설·장비 정보 중 텍스트 정보항목을 모아 하나의 문서로 만든 후, 형태소 분석 및 유사어 확장을 통해 유사도 산출을 위한 단어(Term)집합을 만든다.
   - 본 연구에서는 알고리즘의 성능측정을 위해 한글장비명, 영문장비명, 모델명, 제조사명등 장비의 특징을 구별할 수 있는 텍스트 정보항목을 제외하고 장비설명만을 활용하였다.
   - 연구시설·장비정보가 보유한 텍스트는 평균 400자(한국어기준) 정도이며 연구분야에 특화된 비표준 단어가 많이 사용되기 때문에 학습에 적합하지 않다. 이를 해소하기위해 동의어, 유사어를 해당 단어 다음에 삽입한다.
2) 키워드를 백터공간에 배치 한 후 문서(장비)-문서(장비)간의 동일키워드 빈출 빈도 및 동일 단어간의 거리를 지속적으로 학습하여 유사도를 산출한다.
   - 이때 단어간의 유사도 산출을 위해선 잠재의미분석기법(LSA)을 사용한다. 개별단어-개별문서의 단어빈도-역문서빈도행렬(Term Frequency - Inverse Document Frequency, TF-IDF[13])을 생성한다. 이때 매우 큰 희소행렬이 생성됨으로 특이값 분해(Singular Value Decomposition, SVD)를 적용하여 차원을 축소한다.
   - 분해된 SVD 행렬을 사용하여 단어간, 문서(장비)간 코사인 유사도를 계산한다.

Fig. 1은 위의 절차를 설명하고 있다. 여기서 Wi는 SVD로 분해된 Di의 단어행렬이며 Si는 특이값(sugular value), DTi는 문서행렬의 전치행렬이다. 이를 활용하여 단어간의 코사인 유사도 및 문서간의 코사인 유사도를 구한다.

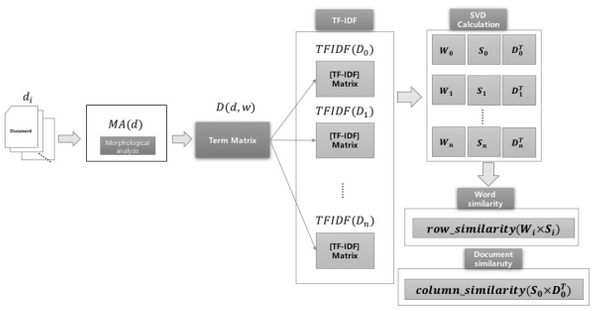

Fig. 1. Analysis Procedure for Similarity

3.4 유사도 산출 알고리즘의 구현

장비의 텍스트정보만을 모은 하나의 문서를 생성한 후 '한국어 형태소분석기(UTagger)[14]로 명사구-술어명사, 형용사-관형형어미-명사구를 추출한다. 이는 문서 내 의미 있는 단어만을 추출하여 연산시간을 줄이기 위함이다. Table 2와 3은 과정을 예시로 설명하고 있다.

Table 2. Object noun phrase + predicate noun extraction[on]

input text	result
The Future Department plans to utilize the technical commercialization expert group that manages the whole project to create visible commercialization performance within two years.	[On] Project Overall Management [On] Using a group of technical commercialization experts [On] Promoting the Performance of Commercialization

Table 3. [Adjective + tubular mother + noun phrase] Syntax extraction[jn]

input text	result
Production of customized biomaterials at low cost	[jn] low cost
to compete in artificial machine development using intense 3D printing technology.	[jn] 3D printing technology

Table 4는 추출된 형태소를 다시 문서로 만드는 과정을 예시로 설명하고 있다.

