Semantic Processing of Engineering Documents in PLM Environment

* KAIST 산업 및 시스템공학과

* 서효원 교수
* 전상민 박사과정/한국타이어
*김경근 박사과정/국방과학연구소
*최승아 석사과정

Contents

Contents

1. Background (AS-IS)

- 제품 개발 시 Engineering 문서 폭증
  - 요구사항 → 설계 → 해석 → 제조 → 시험 → 양산
  - 어디서? 어떻게? 원하는 문서를 빠르게 얻을 수 있을까?

- PLM이 보편화/안정화/고도화 단계
  - 문서의 저장/관리 보다 탐색/검색이 더 부각

- 기존 Engineering 문서의 검색
  - Keyword 검색 → 선택의 폭 너무 넓음

1. Background (TO-BE)

- 효율적인 Engineering 문서 검색을 위해,
  - 문서 Package관리가 아닌 Text 기반 Contents 관리
  - Keyword 검색이 아니 의미기반 검색

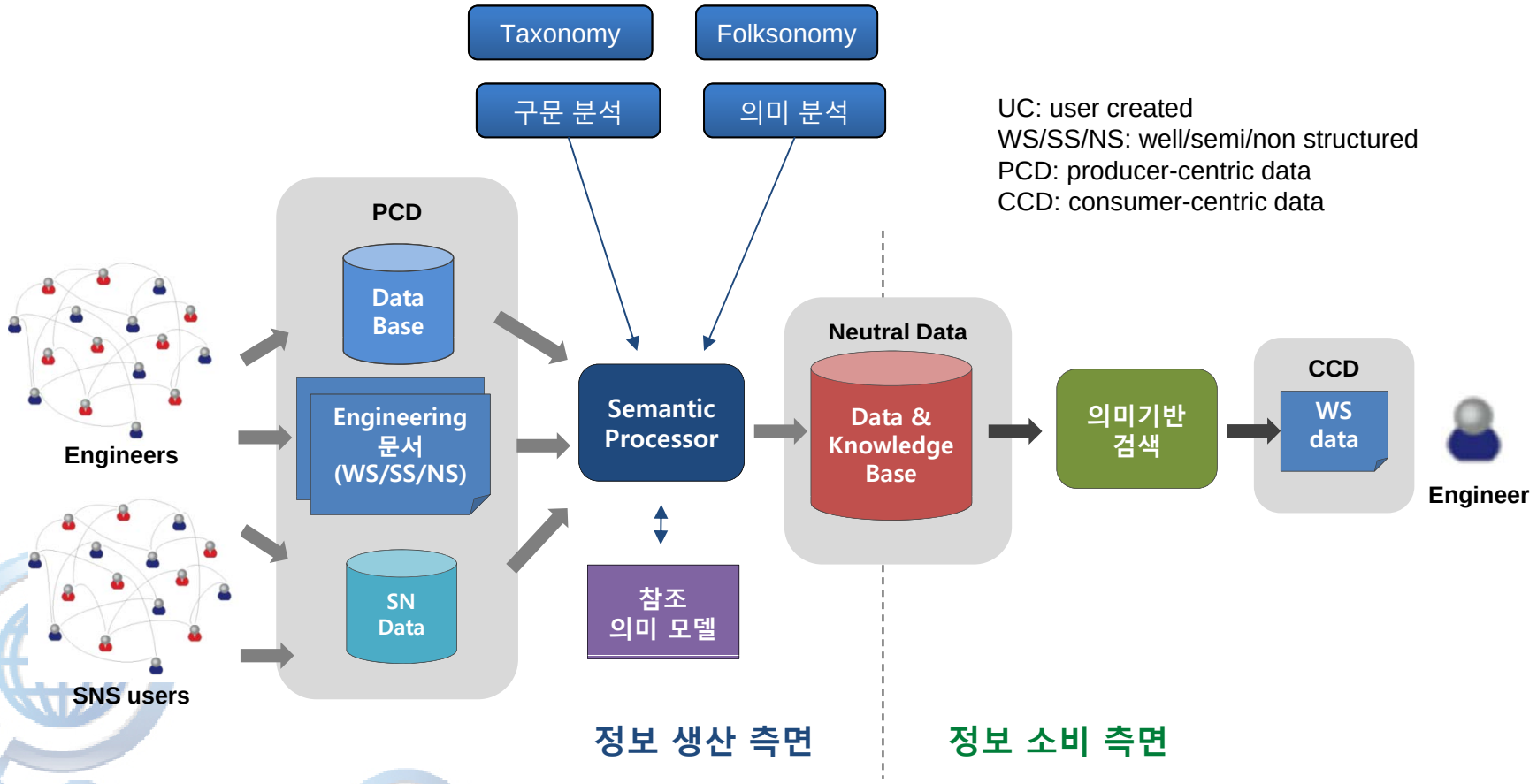
- 의미기반 검색을 위해,
  - 정보의 Semantics 구축 필요
  - 이를 기반으로, 문서의 Semantic Processing 진행

  *Semantic Processing = Syntax Processing(NLP) + Semantic Processing(Ontology)

- Semantic Processing 기반 Engineering 문서 관리
  - 정보 검색의 효율성 (↑)
  - 정보 재 활용성 (↑)
  - 정보의 통합성 (↑)

Contents

2. New Approach



S-NL (약식 자연어 처리) | 분야별 참조모델 온톨로지 | 의미표현 | 의미 유사도 평가 | 자기기반 검색

Contents

Research Trend (1/2)

1. Wu Ying-Han; Shaw Heiu-Jou, "Document based knowledge base engineering method for ship basic design", OCEAN ENGINEERING Volume: 38 Issue: 13 Pages: 1508-1521, SEP 2011

2. Wang Han-Hsiang; Boukamp Frank; Elghamrawy Tar, "Ontology-Based Approach to Context Representation and Reasoning for Managing Context-Sensitive Construction Information", JOURNAL OF COMPUTING IN CIVIL ENGINEERING Volume: 25 Issue: 5 Pages: 331-346, SEP-OCT 2011

3. Liu S.; McMahon C. A.; Culley S. J., "A review of structured document retrieval (SDR) technology to improve information access performance in engineering document management", COMPUTERS IN INDUSTRY Volume: 59 Issue: 1 Pages: 3-16, JAN 2008

4. S. Liu, C.A. McMahon *, M.J. Darlington, S.J. Culley, P.J. Wild, "A computational framework for retrieval of document fragments based on decomposition schemes in engineering information management", Advanced Engineering Informatics 20 (2006) 401–413

5. Zhanjun, L., Karthik, R., A., (2007), " Ontology-based design information extraction and retrieval " Artificial Intelligence for Engineering Design, Analysis and Manufacturing 21, pp. 137–154.

6. Zhanjun Li, Victor Raskin, Karthik Ramani, "Developing Engineering Ontology for Information Retrieval", Journal of Computing and Information Science in Engineering, 3.2008, vol 8

7. Zhanjun Li, Maria C.Yang, Karthik Ramani, "A methodology for engineering ontology acquisition and validation", Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 2009, vol 23

Research Trend (2/3)

8. Zhanjun Li Min Liu David C. Anderson Karthik Ramani, "Semantic-based design knowledge annotation and retrieval", Proceedings of IDETC/CIE 2005 ASME 2005 International Design Engineering Technical Conferences & Computer and information in Engineering Conference September 24-28, 2005, Long Beach, California, USA

9. Deeptimahanti Deva Kumar, Ratna Sanyal(2008) " Static UML Model Generator from Analysis of Requirements(SUGAR)" *2008 Advanced Software Engineering & Its Applications*, pp. 77–84.

10. Lin, JX ; Fox, MS ; Bilgic, T(1996) " A Requirement Ontology for Engineering Design" *Concurrent Engineering-Research and Iapplications, Vol 4, Issue3, pp. 279-291.*

11. Soner, K., Ozgur, A., Orkunt, S., Samet, A., Nihan, K.C., Ferda, N.A., (2012), " An ontology-based retrieval system using semantic indexing," Information Systems, 37, pp. 294-305.

12. Lin, M., H., (2009), " An optimal workload-based data allocation approach for multidisk databases" Data and knowledge Engineering, 68, pp. 499–508.

13. Patricia, L., (2000), " Information extraction from documents for automating software testing," Artificial Intelligence in Engineering, 14, pp. 63-69.

14. Module-based Failure Propagation (MFP) model for FMEA, Int J Adv Manuf Technol, Kyoung-Won Noh, Hong-Bae Jun, Jae-Hyun Lee, Gyu-Bong Lee, Hyo-Won Suh, 2011

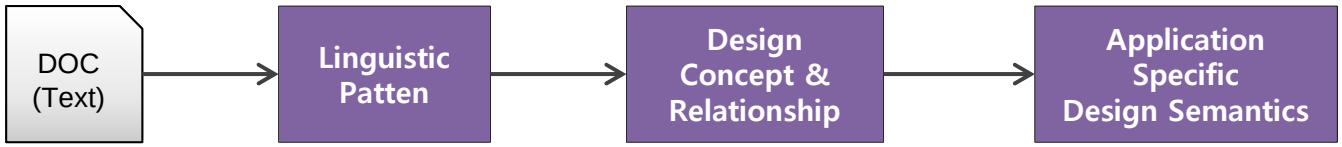
15. A Functional Basis for Engineering Design: Reconciling and Evolving Previous Efforts, NIST Technical Note 1447, Julie Hirtz, Robert B. Stone, Daniel A. McAdams, Simon Szykman, and Kristin L. Wood, 2002

Ontology-based design information extraction and retrieval

ZHANJUN LI and KARTHIK RAMANI
Artificial Intelligence for Engineering Design, Analysis and Manufacturing (2007), 21, 137–154.

1. Abstract

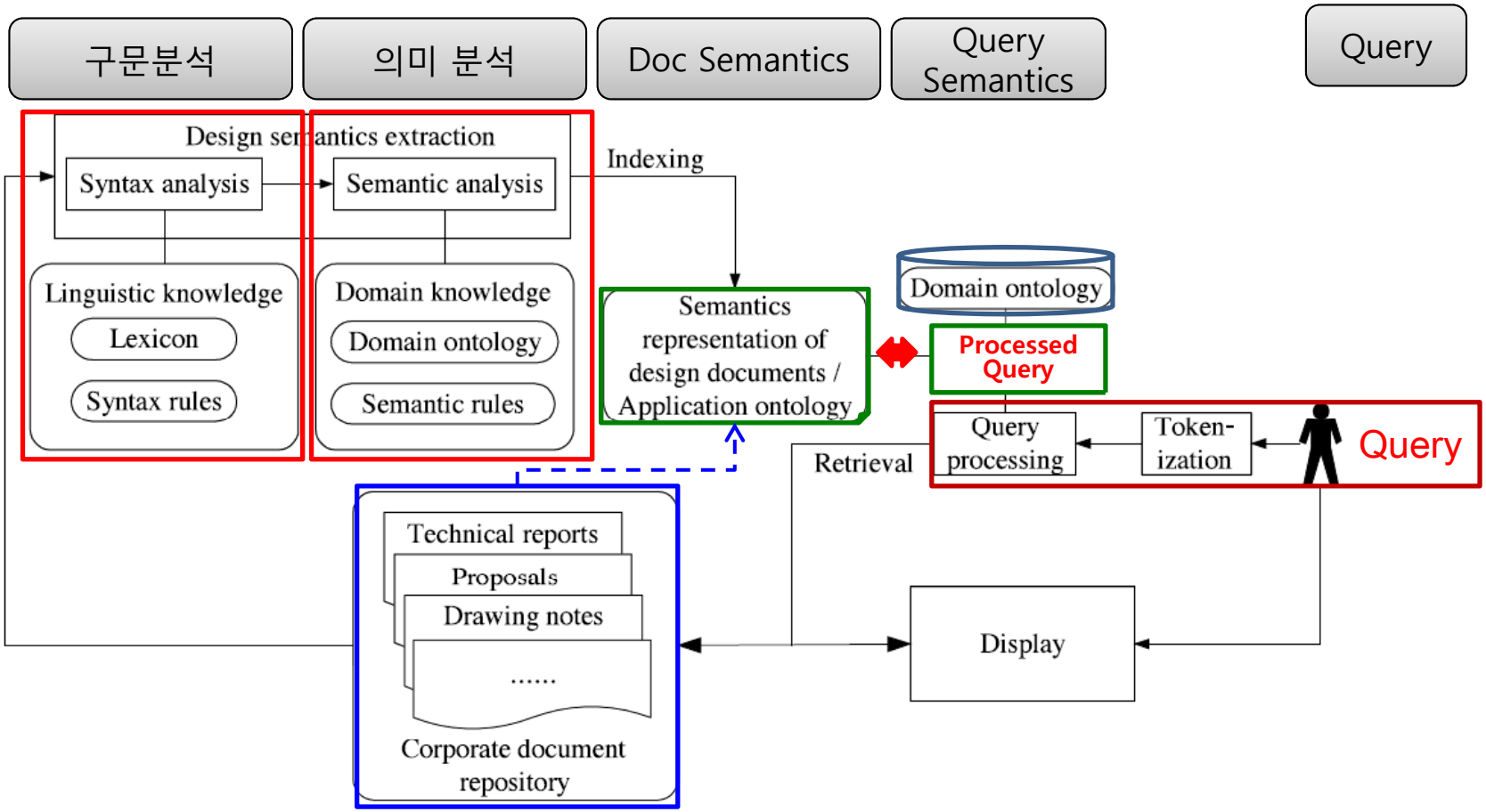
- Increasing complexity of product design process
  → the number of design documents has exploded
- To design information retrieval
  - Shallow natural language process(NLP)
  - Domain-specific design semantics/ontology
    ✓ Text/unstructured → structured/semantic-based representation



