Regular paper | Journal of the KIECS. pp. 571-576, vol. 19, no. 3, Jun. 30. 2024, t. 125, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2024.19.3.571

온톨로지 이질성 문제를 해결하기 위한 온톨로지 매칭 방법

단홍조우* · 이용주**

Ontology Matching Method for Solving Ontology Heterogeneity Issue

Hongzhou Duan* · Yongju Lee**

요 약

온톨로지는 도메인 전문가에 의해 만들어지지만, 동일한 내용이라도 전문가마다 도메인 지식에 대한 이해가 다르기 때문에 상이하게 표현될 수 있다. 아직 온톨로지 표준화가 부족하기 때문에 동일한 도메인 내에 여러 개의 온톨로지가 존재할 수 있으며, 이로 인해 온톨로지 이질성이라는 현상이 발생한다. 따라서 우리는 온톨로지 이질성 문제를 해결하기 위해 SCBOW(: Siames Continuois Bag Of Words)와 BERT(: BiDirectional Encoder Representations from Transformers) 모델을 결합한 새로운 온톨로지 매칭 방법을 제안한다. 온톨로지를 그래프로 표현하며, 온톨로지 매칭 문제에서 발생할 수 있는 일대다 문제를 해결하기 위해 SimRank 알고리즘을 사용한다. 실험 결과 우리의 접근 방식이 전통적인 매칭 알고리즘보다 약 8%의 성능 향상을 보였다. 제안 방법은 온톨로지 매칭에 사용되는 정렬 기술을 향상하고 개선할 수 있다.

ABSTRACT

Ontologies are created by domain experts, but the same content may be expressed differently by each expert due to different understandings of domain knowledge. Since the ontology standardization is still lacking, multiple ontologies can be exist within the same domain, resulting in a phenomenon called the ontology heterogeneity. Therefore, we propose a novel ontology matching method that combines SCBOW(: Siames Continuois Bag Of Words) and BERT(: Bidirectional Encoder Representations from Transformers) models to solve the ontology heterogeneity issue. Ontologies are expressed as a graph and the SimRank algorithm is used to solve the one-to-many problem that can occur in ontology matching problems. Experimental results showed that our approach improves performance by about 8% over traditional matching algorithm. Proposed method can enhance and refine the alignment technology used in ontology matching.

키워드

Ontology Matching Method, Heterogeneity Problem, SCBOW, BERT, SimRank, Knowledge Base
온톨로지 매칭 방법, 이질성 문제, SCBOW, BERT, SimRank, 지식 베이스

* 경북대학교 IT대학 컴퓨터학부
(caixiuming1984@163.com)
** 교신저자 : 경북대학교 IT대학 컴퓨터학부
• 접 수 일 : 2024. 04. 13
• 수정완료일 : 2024. 05. 13
• 게재확정일 : 2024. 06. 12
• Received : Apr. 13, 2024, Revised : May. 13, 2024, Accepted : Jun. 12, 2024
• Corresponding Author : Yong-Ju Lee
Dept. Computer Science and Engineering, Kyungpook National University
Email : yongju@knu.ac.kr

Ⅰ. 서 론

최근 WordNet, Freebase, Wikidata와 같은 대규모 지식 베이스에 지식 그래프 임베딩 및 그래프 합성곱 신경망을 활용한 연구가 활발히 진행되고 있으나, 지식 베이스에 대한 엔티티 매칭 연구는 상대적으로 거의 연구가 미비한 상태이다[1]. 대규모 지식 베이스는 아직 해결되어야 할 이슈들이 많지만, 그들 중 하나는 서로 다른 온톨로지를 사용하는 어휘 이질성(heterogeneity) 문제이다[2]. 온톨로지는 도메인 전문가들에 의해 만들어지지만 동일한 내용일지라도 전문가마다 도메인 지식에 대한 이해가 다를 수 있기 때문에 온톨로지가 상이하게 표현될 수 있다. 현재까지 온톨로지에 대한 표준화도 미비한 상태이기 때문에 동일한 도메인 내에 여러 개의 온톨로지가 존재할 수 있고, 이로 인해 온톨로지 이질성이라는 현상이 발생한다.

그림 1은 하나의 예를 통해 온톨로지 이질성 문제를 보여준다. 여기서 온톨로지 매칭 결과 다음과 같은 두 쌍은 동의어로 인식된다.

O1. Vehicle ↔ O2. Means of Transport

O1. Car ↔ O2. Automobile

그러나 O1의 'Benz C350'과 O2의 'Mercedes Benz C Class 350'은 매칭이 되지 않는다. 이는 기존의 온톨로지 매칭 기법들이 동의어 집합을 사용하여 매칭을 수행하는데, 'Benz C350'과 'Mercedes Benz C Class 350'은 같은 동의어 집합에 속하지 않기 때문에 매칭이 되지 않는다.

