온라인 공개 국방기술정보를 이용한 콘텐츠 분석

Analysis of defense technology based on online public open contents

정 휘 웅*, 김 경 선*, 최 중 환**
다이퀘스트*, 국방기술품질원**

Hwi woong Jeong*, Kyungsun Kim*,
Joonghwan Choi**
Diquest*,
Defense agency for Technology and Quality

요약

국방기술 콘텐츠는 최신성과 은닉성을 전제로 한다. 국방기술 콘텐츠는 그 전문성과 다양한 약어의 활용, 그리고 현재 국가별로 진행되고 있는 다양한 정보들을 바탕으로 정보를 분석하게 된다. 그러나 이러한 국방 기술 정보들인 대부분 축적만 될 뿐 제대로 추출되거나 분석이 되지 못하고 있으며, 실제 현업에서 활용하기에는 부족한 실정이다. 본 고에서는 이러한 문제점을 바탕으로 현재 온라인상에 존재하고 있는 국방 관련 정보를 분석하였다.

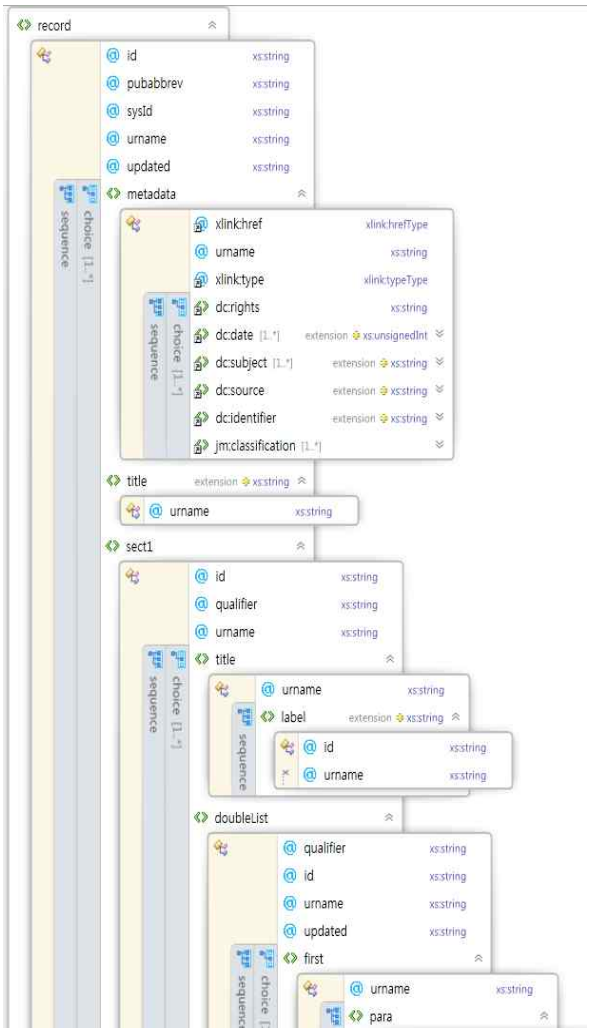
Ⅰ. 서론

국방 콘텐츠는 온라인 상에서 입수하기가 어려울 뿐만 아니라 실제 환경에서 사용되는 경우도 많지 않다. 아울러 국방 콘텐츠를 이용한 정보 분석은 많은 인력과 비용이 들어갈 뿐만 아니라 보안성을 생명으로 하기 때문에, 이에 대한 분석은 국가 안보에 직결되는 중요한 이슈라 할 수 있다. 본 고에서는 2장에서 공개형 국방 콘텐츠 구조에 대하여 설명하며, 3장에서는 콘텐츠를 이용한 사양 정보 추출 실험을, 4장에서는 향후 연구 방향에 대하여 제시하겠다.