Table 4. Process of Extracting Document Using the Extracted Keywords

Equipment name	Text set	Morphological analysis	Object noun phrase + predicate noun pair extraction
liquid chromatographic graph	Mixed compounds in solvents...	Mixed state | Compounds| Nonvolatile.	[on] Mixture separation | n1 blend state | n1 compound .
liquid chromatographic graph	Analyzing and refining various materials...	Purification means| chromatographs...	n1 Diversity | n1 Materials | n2 Analysis and Purification Water | n2 Liquid .
high performance liquid chromatography	Features This system is a sample...	Features| System| sampleinjection.	n1 Various Materials | n2 Analysis and Refining | n1 Features | n1 system | n1 sample | n1 injection | n2 Liquid

Table 5. Expansion of the Document by searching for and inserting the synonyms of the keywords

Equipment name	Document	Expand synonyms and synonyms
high speed liquid chromatography	n1 Hangul name | n4 high-speed liquid chromatography	High Performance Liquid Chromatography | High-speed Protein Liquid Solution _ Chromatography | High-speed Liquid Chromatography | Nano Liquid Chromatography | |Manufacturing Liquid Chromatography | Large-capacity Accumulation Chromatography | Heavy Pressure Chroma

Table 5는 문서내 키워드의 동의·유사어를 삽입하여 문서량을 확대하는 예시를 설명한다.

확장된 문서를 가지고 개별 문서의 TF-IDF 계산하는 과정을 Fig. 2에서 설명한다.

- 문서내에 특정한 키워드가 나타나는 빈도(TF, Term Frequency)를 산출한다.
- 키워드자체가 문서전체에 나타나는 빈도(DF, Document Frequency)를 산출한 후 역빈도(IDF, Inverse Document Frequency)를 계산하고 키워드빈도와 역빈도를 곱하여 TF-IDF를 산출한다.
- 이는 자주 빈출되는 단어 중 그 문서의 성격을 나타내는 단어를 부각시키기 위함이다. TF-IDF를 구하는 수식은 [정의 1]과 같다.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \quad \text{[정의 1]}$$

	NO1	NO2	NO3
equivalent	0	0	0
resolution	0	0	0
pretreatment	0	0	2.83
spectroscopy	0	4.28	0

Fig. 2. Examples of TF-IDF Calculatuon

계산된 TF-IDF에 특이값분해(SVD)를 적용하여 차원을 축소한 후 장비-키워드간 벡터공간내 내적(Inner Product)을 계산하여 유사도를 산출한다. Fig. 3과 같이 SVD를 적용하여 차원을 축소하며 요소별로 분리한다. Wi는 SVD로 분해된 Di의 단어행렬이며 Si는 특이값, DTi는 문서행렬의 전치행렬이다.

그 다음 SVD를 통해 분해된 행렬을 이용하여 장비 및 단어 간 유사도를 계산한다. 코사인 유사도는 두 비교대상간 내적을 총 백터크기로 나누어서 구하며 [정의 2]와 같다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad \text{[정의 2]}$$

Fig. 4는 문서간의 유사도를 구하기 위해 S와 DT를 산술하여 유사도를 구하는 과정을 보여주고 있으며 Fig. 5는 단어간의 유사도를 구하는 과정을 보여주고 있다.

double[][] tf_idf

	NO1	NO2	NO3
equivalent	0	0	0
resolution	0	0	0
pretreatment	0	0	2.83
spectroscopy	0	4.28	0

new SingularValueDecomposition() ↓ getW() getS() getDt()

double[][] W

	Dimension 1	Dimension 2	Dimension 3
equivalent	0	-0.07	-0.016
resolution	0	0	0
pretreatment	0	0	0
spectroscopy	0	0	2

double[][] S

	Dimension 1	Dimension 2	Dimension 3
Dimension 1	18.8	0	0
Dimension 2	0	13.8	0
Dimension 3	0	0	11.2

double[][] Dt

	NO1	NO2	NO3
Dimension 1	0.005	-0.004	0.010
Dimension 2	0.002	-2.3E-4	0.002
Dimension 3	0.999	0.009	-0.013

Fig. 3. Examples of SVD Calculation

(S X DT)T

	NO1	NO2	NO3
Dimension 1	12.00	-3.17	1.40
Dimension 2	5.20	-3.20	0.03
Dimension 3	35.04	-47.45	-4.40

Calculator.getCosSim() ↓

double[][] docCosSim

	NO1	NO2	NO3
NO1	12.00	-3.17	1.40
NO2	5.20	-3.20	0.03
NO3	35.04	-47.45	-4.40

Fig. 4. Examples of Document Similarity Calculation

$(W \times S)^T$

	Dimension 1	Dimension 2	Dimension 3
equivalent	0.71	0.26	0.29
resolution	1.97	1.10	1.78
pretreatment	1.51	0.40	1.31
spectroscopy	0.10	0.01	0.03