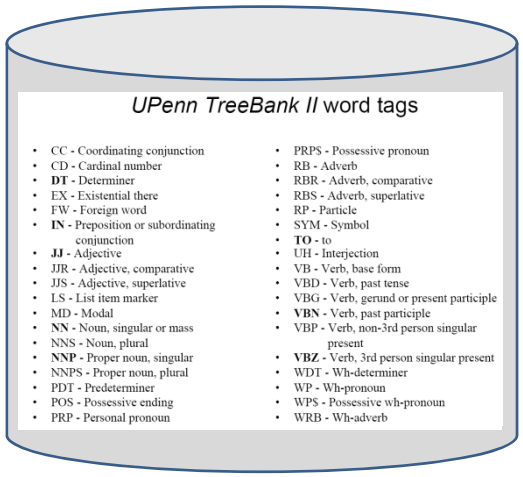
- To improve the performance of design information retrieval
  - Developed ontology-based query processing
    ✓ Users' requests are interpreted based on domain-specific meanings

2. System Architecture & Functional Diagram

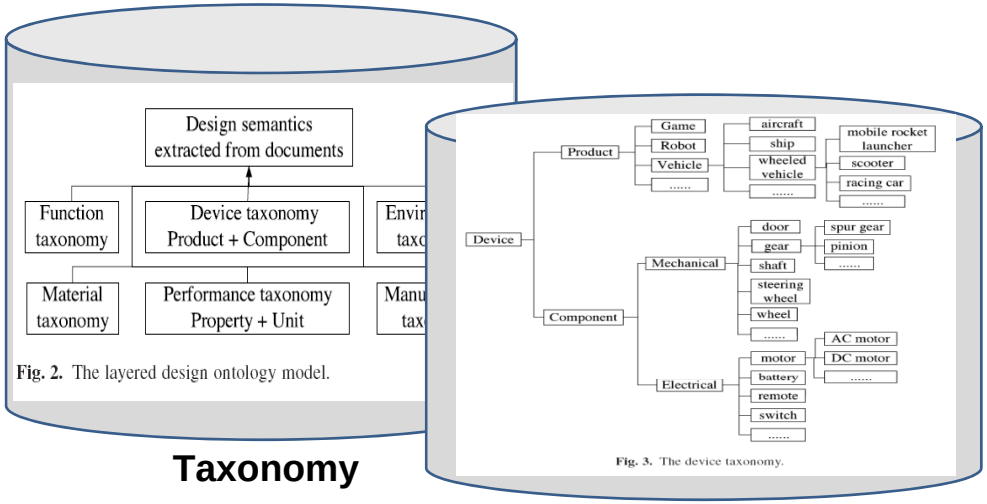
ODART: Ontology-based Design document Analysis and Retrieval Tool



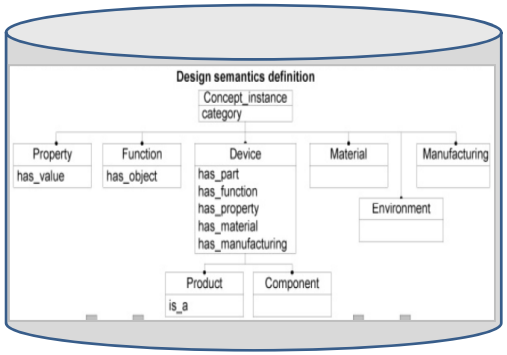
3. Ontology Modeling



< Linguistic Knowledge>



Taxonomy



Reference Model

< Domain Knowledge>

4. Design Semantic Extraction

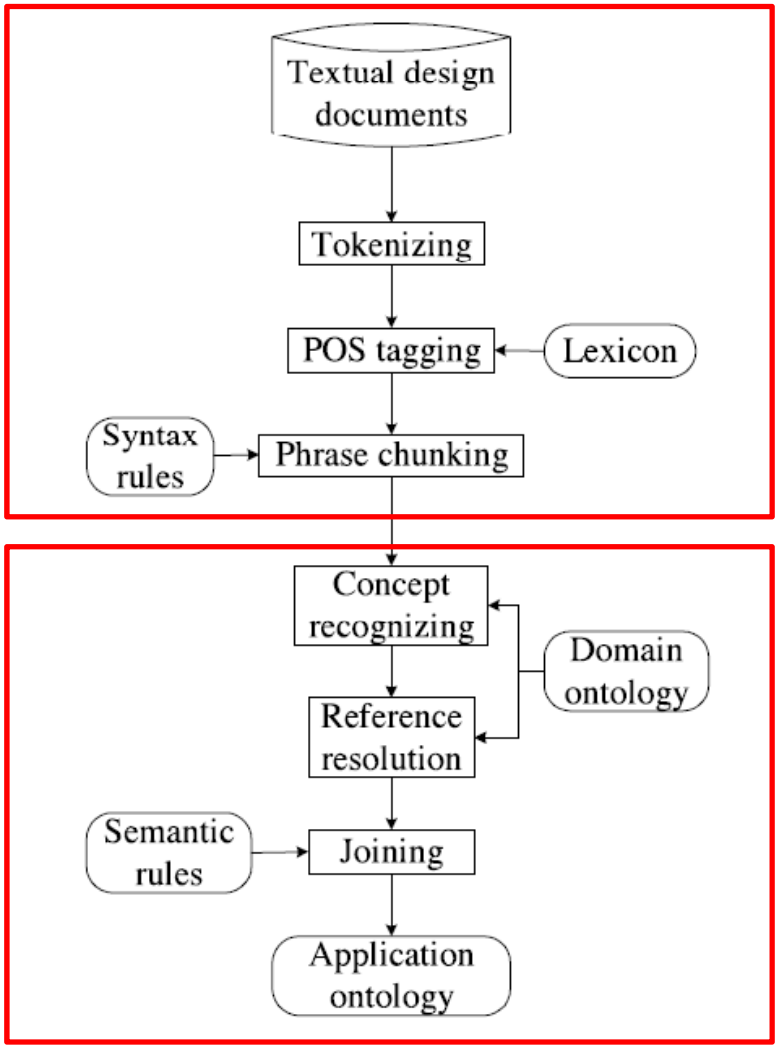
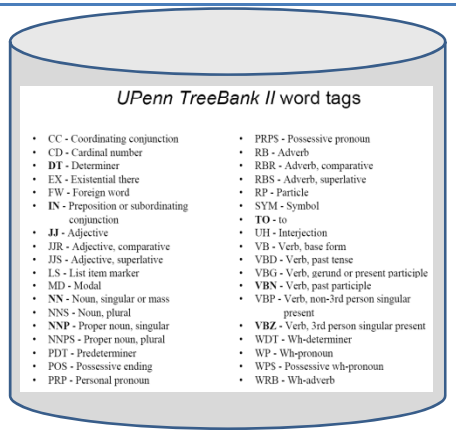
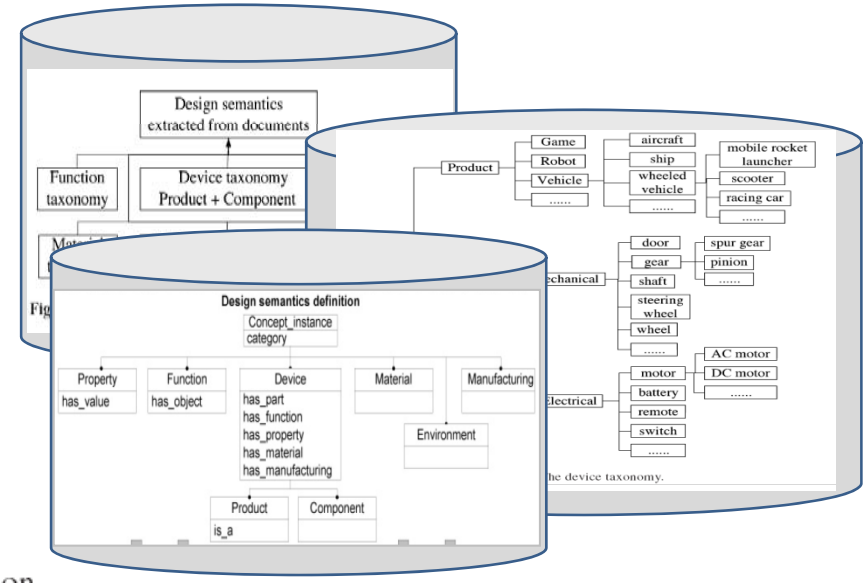


Fig. 6. The modules and procedure of design semantics extraction.



< Linguistic Knowledge>



< Domain Knowledge>

4. Design Semantic Extraction

Our action toy is a *mobile rocket launcher*. The *physical prototype* is capable of *moving* forward or backward driven by a *DC motor* with a *1:64 gear ratio*. A *second DC motor turns a pulley* that allows the *turret* to *rotate* 360 degrees clockwise or counterclockwise. The *rocket mount* can be *raised* and *lowered* by *hand* by placing the *base of the prop arm* into the slots……
The motors and the batteries are purchased from the local store. The *pulley* is made of *plastics* by STL……

a second DC motor turns a pulley that
DT JJ NN NN VBZ DT NN TDT
NP *AVP* *NP* *TDT*
CCI **FCI** **CCI** **TDT**
allows the turret to rotate 360 degrees
VBZ DT NN TO VBZ CD NNS
AVP *NP* *AVP* *MVP*
FCI **CCI** **FCI** **UCI**
clockwise or counterclockwise
RB CC RB
RB *CC* *RB*
RB **CC** **RB**

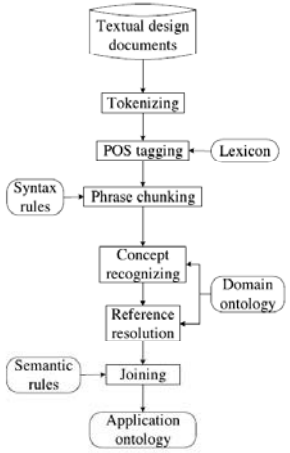
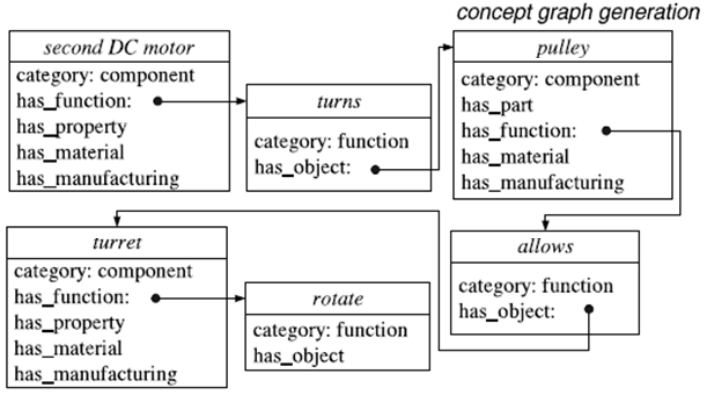


Fig. 6. The modules and procedure of design semantics extraction.