본 연구에서는 이러한 온톨로지 이질성 문제를 해결하기 위해 SCBOW(: Siamese Continuous Bag Of Words)와 BERT(: Bidirection Encoder Representations from Transformers) 모델을 결합한 새로운 온톨로지 매칭 방법을 제안한다. 또한, 우리는 온톨로지를 그래프로 표현하고, 온톨로지의 매칭 문제에서 발생할 수 있는 일대다(one-to-many) 문제를 해결하기 위해 SimRank 알고리즘을 사용하였다. 따라서, 본 연구의 목적은 특정 전문 영역 내에서 서로 다른 온톨로지 간의 온톨로지 이질성 문제를 효과적으로 해결하는 새로운 온톨로지 정렬 모델을 제안하는 것이라 할 수 있다.

Ⅱ. 관련 연구

지식 그래프 및 온톨로지는 지식 표현 방법으로써 RDF(: Resource Description Framework), RDF-S(RDF Schema), OWL(: Web Ontology Language) 등의 언어를 이용해 기술한다. RDF는 단점으로써 어휘를 정의하는 능력이 많이 부족하기 때문에 클래스 간 상하관계를 정의할 수 있는 RDF-S를 같이 사용한다. 그러나 자원 간의 동일성과 이질성, 그리고 집합관계나 제한조건은 RDF나 RDF-S로 표현할 수 없다. 이에 W3C는 좀 더 표현력이 풍부하고 다양한 개념을 표현할 수 있는 OWL을 표준 온톨로지 언어로 제안하고 있다.

단어 임베딩(word embedding)은 유사한 단어들이 인접한 거리에 위치하도록 각 단어에 해당하는 벡터 값을 찾는 것이다. 대표적인 기법으로 Word2Vec[3], GloVe[4], FastText[5] 등이 있으나 이들은 단어가 아닌 문장 표현을 하기 위해서는 최적화되어 있지 못하다. 문장 임베딩(sentence embedding)을 위해서는 단어 임베딩을 평균화하는 것이 효율적인 방법이 될 수 있다. 따라서 SCBOW[6]는 평균화를 위해 단어 임베딩을 직접 훈련하여 문장 임베딩 문제를 해결하였다. BERT[7]는 Transformer의 encoder만을 사용하여 문장을 양방향으로 학습시킨 언어 모델을 말한다. GPT는 한 방향으로만 문장을 학습하고 ELMo는 왼쪽-오른쪽과 오른쪽-왼쪽을 고려하는 LSTM(: Long Short-Term Memory) 두 개를 사용하여 학습했지만, BERT는 MLM(: Masked Language Modeling)과 NSP(: Next Sentence Prediction) 전략으로 하나의 네트워크가 양쪽 방향 다 고려하는 양방향 학습을 통해 문맥 정보를 잘 반영할 수 있다.

온톨로지에서 엔터티 간의 구조적 유사성을 계산하려면 온톨로지에 존재하는 "IS-A"와 "SubclassOf" 같은 관계를 고려하는 것이 필요하다. SimRank 알고리즘[8]은 두 온톨로지에서 추출된 그래프 구조 내 노드 간의 유사성을 발견한다. SimRank의 기본 원리는 두 노드의 이웃 노드가 어느 정도 유사성을 나타내는 경우 이는 두 노드 자체가 일정 수준의 유사성을 가지고 있음을 의미하는 것이다.

III. 온톨로지 매칭 방법

그림 2는 온톨로지 정렬 모델에 대한 프로세스를 개략적으로 표현한 그림인데, 이 모델에서는 텍스트 유사도 계산(text similarity calculation)과 구조적 유사도 계산(structural similarity calculation)의 2단계로 구성되어 있으며 이들 결과를 통합하여 최종 일치 결과를 얻게 된다.