Calculator.getCosSim() ↓

double[][] wordCosSim

	equivalent	resolution	pretreatment	spectroscopy
equivalent	1.0	4.76E-15	-5.11E-16	1.62E-15
resolution	0.0	1.0	0.17	0.45
pretreatment	0.0	0.0	1.0	0.38
spectroscopy	0.0	0.0	0.0	1.0

Fig. 5. Examples of Word Similaruty Calculation

4. 결과 분석

본 연구에서 제안한 유사도 산출방법의 성능이 우수하다면, 기준장비(유사도 측정대상)와 동일분류, 유사모델에 해당하는 장비가 유사도 순위에서 상위에 도출되어야 한다.

전체 유사도 순위를 5개 구간으로 분할 한 후, 구간별 동일분류 및 유사모델의 개수를 2차원 공간에 배치하여 선형추세선 및 기울기를 구하면 Fig. 6과 같은 유형이 도출된다.

Fig. 6에서 X축은 전체장비의 유사도 역순위를 5개구간으로 분리한 것이며, Y축은 기준장비와 동일모델의 개수를 의미한다. 이때 X축의 후순위에 유사한 장비가 많이 배치될수록 추세선의 기울기가 크며, 이는 추세선의 기울기가 성능측정방법으로 사용될 수 있음을 의미한다.

본 연구에서는 성능측정을 위해, 총 5개의 국가연구시설장비 표준분류내 소분류(액체크로마토그래프, 입도분석기, 오실로스코프, 서버, 광학현미경)에서 각 200개의 임의 장비정보를 추출하였으며, 기준장비는 각 표준분류의 임의의 장비를 선택하였다. Table 6은 표준분류별 추출장비정보를 설명하고 있으며 Table 7은 각 분류별 기준장비 및 기준장비와 정성적으로 유사한 모델의 정보현황을 보여주고 있다.

Table 6. Test Dataset for Algorithm Performance

Category	Number
Liquid chromatography mass spectrometer	200
Particle size analyzer	200
oscilloscope	200
server	200
optical microscope	200
Total	1,000

Table 7. Reference Equipment for Comparison

NO	Reference Equipment	model name	Category	Number of identical or similar models
NO1	Mass spectrometer	4000 QTrap	Liquid chromatography mass spectrometer	24
NO2	Particle Measuring Equipment	LS 13 320	Particle size analyzer	22
NO3	IP San Storage	Equallogic PS400e	Server	75
NO4	Fluorescence microscope	Eclipse 80i	Optical microscope	30
NO5	oscilloscope	DPO70804	Oscilloscope	45

시험데이터에 유사도 성능측정방식을 적용한 결과는 다음과 같다.