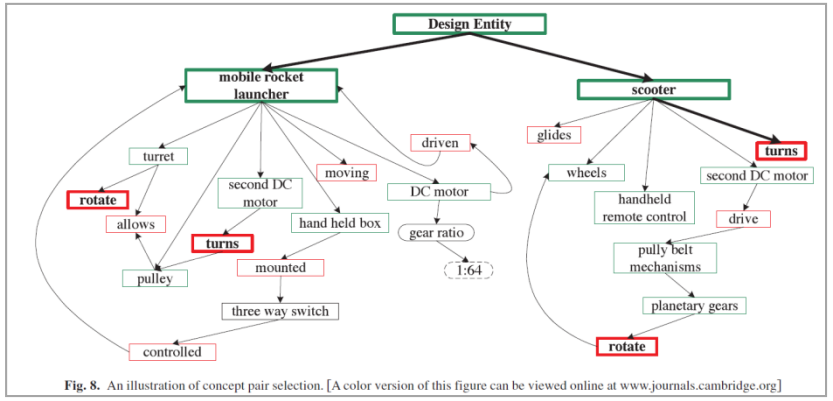
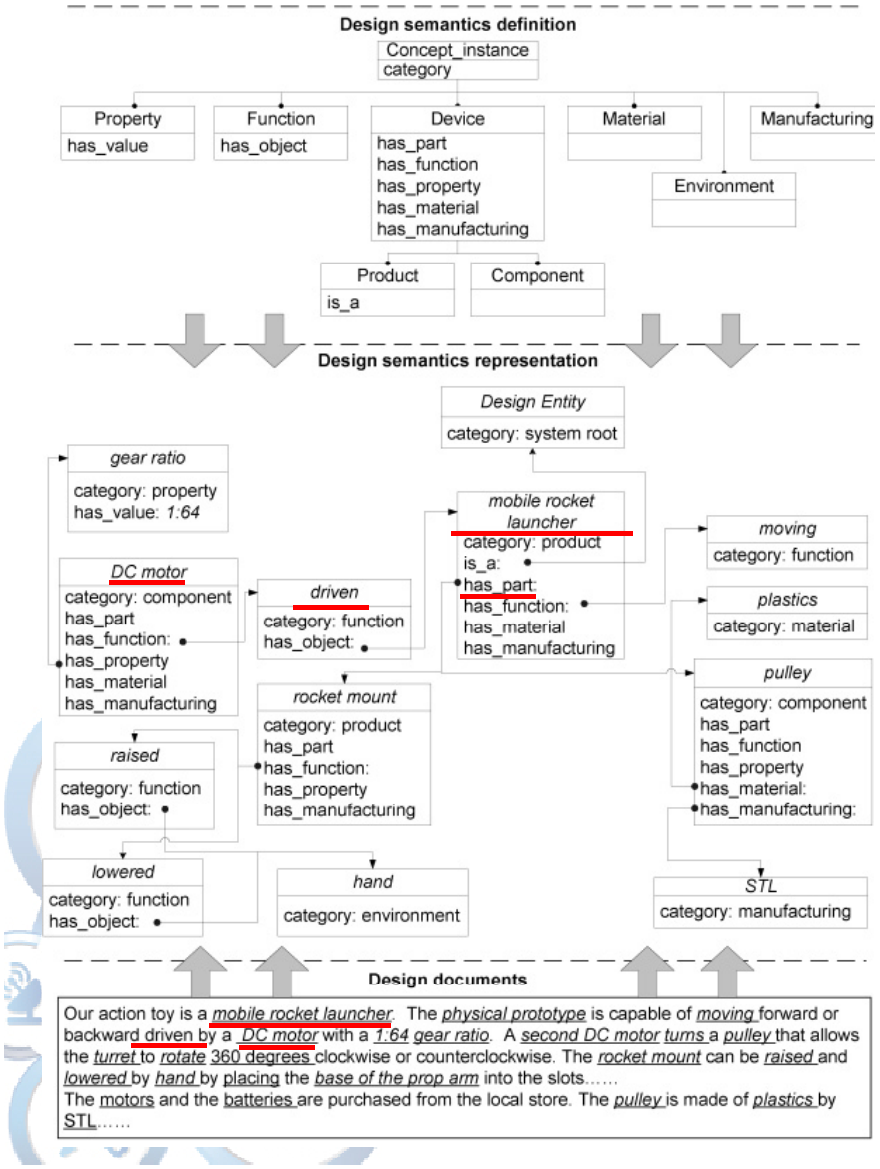
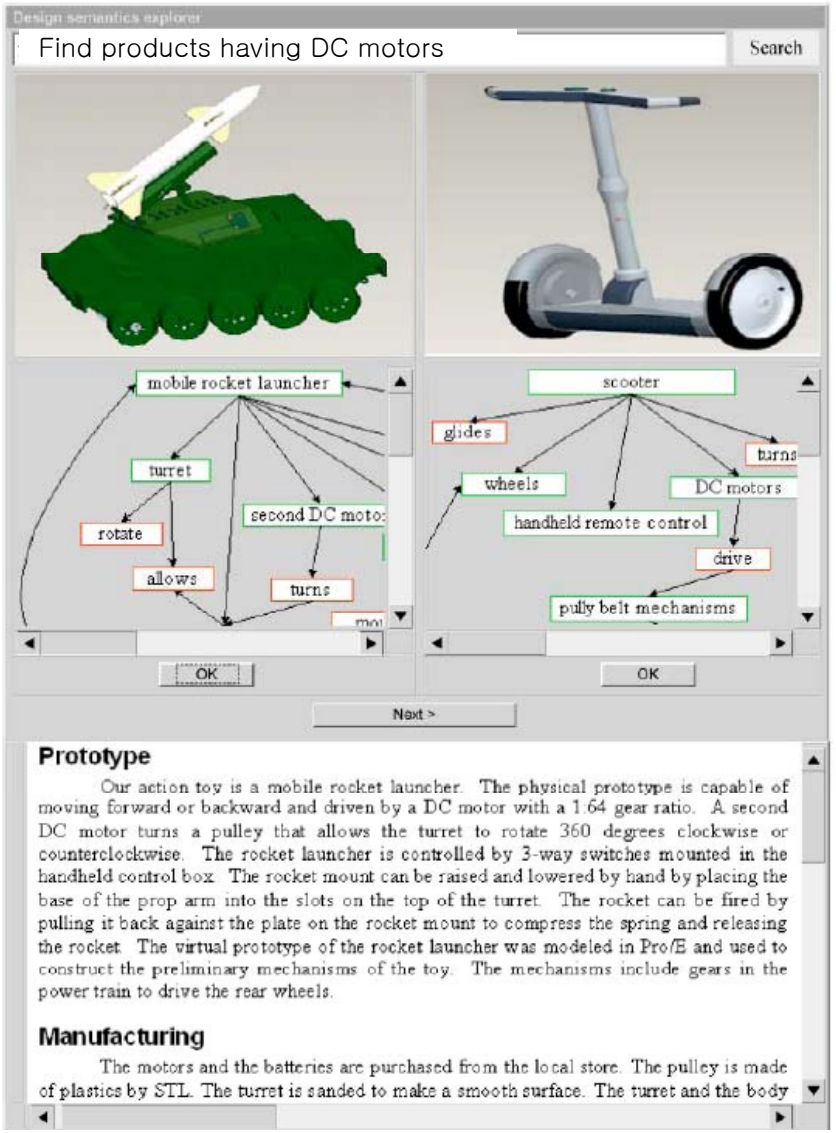


Fig. 8. An illustration of concept pair selection. [A color version of this figure can be viewed online at www.journals.cambridge.org]

5. Evaluation



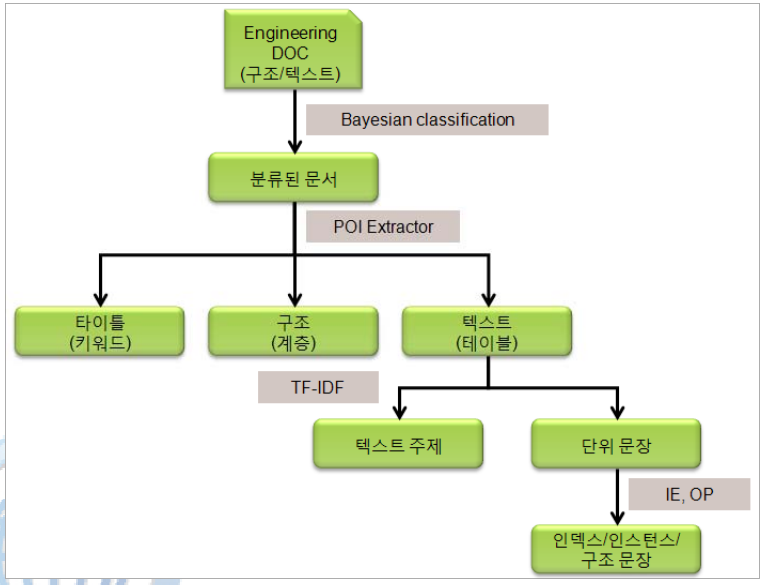


Contents

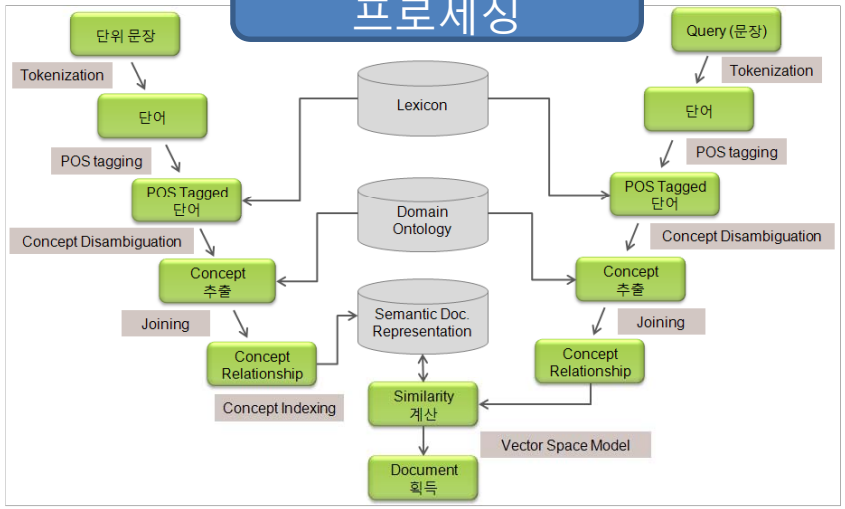
Introduction of Basic Algorithm for semantic document processing