3.1 1단계: 텍스트 유사도 계산

1단계에서는 고품질의 문장 임베딩을 효율적으로 산출하기 위하여 SCBOW 신경망 모델을 사용한다. 여기서 고품질이란 의미상 가까운 문장들에 대한 임베딩은 서로 유사하게 나타나게 하고, 의미상 다른 문장들은 임베딩이 서로 다르게 나타나게 하는 것이다. 문장에 대한 임베딩을 계산하는데 하나의 효율적이고 성공적인 방법은 문장을 구성하는 단어의 임베딩을 평균화하는 것이다. 최근 이러한 작업을 위해 Word2Vec과 GloVe과 같은 연구에서는 사전에 훈련된 단어 임베딩을 사용했지만 이는 문장 표현에 최적화되지는 못했다. 따라서 우리는 단어 임베딩을 평균하여 문장 임베딩을 계산하지만 평균화를 위해 단어 임베딩을 직접 최적화시킨다. 즉, 본 연구에서는 훈련 데이터에서 서로 옆에 나타나는 문장을 예측할 수 있도록 지도 훈련 기준을 구성한다. 구체적으로 설명하면, 한 쌍의 문장(x_i, x_j)에 대해 훈련 데이터에서 문장이 서로 인접할 가능성을 반영하는 확률 $p(x_i, x_j)$를 식(1)과 같이 정의한다.

$$p(x_i,x_j) = \begin{cases} \frac{1}{|X^+|} & \text{if } x_i \in X^+ \\ 0 & \text{if } x_j \in X^- \end{cases} \quad \cdots (1)$$

여기서 분모의 합은 이론적으로 가능한 모든 문장 X 걸쳐 있어야 하지만 실제로는 불가능하다. 따라서 집합 X를 훈련 데이터에서 문장 x_i 옆에 나타나는 문장의 집합 X^+와 훈련 데이터에서 문장 x_i 옆에 관찰되지 않는 무작위로 선택된 n개의 문장 집합 X^-의 합집합으로 대체한다. 또한, 이 단계에서 온톨로지의 텍스트 요소에 의미론적 표현을 할당하기 위해 BERT 방법을 활용한다. BERT는 트랜스포머(transformer)를 기반으로 구성되며 레이블링이 되지 않은 데이터로 사전 학습된 모델이다. BERT의 입력 표현(input representation)은 Token Embeddings, Segment Embeddings, Position Embeddings의 세 가지 임베딩 값의 합으로 구성된다. 학습 과정은 2단계로 진행되며 사전학습(pre-training)과 미세조정(fine-tuning) 과정으로 진행된다.

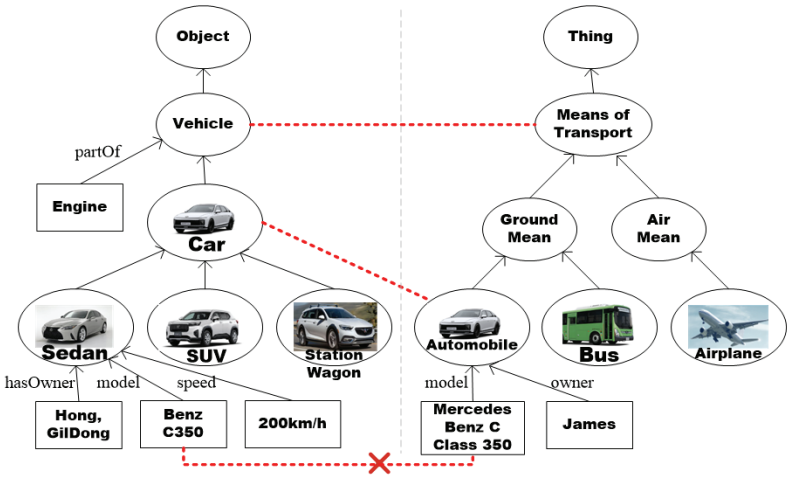

그림 1. 온톨로지 이질성 문제의 예
Fig. 1 An example of ontology heterogeneity problem

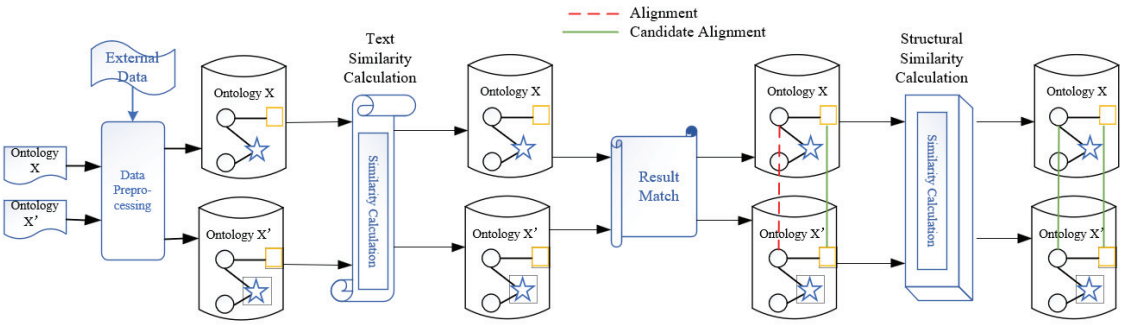

그림 2. 온톨로지 정렬 모델을 위한 프로세스
Fig. 2 Process for ontology alignment model

$$S(x,y)=\begin{cases} sim(x,y) & \text{if}(x,y)\in A \\ \dfrac{C}{|N(x)||N(y)|}\displaystyle\sum_{i=1}^{|N(x)|}\sum_{j=1}^{|N(y)|}S(N_i(x),N_j(y)) & |N(x)||N(y)|\neq 0 \\ 0 & |N(x)||N(y)|=0 \end{cases} \quad \cdots (2)$$