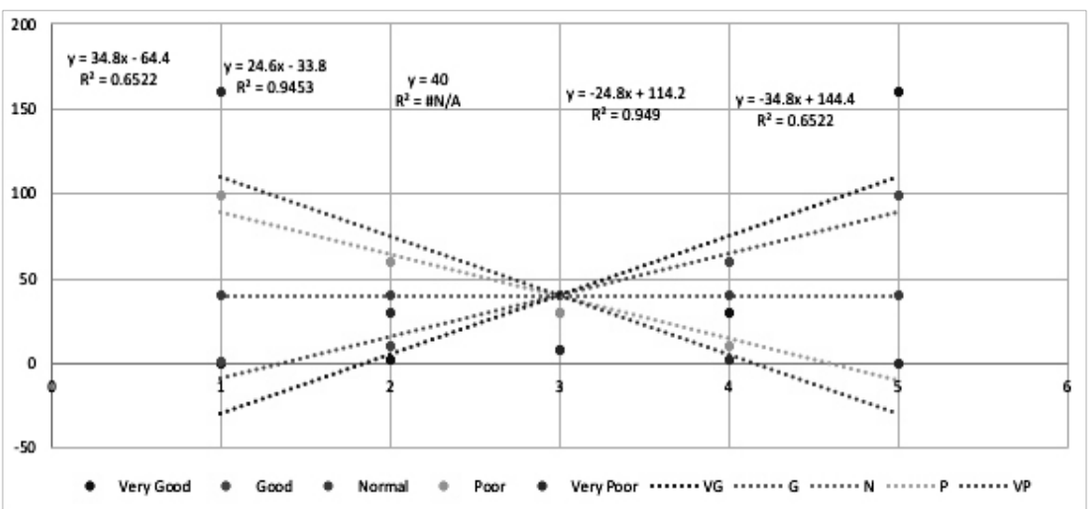

Fig. 6. Examples of Graph

Table 8. Same Classification Distribution in Similar Equipment

Section \ Equipment	NO1	NO2	NO3	NO4	NO5
1~200	112	97	114	114	107
201~400	44	46	26	37	35
401~600	25	22	21	17	30
601~800	10	19	23	18	16
801~1000	9	16	16	14	12
Total	200	200	200	200	200

먼저 동일 소분류 분포를 분석하였다. Table 8에서 보는 바와 같이 동일 소분류에 속한 장비가 상위구간(유사도가 높은 구간)에 많이 분포하는 것을 볼 수 있으며 Fig. 7에서 보는 바와 같이추세선의 평균기울기는 21.12로(최고 40, 최저 -40) 본 연구에서 제안한 알고리즘으로 유사한 장비를 추출할 수 있음을 보여준다.

Table 9. Same or Similar Model Distribution in Similar Equipment

Section \ Equipment	NO1	NO2	NO3	NO4	NO5
1~200	8	13	38	19	19
201~400	7	5	15	5	5
401~600	3	2	10	3	3
601~800	4	1	4	1	9
801~1000	2	1	8	2	9
Total	24	22	75	30	45

또한 유사모델의 분포를 분석하였을 때 Table 9에서 보는 바와 같이 유사도 상위 구간에서 동일모델이 많이 분포되는 것을 볼 수 있으며 Fig. 8과 같이 그래프의 기울기 또한 양(+)의 값을 갖는 것을 볼 수 있다.

아울러 하위구간으로 검출되는 유사분류 정보에 대해 검토해 볼 필요가 있다. NO3장비(Server)에서 하위구간(801-1000)에 포함된 동일소분류 장비 10점의 정보를 분석한 결과, 아래의 Table 10에서 보는 바와 같이 3가지 유형의 정보가 검출되었다.

Table 10. Ideal value type

Case	Name	Manufacture	Category	Description
1	Disk	Hitachi	Server	null
2	storage	Hitachi	Server	-2146826259
3	Digital Telemetry System	Telemetrie Elektronik	Server	wireless data transfer system……

1번 유형은 텍스트 정보가 없는 경우, 2번 유형은 잘못된 값이 입력된 경우, 3번 유형은 소분류를 잘못 입력한 정보이다. 이러한 유형들은 공공목적의 데이터베이스에서 많이 발생되는 현상으로, 많은 인력(비용)을 투입하여 데이터를 수정하고 있다. 이는 기 구축된 정보의 오류를 검출할 수 있는 방법으로 본 연구결과가 사용될 수 있음을 알려준다.

5. 고 찰

본 연구를 수행한 국가연구시설장비정보 데이터베이스의 사용자는 고품질의 데이터를 활용한 서비스(유사도 검색)를 요구하고 있다. 하지만 정보수집체계의 한계로 정보의 품질이 좋지 않아 사용자의 요구를 만족시킬 수 없었다. 이를 위해 정보간의 유사도를 정량적으로 산출하여 유사정보간의 추천을 통해 정보를 보완해야 하지만 표준화된 정보항목의 부족으로 유사관계를 추출할 수 없었다.