1. 알고리즘 Outline

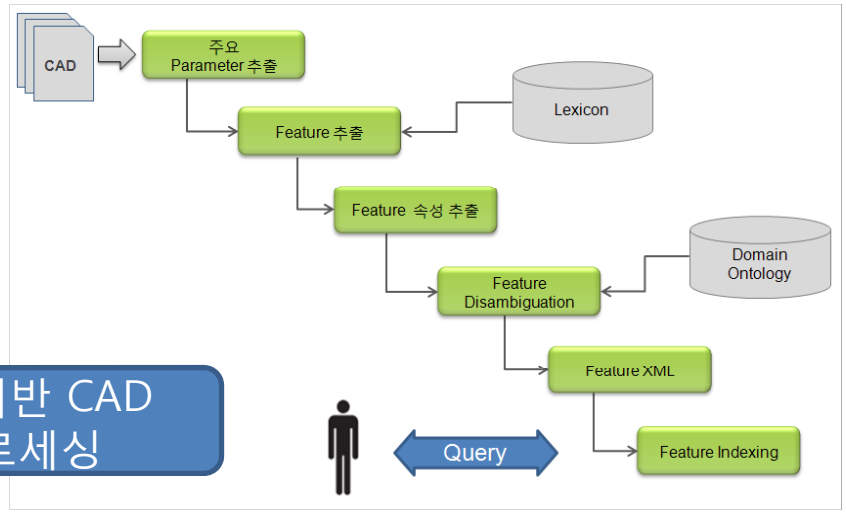
문서 프로세싱



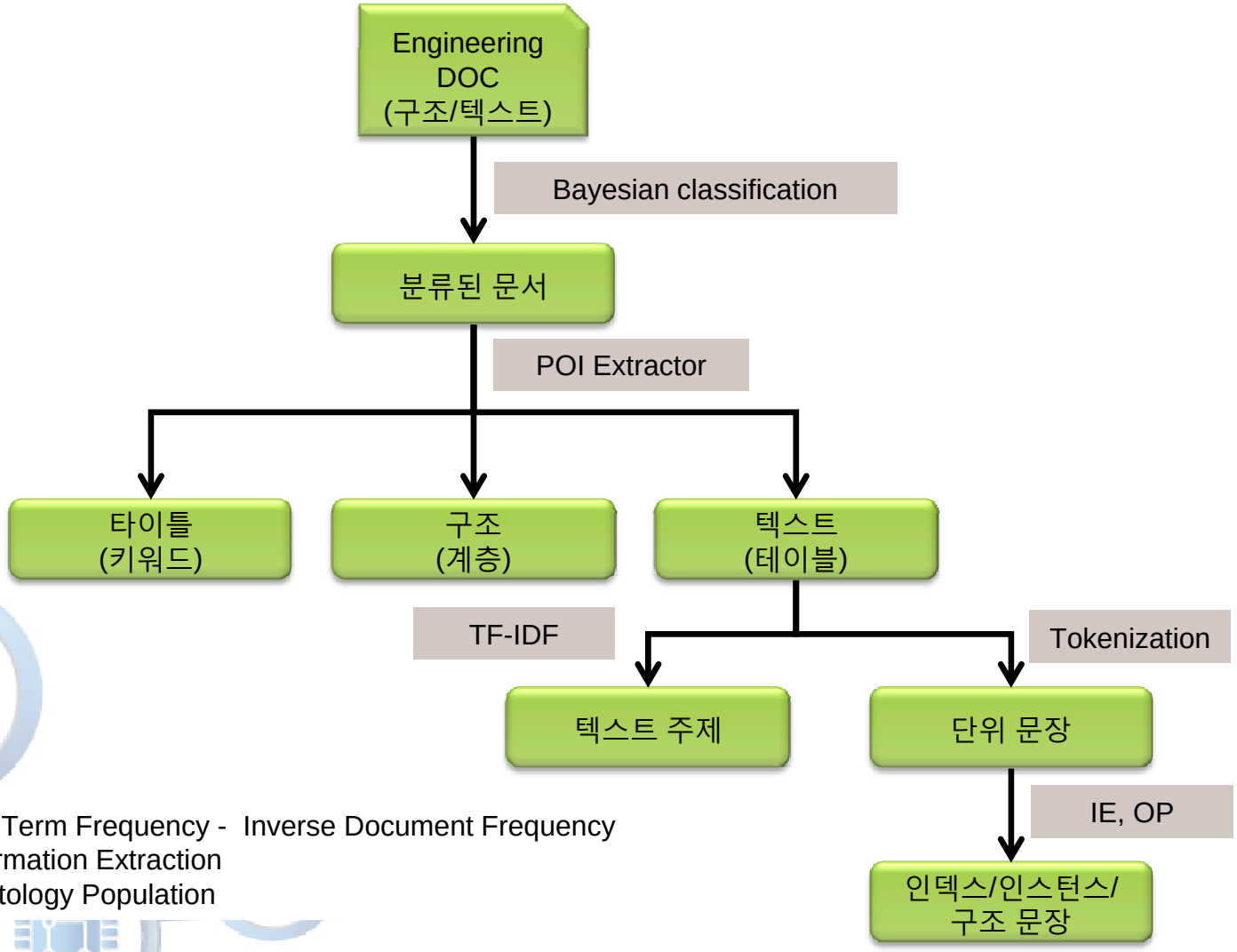
의미기반 텍스트 프로세싱



의미기반 CAD 프로세싱



2. 문서 프로세싱



•TF-IDF : Term Frequency - Inverse Document Frequency
• IE : Information Extraction
• OP : Ontology Population

2.1. 주요 알고리즘

Input

문서 ex) FMEA Doc.

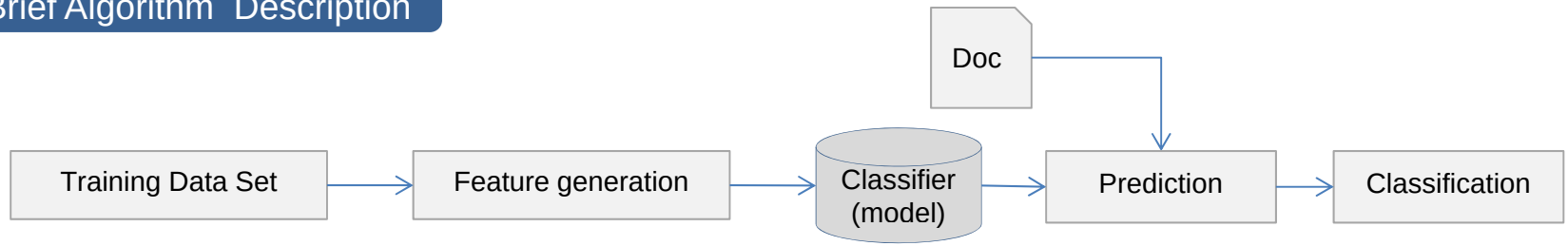
Algorithm

Bayesian classification

Output

분류된 문서 ex) Class 1

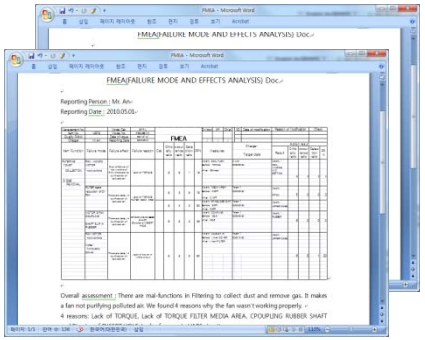
Brief Algorithm Description



Example

Training data sets

w1	w2	…	Class
FMEA	Failure Analysis		1
Evaluation	Measure		1
Concept	Design		2
Detail	Design		2
CAD	Drawing		3
…	…		..



(FMEA, Failure Analysis…)

Output : Class 1

사용 Reference

Training data sets

2.2. 주요 알고리즘

Input

분류된 문서

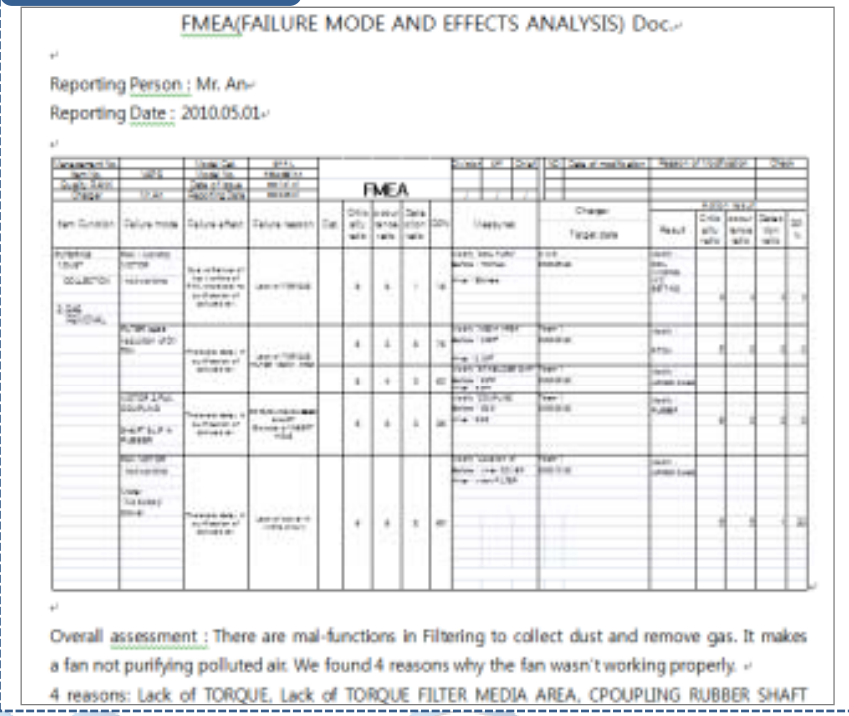
Algorithm

Apache POI Extractor API

Output

타이틀, 구조, Text 추출

Example



FMEA(FAILURE MODE AND EFFECTS ANALYSIS) Doc.

Reporting Person : Mr. An

Reporting Date : 2010.05.01

FMEA

Overall assessment : There are mal-functions in Filtering to collect dust and remove gas. It makes a fan not purifying polluted air. We found 4 reasons why the fan wasn't working properly.

4 reasons: Lack of TORQUE, Lack of TORQUE FILTER MEDIA AREA, CPOUPLING RUBBER SHAFT

Output :

-Title : FMEA(FAILURE MODE AND EFFECTS ANALYSIS) Doc.

