An Licensing Question and Answer Tracking Database using Unsupervised Keyword extraction

비지도학습 기반 키워드 추출을 이용한 추적가능 인허가 질의 응답 DB 구축

은형석[a,b*], 김동희[a*], 이윤희[a], 지은경[b], 권영미[c]

[a] 한국전력기술 원자로설계개발본부 [b] 한국과학기술원 전산학부 [c] 충남대학교 전파정보통신공학과

* Co-first author: hseun@kaist.ac.kr, doris@kepco-enc.com

연구 배경 및 필요성

인허가 질의 전산화 및 추적성 부재

- 원전 안전 관련 설계에 관한 규제기관의 인허가 질의응답 항목에는 한국표준원전의 설계 착수부터 새울 3,4호기 원전까지의 누적 데이터가 포함. 이 질의응답 자료를 누적관리하여 인허가 응답의 일관성을 유지하고 있음. 그러나 설계사가 소지한 표준원전부터의 인허가 질의 응답 항목은 상당수 비전산화 되어있었으며, 단순히 파일로만 관리되어 사용자의 검색능력에 의존해왔음.
- 만약 질의응답 정보를 색인화하고 정보를 연결할 수 있으면 다른 질의응답간의 추적성을 확보 가능하고 색인화된 정보를 이용하여 각종 통계를 산출할 수 있을 것이나, 기반 DB가 미구축되어 있었음.
- 설계정보가 포함된 인허가 질의응답 특성상, 외부망으로 연결하여 생성형 AI 및 지도학습을 사용하는 것이 어려움.

KR-WordRank

- **WordRank** : 문서에 대한 추출 요약 알고리즘
  - Chen et al.(2011) : 외부 경계 값 (주어진 단어의 좌우 주변에 다른 단어가 나타날 가능성)과 내부 경계 값(주어진 단어를 이루는 연속적인 글자의 응집성)을 모두 이용하여 단어를 인식하는 비지도학습 방법
  - Mutual reinforcing relationship 방법을 이용하여 중심성이 높은 후보 마디를 단어로 추출
  - 학습데이터를 사전에 구축하지 않고도 단어를 인식
- **KR-WordRank** : WordRank의 한국어 적용
  - Kim et al.(2014) : 기존의 WordRank보다 정교한 한국어 단어 추출이 가능함을 입증
  - 빅데이터 환경에서 작성된 온라인 문서의 요약 혹은 단어 추출과 같은 응용 연구에 활용될 수 있을 것

계측제어 SW간 중요도 정보 부재

- SC(Safety-critical) 시스템, ITS(Important-To-Safety) 시스템, 비안전 시스템의 SW 분류 → 계측제어계통의 소프트웨어 간에는 중요성에 질적인 차이가 존재함.
  - 예: PPS(Plant Protection System)와 DPS(Diverse Protection System)의 중요성은 기본 관계와 백업 관계에서 유추할 수 있나 ESF-CCS(Engineered Safety Features-Component Control. System) 및 CPCS(Core Protection Calculation System)의 중요성은 사용자 관점에 따라 달라질 수 있음
- 이러한 내용은 주관적이고 상대적이므로, **계통 간의 중요도를 정량화 하거나 분석하려는 시도는 한번도 이루어지지 않았음.**

LQATD (Licensing Question and Answer Tracking Database) 구축

Overall Approach

인허가 질의응답 데이터의 전산화

- 한국전력기술에서 보유하고 있는 3237개의 인허가 질의응답 데이터를 전산화. 전산화 대상 데이터 세트는 계측제어(I&C) 부서 또는 일부 안전 해석(SA) 부서와 관련된 문의로 한정. 분실되었거나 진행 중인 질의는 제외.

Projects	인허가 질의/응답 수
울진 5,6호기	906
영광 5,6호기	130
신고리 1,2호기	386
신월성 1,2호기	67
신고리 3,4호기	690
신고리 3,4호기 12발 제어봉 용역	29
신한울 1,2호기	765
새울 3,4호기	264
Total	3237 쌍의 인허가 질의 및 응답

전처리 및 키워드 추출

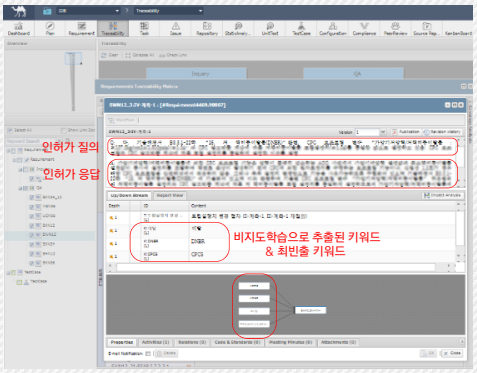
- 내부 보안망에서 비지도학습 KR-WordRank을 이용하여 질의내용, 대답내용을 각각 키워드 추출.
- 전처리 : 불용어처리를 위한 별도의 음절데이터를 사용하여 정보가 없는 단어를 제거. 원전별로 계통명칭이 다른 경우에는 하나의 대표명칭으로 통일 (ex. DPPS → PPS, CFMS → SPADES, RCOPS → CPCS)
- 질의, 대답 각각 2개씩 키워드 추출. 또한 가장 빈번하게 사용되는 단어 각각 2개를 추가 추출. 질의/응답 모두 별도로 키워드 추출을 하였기 때문에 질의응답별 8개의 키워드 추출