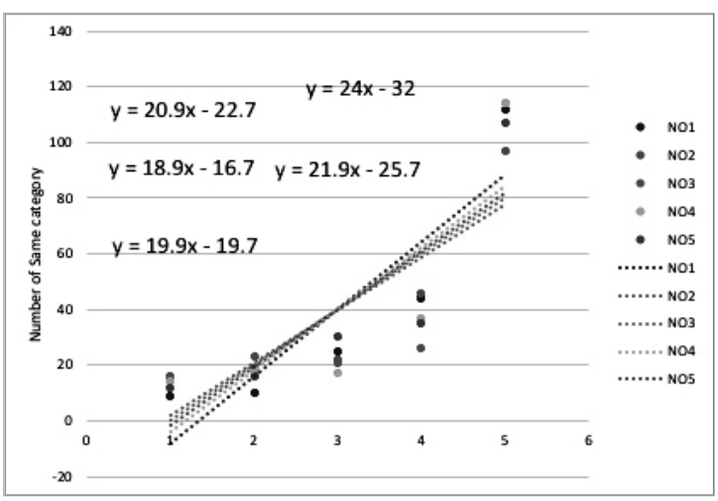

Fig. 7. Analysis

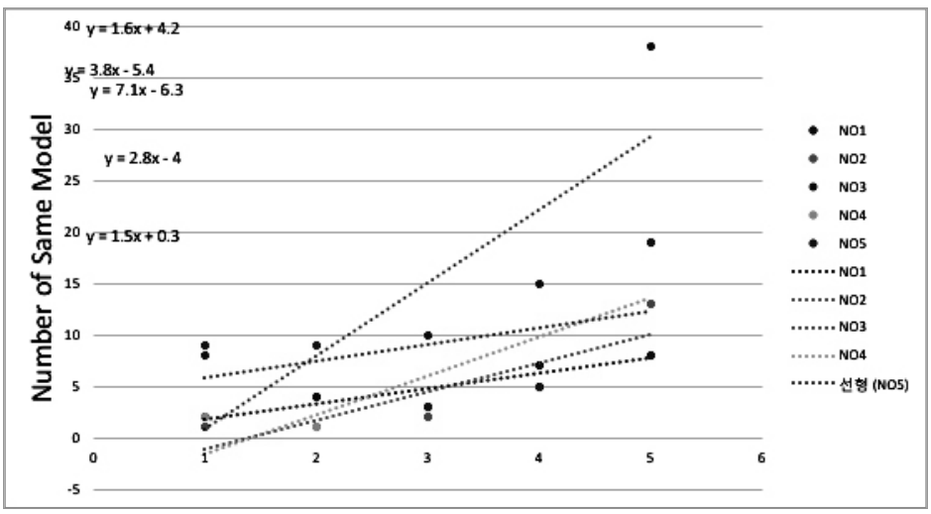

Fig. 8. Analysis

본 연구에서는 부족한 데이터를 활용하여 정보간의 유사도를 산출하기 위해 정보가 가진 텍스트정보를 활용하여 유사도를 측정하는 방식에 대해 연구하였다. 이를 위해 데이터베이스에서 텍스트정보를 추출하는 방법, 머신러닝을 활용하여 단어간의 유사도를 측정하는 방법, 또한 이를 활용하여 정보의 유사도를 측정하는 방법을 제시하였다.

본 연구는 한국어를 기반으로 수행하였기에 언어가 다를 경우 같은 의미를 가지더라도 유사도를 측정할 수 없는 한계가 있었다. 이는 어휘망과 시소러스 등을 결합하면 충분히 해결될것이라 생각한다. 향후 이러한 텍스트유사도와 사용자경험, 분류정보 등을 결합하여 좀 더 정확한 유사도 측정방법을 산출할 수 있을 것이라 생각한다.

6. 기대효과 및 향후 연구방향

6.1 연구의 기대효과

본 연구를 통해 특정 데이터베이스(본 연구에서는 연구시설·장비)가 보유한 텍스트 정보로 정보간의 유사도를 측정할 수 있는 방법을 도출하였다. 특히 본 연구를 통해 다음과 같은 분야에 활용가능하다.

- 연구시설·장비 관련 정보서비스 제공중이던 키워드 빈출빈도 기반 검색방식을 유사도 기반 검색으로 대체하여 서비스 이용자에게 좀 더 정확한 정보를 제공할 수 있다.
- 연구시설·장비 정보등록시 표준분류의 추천이 가능하며, 오분류된 정보 검출이 가능하다.

아울러 본 연구를 응용하면 다음과 같은 정보 서비스 개발이 가능하다.

- 현재 비교대상 장비(모델)를 선택해야만 활용가능한 중복성 검토기능을 적정가격 산정기능을 키워드를 통해서도 서비스 할 수 있는 기반을 마련하였다.
- 연구장비의 도입연도별 주요 키워드를 도출하여 어떠한 연구 또는 장비가 유행하는지 트렌드 분석이 가능해진다.
- 연구자가 보유한 장비의 텍스트를 하나의 분석대상 문서로 만든 후 이에 대한 유사도를 분석하면 유사한 연구를 수행하는 연구자를 연도별로 도출 할 수 있다.
- 국가연구개발과제로 구축한 연구장비의 텍스트를 하나의 분석대상 문서로 만들고 유사도를 측정하면 유사한 과제를 유추 할 수 있다.
- 장비별 도출된 키워드를 검색(해시테그) 또는 정보 군집화 방법으로 활용할 수 있다.