-Structure :

Meta-data : Image : 0 개, Table : 1개, Text : 11줄,
작성자 : Mr.An, 문서생성일 : 2010.3.20
…

-Text : FMEA(FAILURE MODE AND EFFECTS ANALYSIS) Doc.
Reporting Person : Mr. An
Reporting Date : 2010.05.01
-Overall assessment : There are mal-functions in Filtering to collect dust and remove gas. It makes a fan not purifying polluted air…

사용 Reference

2.3. 주요 알고리즘

Input

문서(Text)

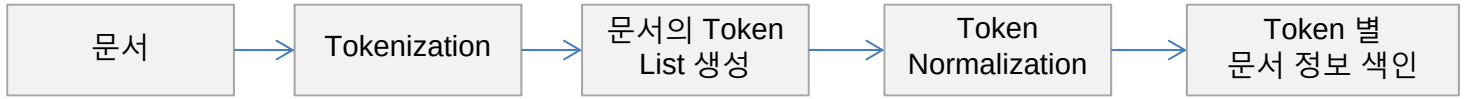
Algorithm

Inverted Index

Output

Inverted index

Brief Algorithm Description



Example

FMEA (FAILURE MODE AND EFFECTS ANALYSIS) Doc. Reporting Person : Mr. An. Reporting Date : 2010.05.01. There are mal-functions in Filtering to collect dust and remove gas. It makes a fan not purifying polluted air…

Stopword list

-A
-And
-Around
-Every
-For
-From
-In
-Is
-It
…

	Term	Document ID
1	FMEA	1
2	FAILURE	1.2
3	Design	1,3
4	Reporting	1,2,4
…	…	…
…	…	…

사용 Reference

2.4. 주요 알고리즘

Input

문서(Text)

Algorithm

TF-IDF

Output

Text Weighting (중요도)

Brief Algorithm Description

$$w_{t,d} = \log(1 + tf_{t,d}) \times \log_{10}(N / df_t)$$

Example

FMEA (FAILURE MODE AND EFFECTS ANALYSIS) Doc. Reporting Person : Mr. An. Reporting Date : 2010.05.01.
There are mal-functions in Filtering to collect dust and remove gas. It makes a fan not purifying polluted air…

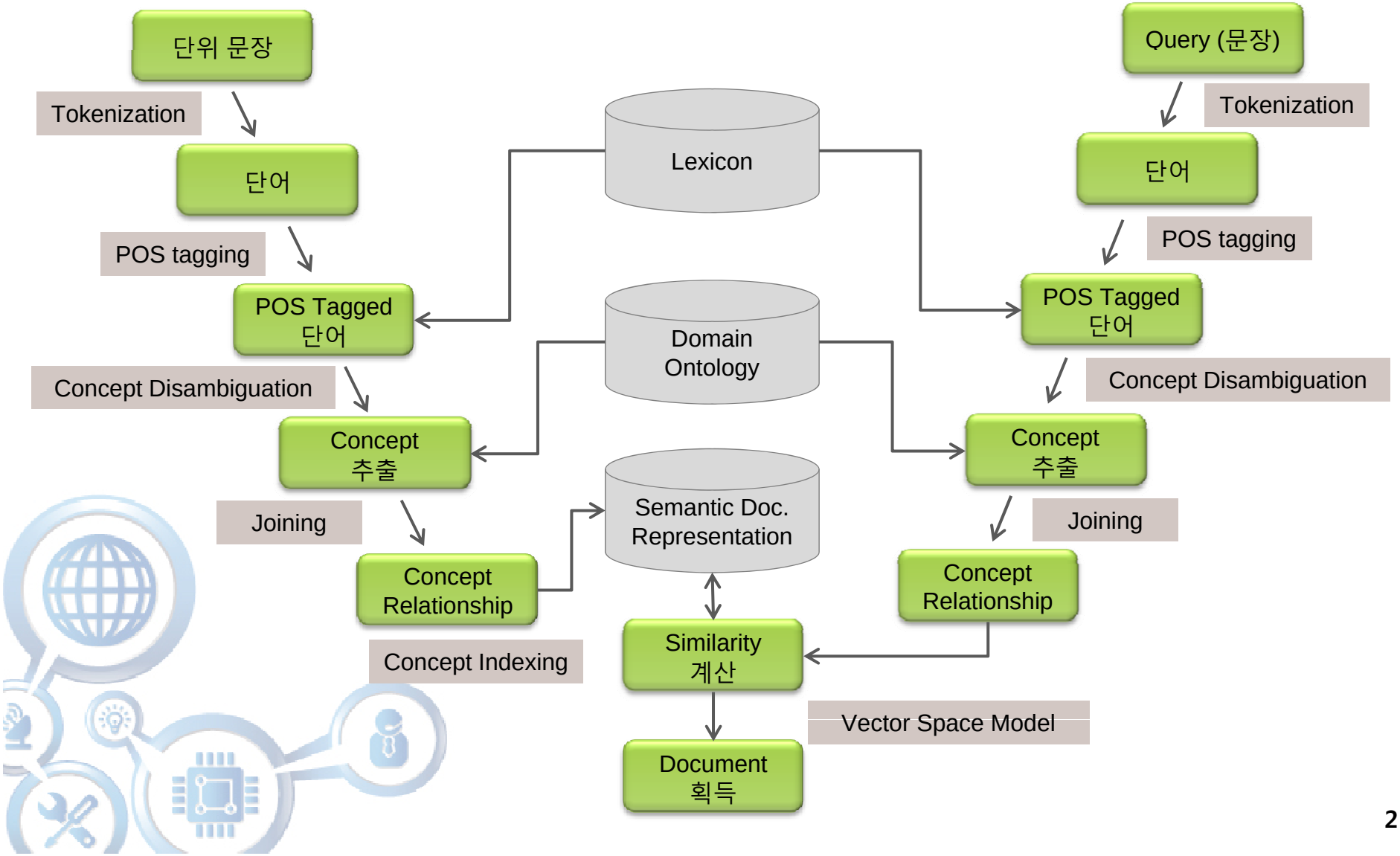
	Doc.1	Doc.2	Doc.3	…
FMEA	4	2	1	
Failure	5	2	4	
Lack	3	0	0	
Fan	13	0	0	
…	…	…	…	
…	…	…	…	

	Doc.1	Doc.2	Doc.3	…
FMEA	0.365	0.5228	0.157	
Failure	0.406	0.5228	0.365	
Lack	0.602	0	0	
Fan	1.146	0	0	
…	…	…	…	
…	…	…	…	

Output : 'Fan' is the most important word.

사용 Reference

3. 의미기반 텍스트 프로세싱 & 검색



3.1. 주요 알고리즘

Input

문장

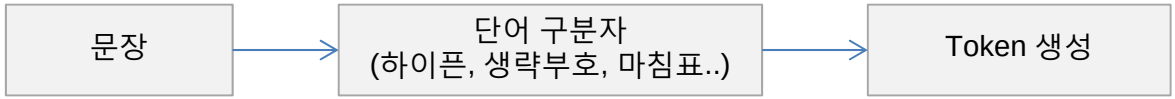
Algorithm

Tokenization

Output

Token(단어)

Brief Algorithm Description



Example

A second DC motor rotates air that makes a smell removed

A/ second/ DC/ motor / rotates/ air/ that/ makes/ a/ smell/ removed/

사용 Reference

3.2. 주요 알고리즘

Input

문장

Algorithm

POS(Part-of-Speech) tagging

Output

POS tagging

Brief Algorithm Description

Example

A/ second/ DC/ motor / rotates/ air/ that/ makes/ a/ smell/ removed.

A<DT> second<JJ> DC<NN> motor > rotates<VBZ>
air<NN> that<TDT> makes<VBZ> a<DT> smell<NN> removed<VBZ>.

DT – Determiner
JJ - Adjective
NN - Noun, singular or mass
VBZ - Verb, 3rd person singular present
TO – to
CD - Cardinal number
NNS - Noun, plural
RB – Adverb
CC - Coordinating conjunction
…

사용 Reference

Lexicon DB

3.3. 주요 알고리즘

Input

한글 문장

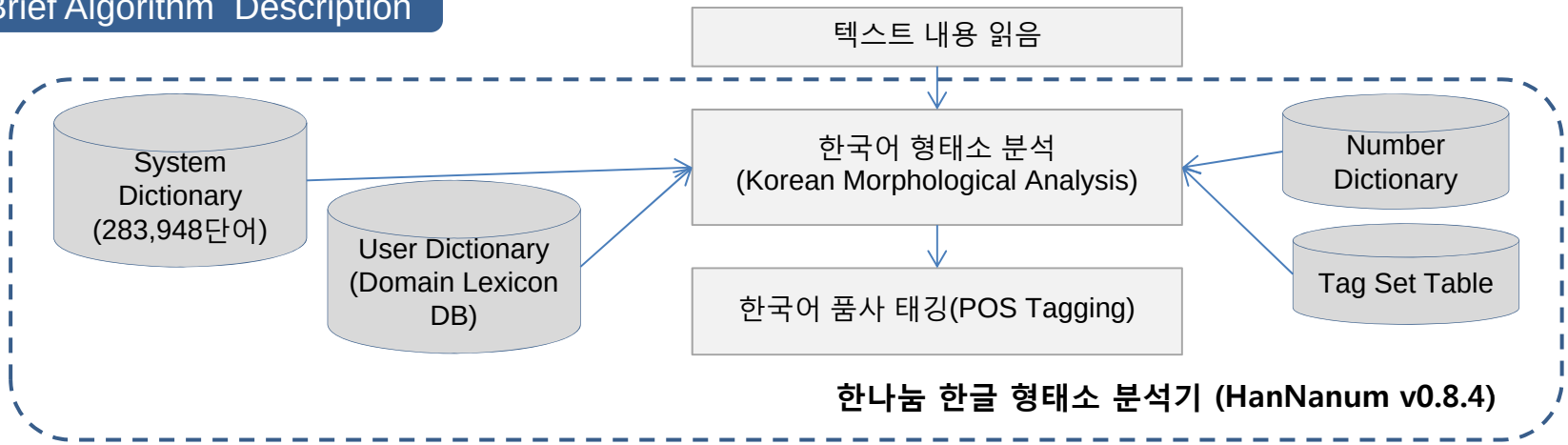
Algorithm

한글 형태소 분석

Output

한글 POS tagging

Brief Algorithm Description



Example

유도탄 중량을 현재 100kg에서 80kg으로 감량 설계 추진 요망

유도탄/NC 중량/NC 100/NN kg/F 에/PV 80/NN kg/F 감량/NC 설계/NC 추진/NC 요망/NC

사용 Reference

System Dictionary, User Dictionary, Number Dictionary, Tag Set Table

3.4. 주요 알고리즘

Input

문장

Algorithm

Concept Disambiguation

Output

문장과 가장 유사한 Concept

Brief Algorithm Description

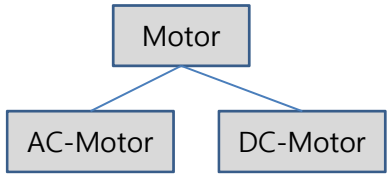
$$Tscore_{ij} = \frac{(\#\text{ of words in the phrase } T_j \text{ matches with}) * \sum_{m=1}^{N} \omega_m}{\#\text{ of words in the phrase}},$$

$$Cscore_i = \max(Tscore_{ij}) \quad 1 \le j \le n.$$

-Wm : weight of phrase
(contain only one word : Wm = 1,
No match : Wm = 0,
Rightmost matched : 0.55
[가장 매칭되는 단어에게 높은 점수]
Rest of : split 0.45 equally)

Example

ex) phrase : “second DC Motor”
0.225 0.225 0.55
Concept 이 “Motor” 일 경우,
Tscore= (1*(0+0+0.55))/ 3 = 0.183
Concept이 “AC Motor”일 경우,
Tscore= (1*(0+0+0.55))/ 3 = 0.183
Concept이 “DC Motor”일 경우
Tscore= (2*(0+0.225+0.55))/ 3 = 0.516 → DC Motor 가 가장 높은 점수로 선택됨.



사용 Reference

3.5. 주요 알고리즘

Input

문장

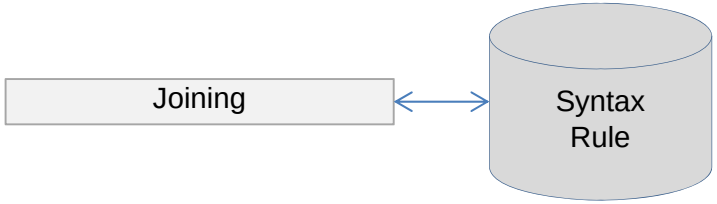
Algorithm

Joining

Output

Concept Relationship

Brief Algorithm Description



Example

Input : DC motor rotates fan.
→DC<NN> motor<NN> rotates<VBZ> fan<NN>.