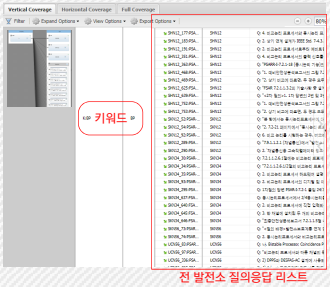
후처리 및 추적성 연결

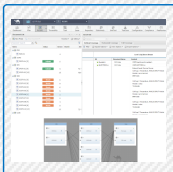
- 후처리 : 중복된 키워드 연결은 필요없으므로, 중복은 제거. 또한 키워드가 1글자이거나 무의미한 단어인 경우도 제거
- 후처리된 키워드와 질의응답은 SILKROAD (상용 형상관리 도구) 에 등록

질의 번호	질의 제목	질의 키워드	응답 키워드
SWN12_1:IV-I&C-1	설정치 변경 절차	CPCS, CEAC, 수정, 상수	Addressable, Constants, PLUS7, 수정
SKN34_54:PSAR-I-7.2-29	다양성	다양성, SAR, 설계, 테스팅	다양성, 트립, 채널, 시스템

LQATD (비지도학습 기반 키워드 추출을 이용한 추적가능 인허가 질의 응답 DB) 구축







기존 I&C 소프트웨어 RTM (Requirement Traceability Matrix) 과도 연결가능

→ 인허가 질의에서 세부 설계까지 확인가능

Keywords	질의/응답 수
시스템	1005
PPS	609
ESF-CCS	330
테스트	328
설계	323
신호	225
트립	224
채널	223
안전	218
소프트웨어	213
CPCS	171
제어	167
설정치	152
프로세서	148
ESFAS	138

〈질의 Top 15〉

Software	Software Importance
PPS	42.3 (BP: 22.2, CP: 20.1)
ESF-CCS	22.9
CPCS	11.9 (CPC: 7, CEAC: 2.2, CPP 2.6)
ESFAS	9.6
DPS	4.8
COLSS	3.6
QIAS-P	3.1
QIAS-N	1.1
SPADES	0.8

〈소프트웨어 중요도 지수〉

- 소프트웨어 중요도 지수 (Target Software Importance) $= \frac{\text{Number of Target Software Queries}}{\text{Total Number of Software Queries}} \times 100\%$
- PPS는 위의 나열된 소프트웨어 중 43.3%로 가장 많은 질문을 받았으며, ESF-CCS에 비해 약 1.8배, CPCS에 비해 3.6배 가량 더 질문을 받았음.
- PPS 서브프로그램인 BP, CP의 질의 비율 : 52.5% / 47.5%
- CPCS 서브 프로그램인 COPP(CPC), CEAP(CEAC), CCP(CPP)의 질의 비율 : 59.3%, 18.5%, 22.2%

결론 및 활용방안

결론 및 향후 연구

- 3237개 질의응답에 대한 전산화를 수행하고 KR-WordRank를 이용하여 질의응답에 대한 키워드 추출을 수행함.
- 추출된 키워드로 추적성을 연결하여 체계적인 인허가 질의응답 데이터베이스를 구축.
- 기존 소프트웨어 문서의 RTM과 결합하여 세부적인 소프트웨어 설계도 확인가능.
- 수많은 주제에 대한 다양한 통계분석이 가능해졌으며, 통계를 기반으로한 '소프트웨어 중요도' 라는 새로운 지수를 제안함.

활용방안 및 기대효과

- 자동 생성된 키워드들로 인해 인허가 질의응답 정보의 추적성 연결 가능

→

- 시각적으로 키워드를 확인하고 다른 질의를 추적할 수 있어 향후 인허가 질의응답의 일관성과 설계 일관성을 확보하여 인허가 질의 대응에 도움을 줄 수 있을 것으로 기대됨
- 소프트웨어 중요도 지수는 주관성을 배제한 통계 결과를 바탕으로 제시되었으므로, 소프트웨어 신뢰성의 정량적 평가에 활용된다면 보다 객관적인 소프트웨어 신뢰성 측정이 가능할 것으로 판단됨.

참고문헌

[1] Chen, S., Xu, Y., and Chang, H., "A simple and effective unsupervised word segmentation approach," In proceedings of the 25th AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 2011.

[2] H. Kim, S. Cho, and P. Kang, "KR-WordRank : An Unsupervised Korean Word Extraction Method Based on WordRank," Journal of Korean Institute of Industrial Engineers, 2014.

[3] SILKROAD RM. http://www.silkroadalm.com/main.do?menu_item=rm_overview.