6.2 향후 연구방향

본 연구에서 구현한 알고리즘은 인공 신경망으로 향후 연구시설·장비 정보관련 딥러닝 구현시 전처리 단계로 활용할 수 있다. 또한 학습기반 알고리즘으로 연구시설·장비의 성능예측 및 정보정확도 검증에 활용할 수 있다.

현재, 사용자가 직접 등록한 연구시설·정보의 정보품질을 개선하기 위해 많은 비용을 투입하고 있다. 이를 해결하기 위해 학습기반 알고리즘을 적용하여 장비명 또는 모델명에서 장비의 주요정보를 유추하여 사용자 정보입력을 최소화 할 수 있는 방법을 연구할 수 있다.

References

[1] Ministry of Science and ICT, R.O.Korea, “National Research Facilities & Equipment Trends 2016”.

[2] NFEC, “A Study on the Similarity Estimation of Research Facilities and Equipments,” PRISM. Polissue 14 2015.

[3] NFEC Research report(Kyung Hee University), “Research Equipment duplication and the improvement of the fair price calculation,” 2014.

[4] Jeong Ok-Nam, “A Study on the Improvement of Similarity Evaluation Model for R&D Project,” Ph.D. dissertation Soongsil University, 2013.

[5] Jianshan Sun, “Leveraging Content and Connections for

Scientific Article Recommendation in Social Computing Contexts," *The Computer Journal*, Vol.57, Issue 9, Sept. 2014.
[6] DinhTuyen Hoang, "Academic event recommendation based on research similarity and exploring interaction between authors," *2016 IEEE International Conference on Systems, Man, and Cybernetics*, 2016.
[7] Qamar Mahmood, Muhammad Abdul Qadir, Muhammad TanvirAfzal, "Application of COReS to Compute Research Papers Similarity," IEEE Access, 2017, Vol.5.
[8] Sukyung Kim, "Advanced Ontology Using the Seesorus and Meta Information of Research Facilities, Korea Basic Science Research Institute," Hanbat National University, 2015.
[9] K.V. Neethukrishnan, "Ontology based research paper recommendation using personal ontology similarity method," 2017 Second International Conference on Electrical,Computer and Communication Technologies (ICECCT).
[10] Qamar Mahmood, "Document similarity detection using semantic social network analysis on RDF citation," *graph2013 IEEE 9th International Conference on Emerging Technologies (ICET)*.
[11] Presidential Decree of South Korea 28799, "Regulations on the Management of National R&D Projects," 2018.4.17.
[12] Lee Kyung Mi, Seo Dong Ryul, Choi Jin Sook, "Extract Class Composition Hierarchy Information in Semi-structured Data Processing," 1997.
[13] C. D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval," Cambridge University Press. 100-123. ISBN 9780521865715. Scoring, term weighting, and the vector space model.
[14] http://nlplab.ulsan.ac.kr/doku.php?id=utagger

김 용 주

https://orcid.org/0000-0002-0625-5317
e-mail : kimyj@nfec.go.kr
2002년 한밭대학교 컴퓨터공학과(학사)
2004년 한밭대학교 컴퓨터공학과(석사)
2014년~2018년 한밭대학교 컴퓨터공학과 박사과정
2009년~현 재 국가연구시설장비진흥센터 연구원
관심분야 : Data Mining, Big Data Analytics

김 영 찬

https://orcid.org/0000-0003-3417-1935
e-mail : yckim@hanbat.ac.kr
1985년 아주대학교 전자공학과(학사)
1985년~1990년 삼성전자 연구원
1987년 한국과학기술원 전기및전자공학과 (석사)
1995년 한국과학기술원 전기및전자공학과(박사)
1996년 University of Arizona, Visiting Scholar
1997년~1998년 한국전자통신연구원 선임연구원
1998년~현 재 한밭대학교 컴퓨터공학과 교수
관심분야 : Database, ORM, Data Mining, Big Data