Syntax Rule : NN^VBZ^NN -> VBZ(NN,NN)
→ Rotate(DC Fan, air)

사용 Reference

3.6. 주요 알고리즘

Input

문서(Text)

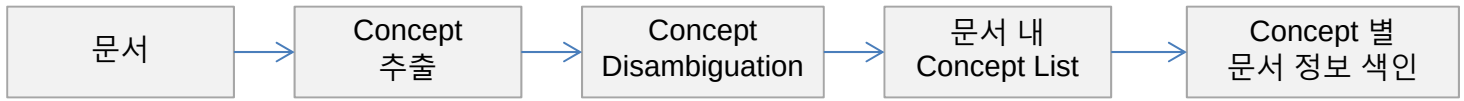
Algorithm

Concept Index

Output

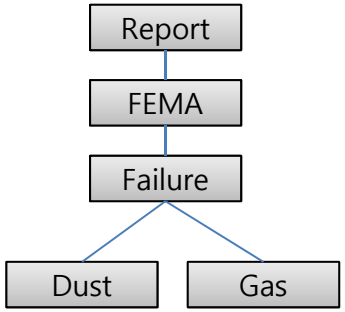
Concept index

Brief Algorithm Description



Example

FMEA (FAILURE MODE AND EFFECTS ANALYSIS) Doc. Reporting Person : Mr. An. Reporting Date : 2010.05.01. There are mal-functions in Filtering to collect dust and remove gas. It makes a fan not purifying polluted air…



	Term	Document ID
1	FMEA	1
2	FAILURE	1.2
3	Dust	1,3
4	Gas	1,2,4
…	…	…
…	…	…

사용 Reference

3.7. 주요 알고리즘

Input

Query, 문서들

Algorithm

Vector Space Model (Similarity)

Output

Query와 가장 유사한 문서

Brief Algorithm Description

$$\mathrm{sim}(D_j, Q_i) = \mathrm{cosine}(D_j, Q_i) = \frac{\vec{D}_j \cdot \vec{Q}_i}{|\vec{D}_j||\vec{Q}_i|} = \frac{\sum_{i=1}^{t} \omega_{i,j} \times \omega_{i,q}}{\sqrt{\sum_{i=1}^{t} \omega_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} \omega_{i,q}^2}},$$

Example

문서 내 각 단어의 Score, ex) FMEA, Failure, Dust, Gas…)

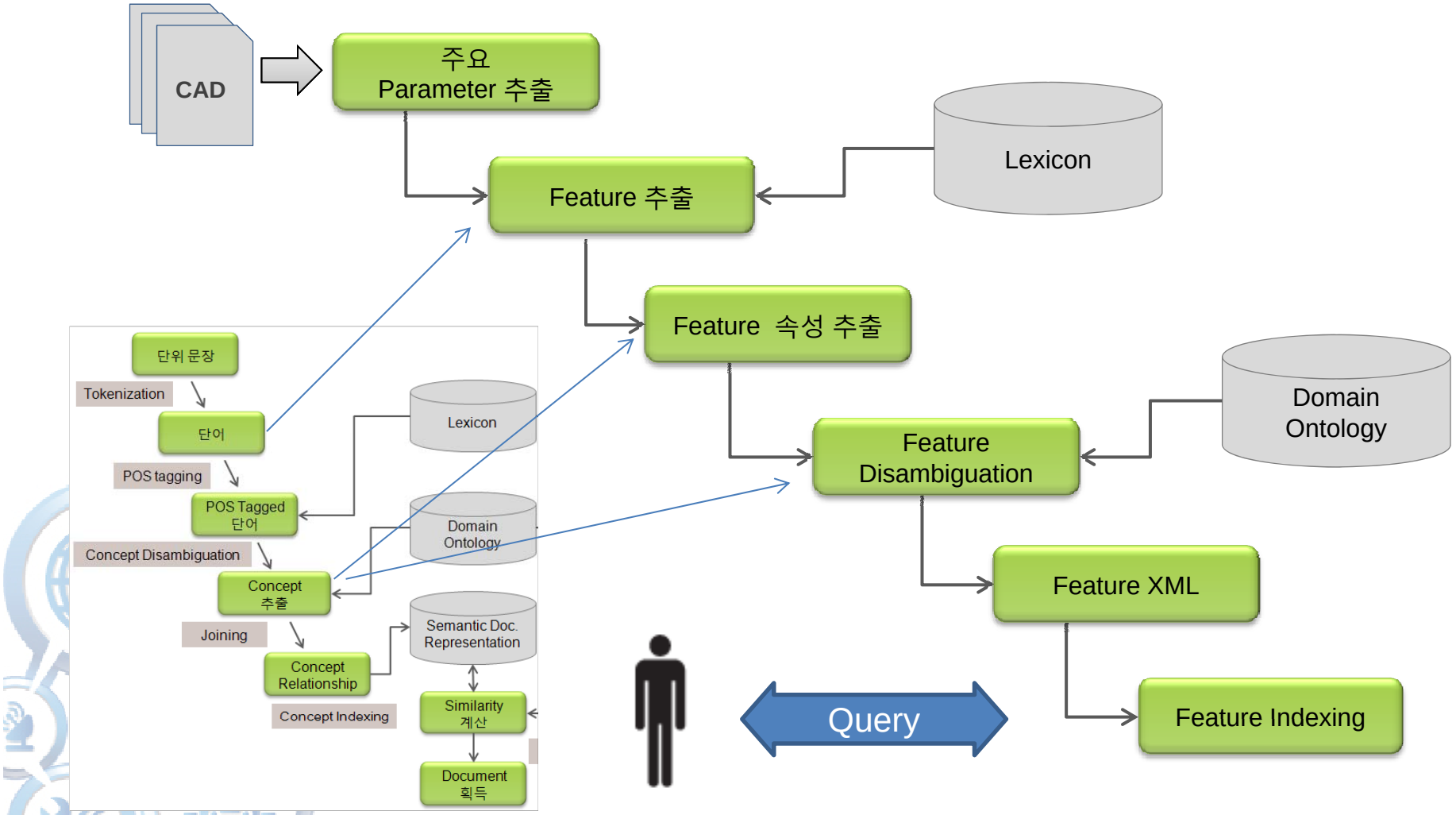
d1 = <9 6 3 2 2 2 0 2 0 0 0 0 >
d2 = <4 3 1 1 0 0 0 2 2 4 2 1 >
d3 = <1 1 1 1 1 1 1 1 1 1 0 0 0 >
d4 = <1 0 0 0 2 0 0 2 1 1 1 1 >

cos(d2, d1) = (36+18+3+2+4) / (7.5*11.9) = 0.71
cos(d2, d3) = (4+3+1+1+2+2) / (7.5*3) = 0.58
cos(d2, d4) = (4+4+2+4+2+1) / (7.5*3.6) = 0.63

D2와 D3가 가장 유사함

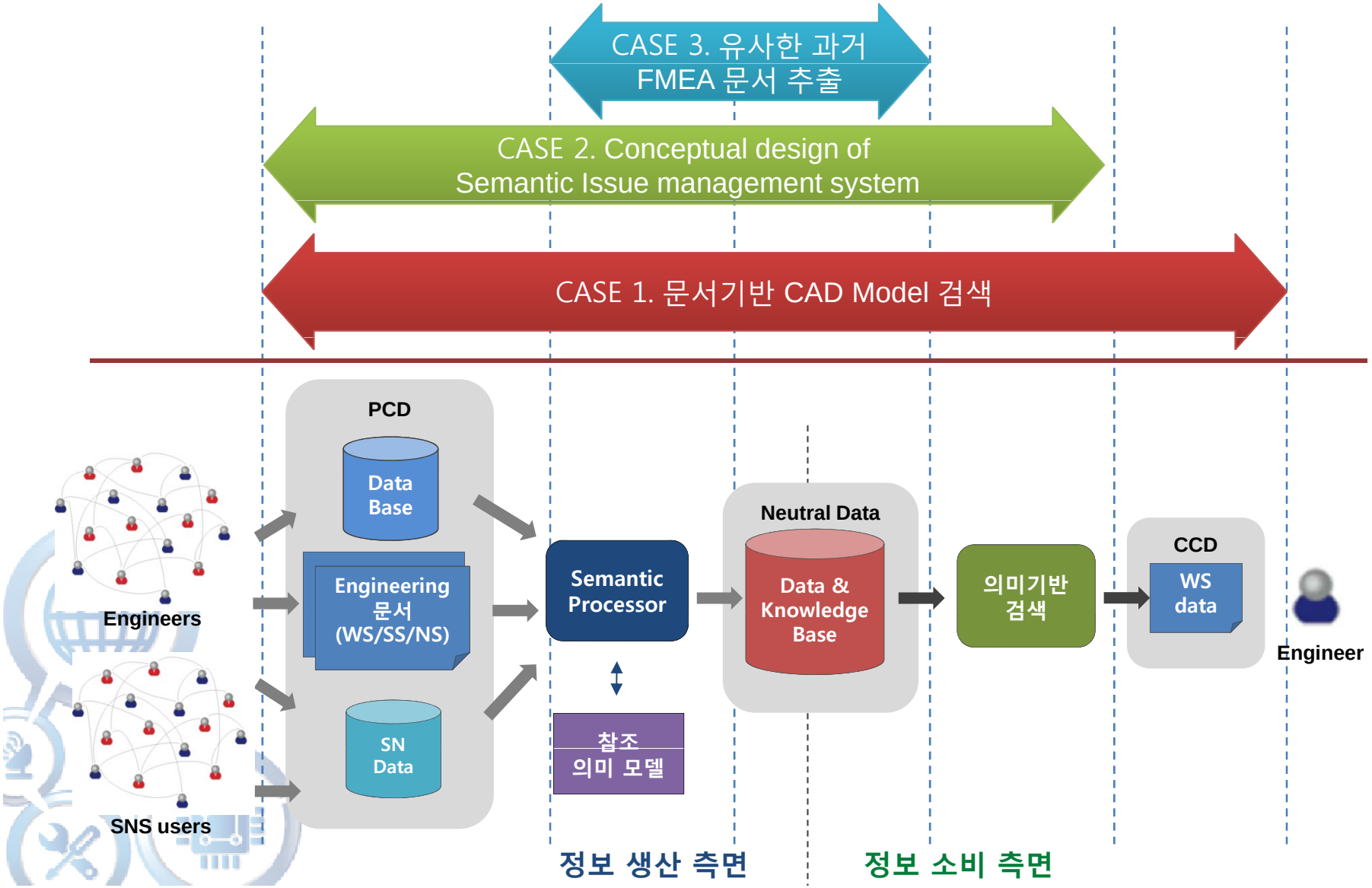
사용 Reference

4. 의미기반 CAD 모델 추출 및 검색



Contents

Case Study Outline



CASE STUDY 1 :