Ⅱ. 공개형 국방 콘텐츠의 구조

온라인상에 존재하는 정보들은 최근 많은 형태에 있어서 XML 기반 형태로 제공되고 있다. XML 기반 말뭉치의 구조는 LISA에서 제시한 구조와 함께, TEI가 가장 대표적이며[1], 이의 메타 정보는 Dublin Core를 핵심으로 꼽을 수 있다. 공개형 국방 콘텐츠의 경우에도 각 콘텐츠의 항목에 있어서 Dublin Core[2] 를 이용한 메타 정보를 표식하고 있으며, 각 문단 및 문장 정보에 대하여 id 정보를 지정함으로써 말뭉치로 구성하였을 때 정보의 추출이 용이하도록 지원하고 있다. 각 데이터의 구조는 xlink의 구조와 함께 각 문단의 정보를 표식하고 있는데, 모든 정보의 구조를 paragraph 단위로 분절하고 있어서 정보의 추출이 용이하게 구성되어 있다.

각 용어의 요소들은 사양 및 각종 정보들 (진수 시기, 초기 비행 시기 등)의 정보를 개별적인 태깅 구성을 통하여 제공하고 있다.



▶▶ 그림 1. 공개형 국방 콘텐츠의 DTD 구조 예시

Ⅲ. 공개형 국방 콘텐츠에서 사양 정보 추출

본 고에서는 XML문서를 바탕으로 국방 관련 사양 정보를 파싱하여 추출하는 실험을 수행하였다. 수집된 3,338건의 문서 중에서 총 1558개의 사양 정보를 추출하였으며, 이에 따른 단위 정보를 함께 추출할 수 있었다.

별도의 추출 모듈을 이용하여 용어 내에서 특정한 메타 정보가 존재하는 정보를 바탕으로 각 항목을 추출하였으며, 오류로 입력된 사양 정보들은 통계에서 제외했다. 가령 용어에 있어서 띄어쓰기가 되지 않아서 인식이 달리 된 경우는 항목 자체를 제거하여 통계의 정확도를 높였다.

표 1. 사양 정보 추출

	kg	kilometre	kilometres	km	m	합
Height	0	0	0	0	167	167
Hovering ceiling	0	0	0	0	20	20
Landing run	0	0	0	0	88	88
Length	0	0	0	0	178	178
Range	0	349	1	157		507
Rotor diameter	0	0	0	0	30	30
Service ceiling	0	0	0	0	104	104
Take-off run	0	0	0	0	105	105
Take-off weight	194	0	0	0		194
Wingspan			0	0	166	166
총합계	194	349	1	157	858	1559

각 추출된 정보들은 세부적인 사양 정보를 포함하고 있으며, 이는 문서별로 저장된 기종 혹은 기술의 정보와 연계되어서 비교 정보로 활용 될 수 있다. 아울러 각 사양정보가 국내 국방 콘텐츠 환경과 연계되어 분석될 수 있도록 함으로써 이의 분석이 보다 용이하게 이루어질 수 있을 것이다. 링크드 데이터 및 외부 연계 정보를 이용하여 콘텐츠간의 연계성을 높일 경우 이의 효율성 역시 매우 높아질 것이다.

Ⅳ. 결론 및 향후 연구

본 고에서는 초기적인 형태의 정보 추출 및 XML 문석 작업을 수행하였다. 미래 전산 환경에서는 추출된 정보를 바탕으로 얼마나 시맨틱 혹은 의미론적 관점에서 적절하게 정보를 나열하는 것이 매우 중요한 이슈로써, 본 고에서 제시한 정보들은 향후 다양한 형태로 정보를 분석하는데 도움을 줄 수 있다. 가령 시맨틱 검색에서 제시된 각 사양 정보들을 바탕으로 정보 구조를 구성한다면 시맨틱 검색 등에서 정보를 분석하는데 그 효율성을 획기적으로 개선시킬 수 있을 것이다. 아울러 이를 통하여 전문 용어를 추출하고 태깅함으로써 정보를 단순히 기록으로만 남기는 것이 아니라, 정확한 정보를 추출할 수 있는 기반이 될 수 있을 것으로 기대된다.

알림

- 본 논문은 국방기술품질원의 보안성 검토결과 적격 판정을 받았습니다.
- 본 연구는 문화관광부 2010년도 콘텐츠산업기술지원사업의 지원으로 이루어졌습니다. (과제번호: 2-10-7602-001-10752-10-001, 디지털 문화콘텐츠 융복합 서비스를 위한 시맨틱웹 매쉬업 플랫폼 기술)

▌참 고 문 헌▐

[1] http://www.oasis-open.org
[2] http://www.dublincore.org