문서 기반 CAD Model 검색

1. 재사용을 위한 CAD Model 검색의 문제점

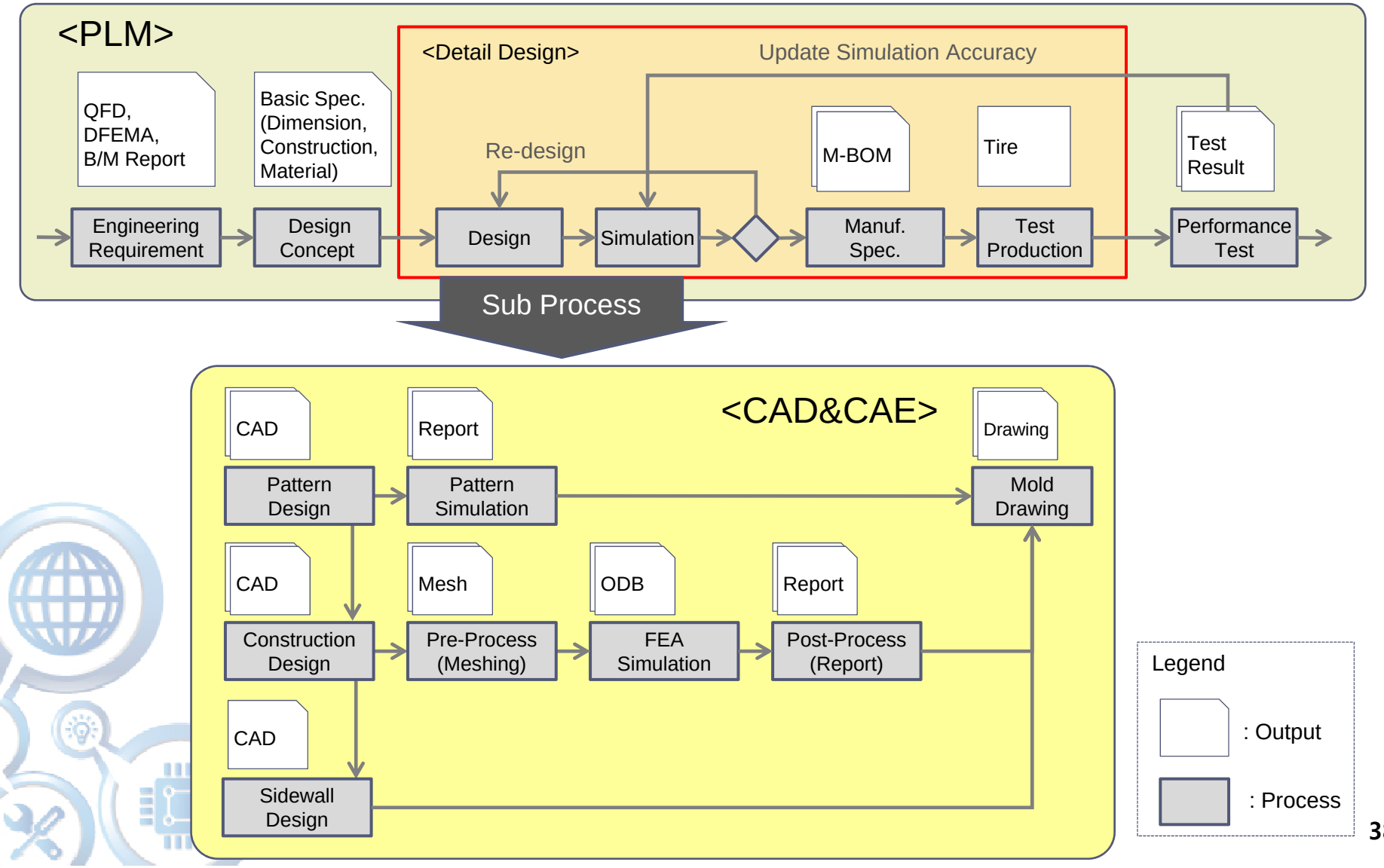
- CAD Model의 재사용
  - 제품 설계 시 80%의 CAD Model이 재사용 되고 있음.
  - 성공적인 재사용은 50%의 비용 절감 효과 발생.

- 현재 CAD Model 검색 방법
  - CAD File 이름을 통해
  - 정보 시스템(PLM,PDM)의 정보(BOM) 기반 검색
    ➔사용자가 의도한 검색에 한계 (검색 결과 중 48%는 사용 불가[1])

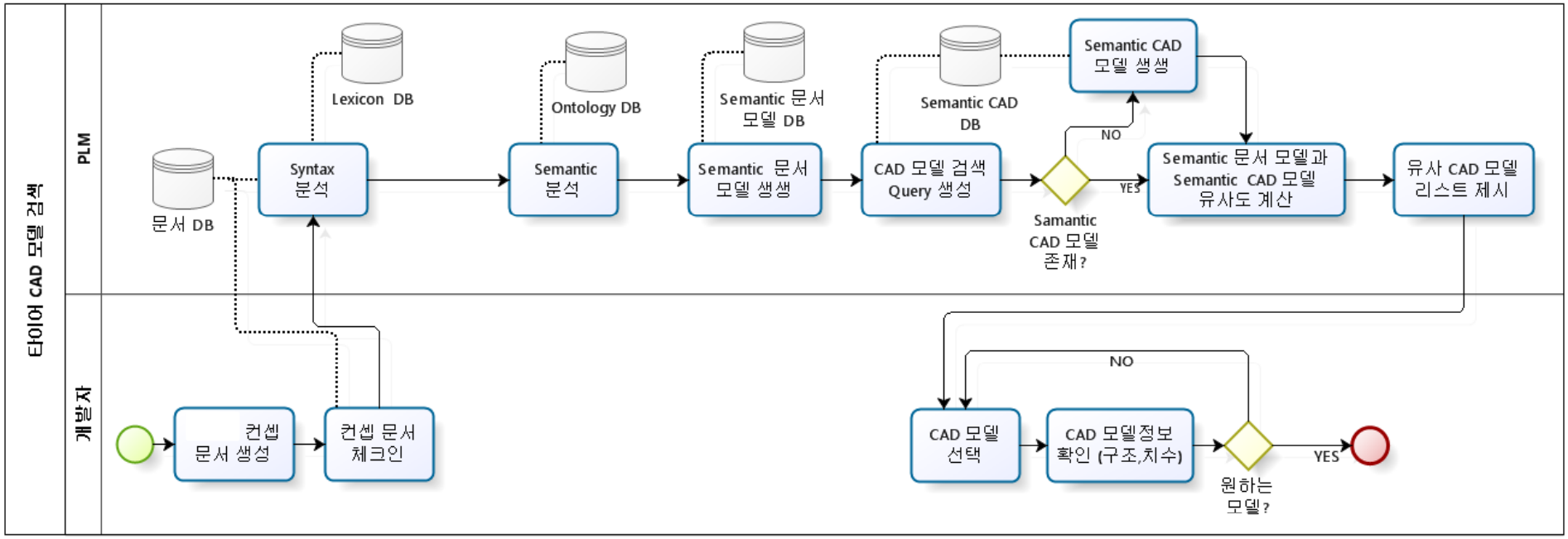
- 보다 효과적인 검색 및 재사용을 위해서
  - 단순 이름 및 정보 기반이 아닌 Semantic CAD Model검색 필요
  - CAD Model에 대한 Knowledge 추출 및 재사용 필요

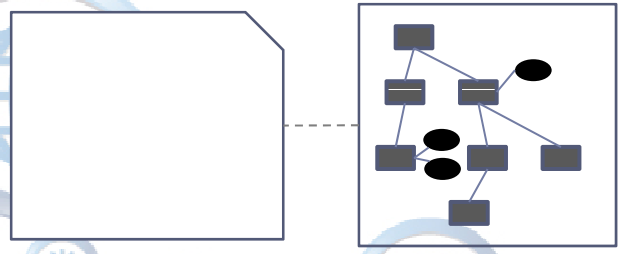
[1] LI, M, Y F Zhang*, J Y H Fuh and Z M Qiu, "Towards effective mechanical design reuse: feature-based CAD model retrieval on general shapes and partial shapes". JOURNAL OF MECHANICAL DESIGN, (2009).

2. CAD Model 재 사용 예



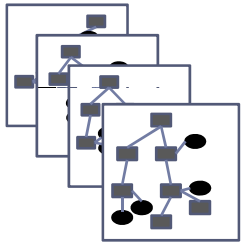
3.문서 기반 CAD 모델 검색 프로세스 (1/2)





Concept 문서

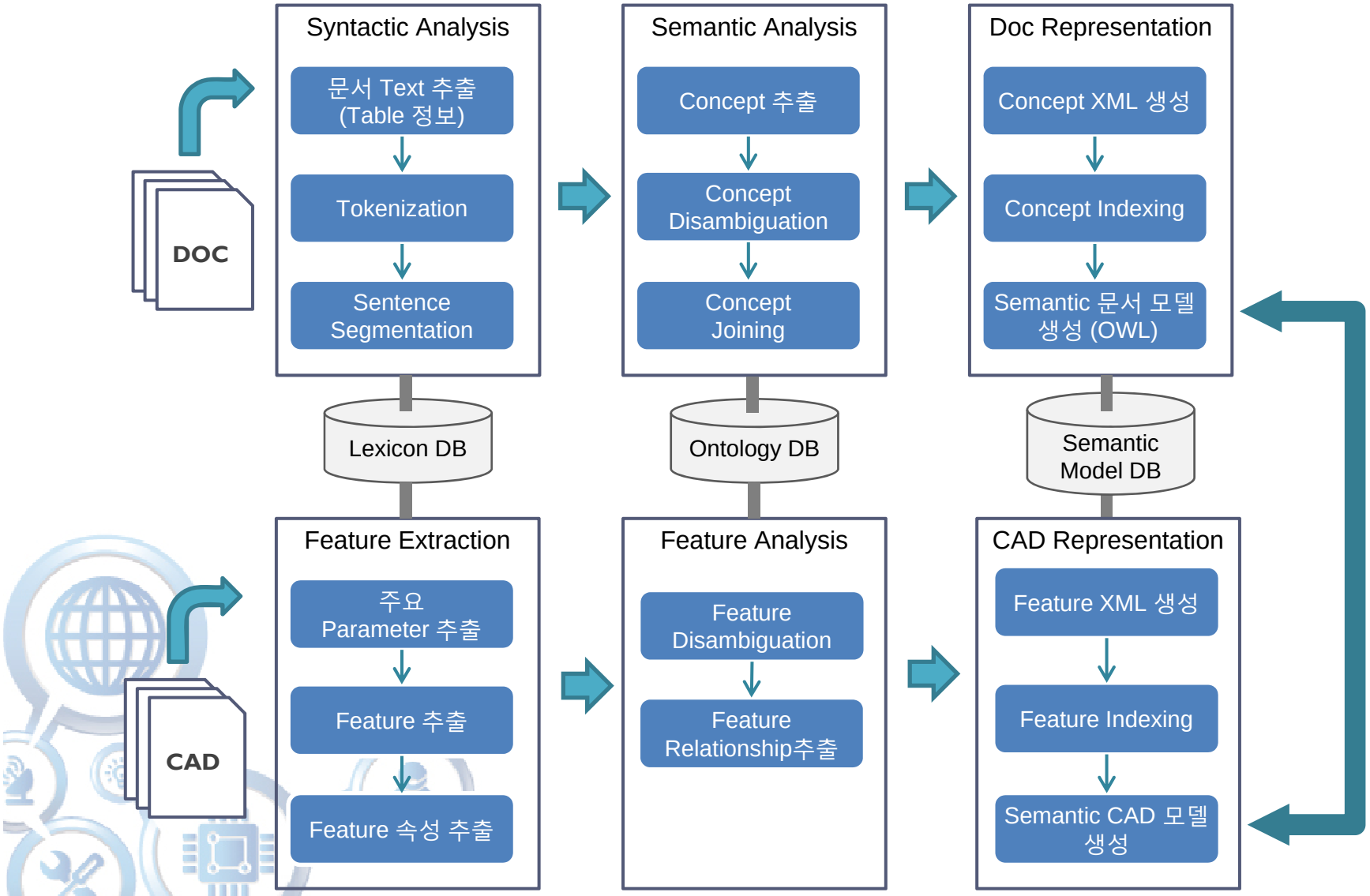
Semantic 문서 모델



Semantic CAD 모델

CAD 모델

3.문서 기반 CAD 모델 검색 프로세스 (1/2)



CASE STUDY 2 :

Conceptual design of Semantic Issue management system

1. AS-IS

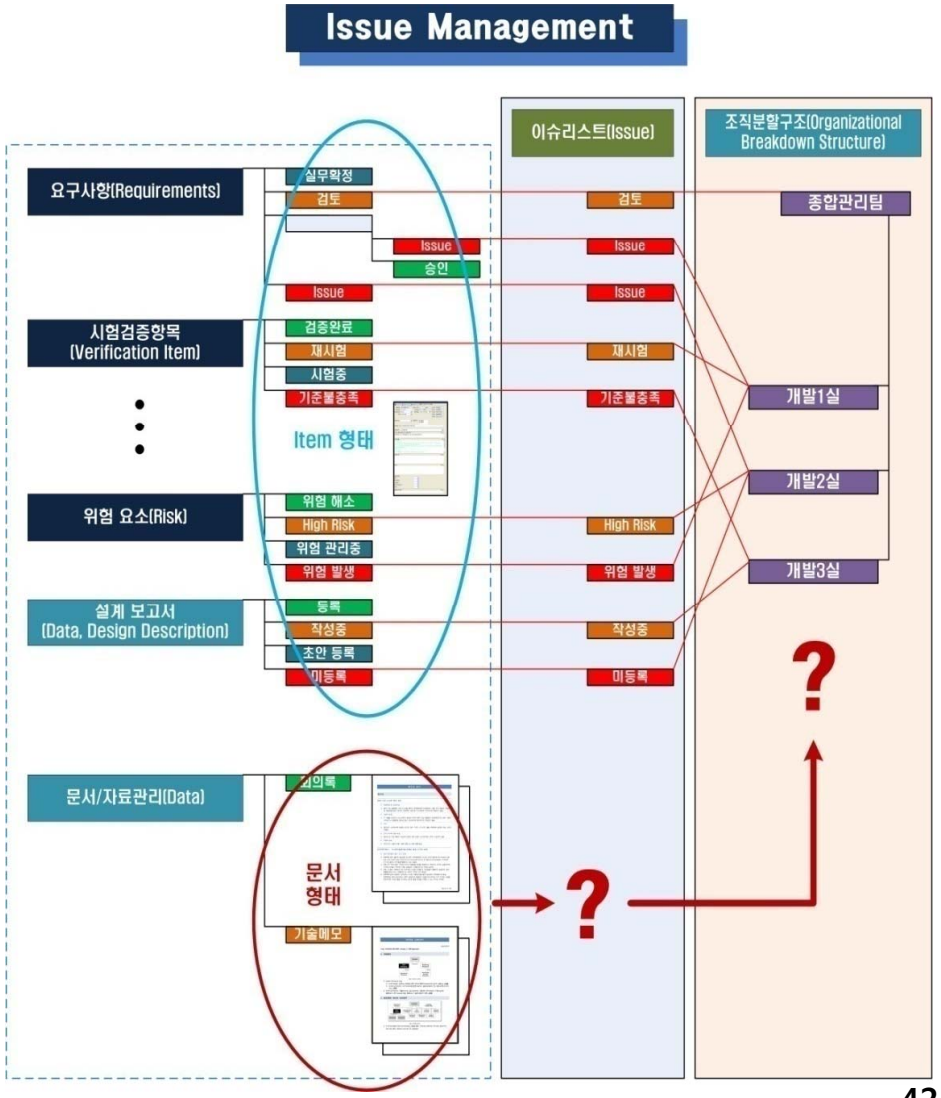
◆ AS-IS

① 매일 매일 시스템 개발 프로젝트와 관련한 다양한 형태의 회의, 토의, 의사결정이 이루어지고 있지만, 이에 대한 내용은 참석자 및 관련자(회의록 등을 메일로 전달받은 사람)들에게만 전달되고 있음.

② 시스템 또는 서브시스템의 규격, 설계사항과 관련하여 토의되는 내용이 대부분이지만, 이와 관련한 분류(시스템 구성에 따른) 및 변경사항 반영은 현재는 모두 데이터베이스 관리자가 직접 수행하여야 하는 상태임.

③ 심각한 overhead 발생.

④ 변경사항 반영이 실시간으로 이루어지지 못하고 있으며, 누락사항 발생도 충분히 가능한 상태임.



2. TO-BE

◆ 제안 사항

① 조직 구성(OBS), 업무분할구조(WBS), 시스템 분해구조(SBS), 요구사항 전개구조(RBS)를 Ontology 모델 등으로 구축. → Reference Ontology Model

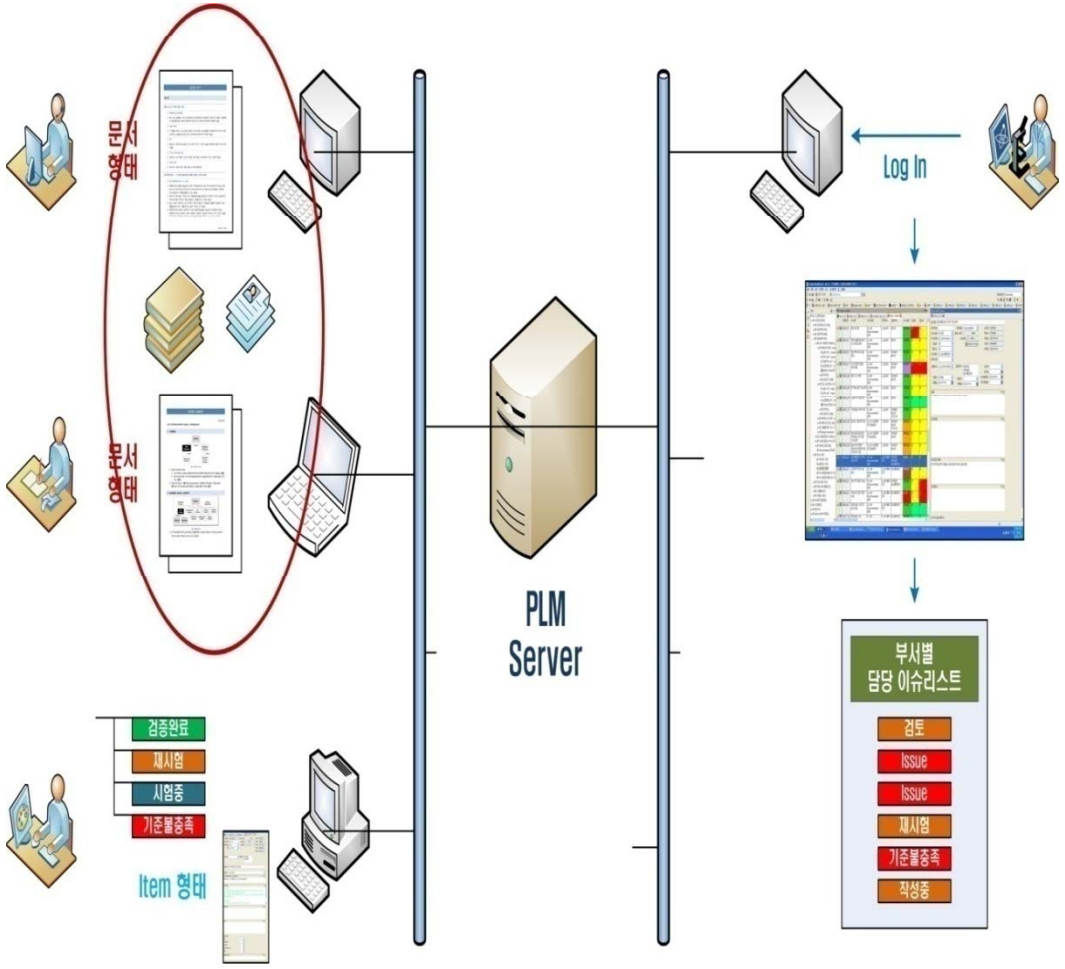
② 회의 및 의사결정 데이터 입력 : 회의록(document) 또는 SE 도구를 이용한 well-structured data.

③ 입력 데이터를 Reference Ontology model을 이용, 각 조직, 업무, 시스템에 대하여 Semantic하게 할당, 재정리.

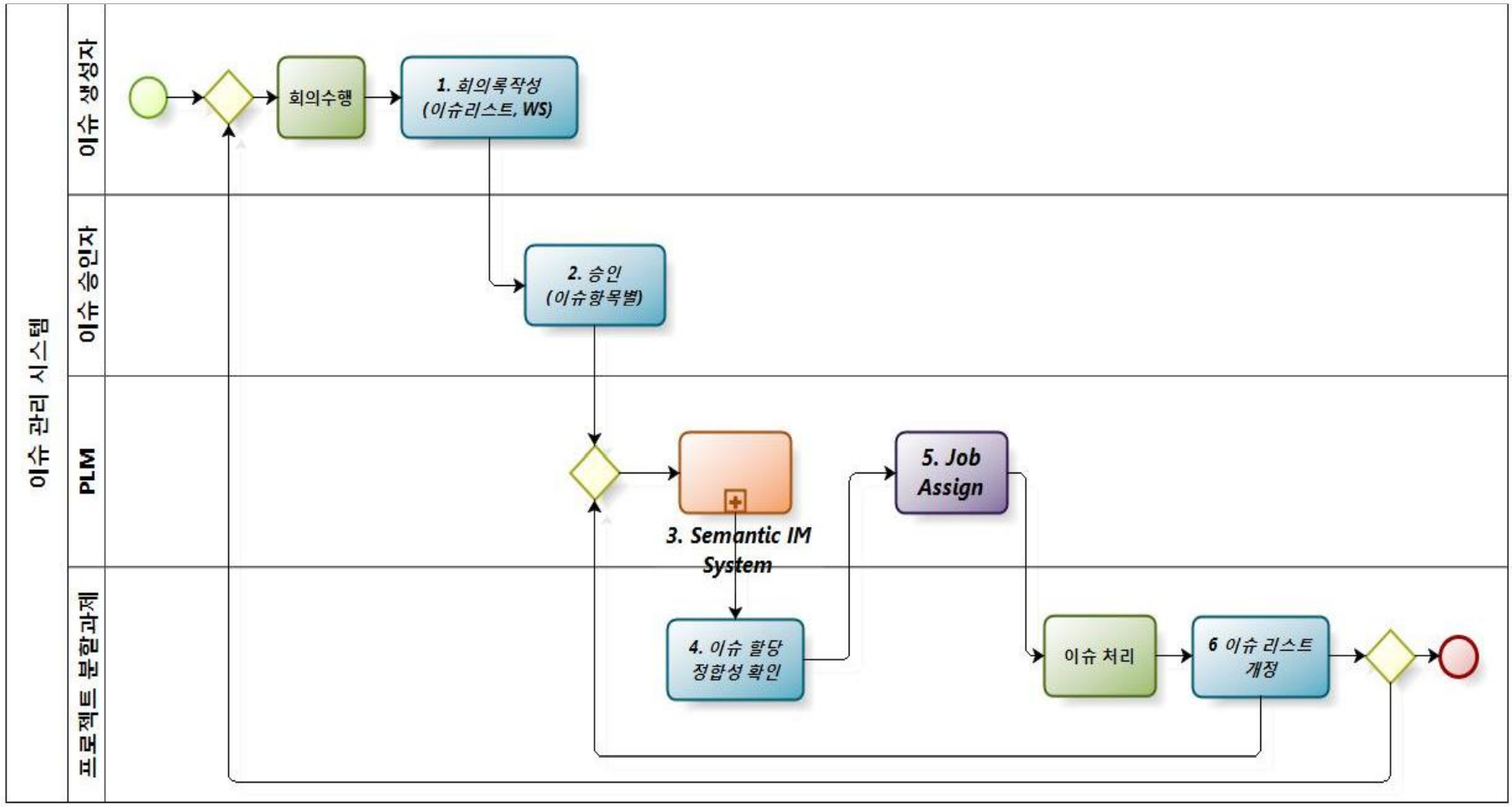
◆ TO-BE

① 엔지니어가 출근하여 PLM시스템에 로그인하게 되면, 항상 자기자신의 업무(또는 조직, 담당 시스템, 요구사항)와 관련하여 토의되거나 의사 결정된 내용을 실시간으로 확인 가능.

② 자기자신이 현재 해결하여야 할 이슈(회의의 경우 Action item)을 실시간으로 확인 가능.



3. TO-BE Work Flow



Contents

Conclusion

- 제품 개발 관련 Engineering 문서 폭증 → 효율적 검색 방안 필요
- PLM이 보편화/안정화/고도화 단계 → 탐색/검색 기능 부각
- Keyword 검색 → 선택의 폭 너무 넓음
- Semantic Processing 기반 Engineering 문서 관리/검색
  - 정보 검색의 효율성 (↑)
  - 정보 재 활용성 (↑)
  - 정보의 통합성 (↑)

Question and Answer

THANK YOU ☺