3.2 2단계: 구조적 유사도 계산

2단계로 SimRank 알고리즘을 활용하여 개체 간의 구조적 유사성을 계산한다. 온톨로지는 먼저 구조 정보를 사용하여 그래프 구조로 변환한다. 그런 다음 이전 단계에서 얻은 텍스트 유사성 정보를 사용하여 두 그래프 구조 간의 연결을 설정하고 이를 하나의 그래프로 결합한 후, SimRank 알고리즘을 두 온톨로지의 엔터티 간 구조적 유사성을 측정하는 데 적용한다. 식(2)는 온톨로지 매칭 과정에서 두 노드 간의 유사도가 어떻게 계산되는지를 나타낸다.

여기서, ① 두 노드가 초기 앵커 일치인 경우($(x,y)\in A$), 워드 임베딩 기법을 사용하여 코사인 유사성을 계산하여 유사성을 결정하는데($sim(x,y)$), 이는 노드 간의 텍스트 유사성을 캡처한다. ② 그러나 두 노드가 초기 앵커 일치가 아닌 경우($|N(x)||N(y)|\neq 0$) 유사성은 다르게 계산된다. 이웃 노드 간의 유사성을 합산한 다음() 결과 유사성을 평균하여 계산되는데$\frac{C\sum\sum S(N(x),N(y))}{|N(x)||N(y)|}$, 이 접근법은 이웃 노드의 유사성을 고려하여 구조적 정보를 고려한다. 여기서, S는 similarity, C는 [0, 1] 사이의 상수값이다. ③ 노드 중 하나가 격리된 경우(즉, 연결된 이웃 노드가 없음: $|N(x)||N(y)|\neq 0$), 두 노드 간의 구조적 유사성은 매우 낮아 유사성 값이 0이 되는데, 이는 격리된 노드와 그래프의 다른 노드 간의 구조적 유사성이 부족함을 반영한다.

IV. 실험 분석

본 연구에서는 SCBOW 모델을 이용하여 초기 단어 벡터를 개선한 후, 의미론적 표현을 할당하기 위해 BERT 방법을 적용하였고, SimRank 알고리즘을 이용하여 구조적 유사성을 보정하여 최종 매칭 결과를 얻었다. 성능평가 방법은 정보검색 분야에서 가장 보편적으로 활용되고 있는 재현율(recall) R과 정밀도(precision) P, 그리고 F-측정값(F-measure) F를 이용한다[9]. 재현율은 온톨로지 매칭 시 적합한 매칭이 얼마나 되는지를 나타내고, 정밀도는 매칭된 결과 중에서 요구한 매칭들이 얼마나 되는지를 나타낸다. 재현율과 정밀도는 모두 높을수록 성능이 좋다고 할 수 있으나 이들은 서로 반비례적인 관계가 있어 한쪽을 높이면 다른 한쪽이 내려가는 것이 보통이다. 따라서 F-측정값은 재현율과 정밀도를 대체하는 하나의 척도로써 재현율과 정밀도의 가중치 조화 평균으로 계산

하는 방식이다[10]. 재현율 R, 정밀도 P, 그리고 F-측정값 F는 식 (3), (4), (5)와 다음과 같이 계산된다.

$$R = \frac{C}{C+M} \quad \cdots (3)$$

$$P = \frac{C}{C+IC} \quad \cdots (4)$$

$$F = \frac{2P \times R}{P+R} \quad \cdots (5)$$

여기서, 매칭된 엔티티들 중 적합한 엔티티 수는 C(correct)이고, 부적합한 수는 IC(incorrect)이며, 매칭되지 않은 엔티티들 중 적합 엔티티 수는 M(missing)으로 표현된다. 적합 엔티티는 실험 데이터를 구축하는 연구가들에 의해 계산되는데 그 수는 C+M과 같이 된다.

실험 데이터로는 FMA(: Foundational Model of Anatomy)[1], SNOMED(: Systematized NOmenclature of MEDicine)[2], 그리고 NCIT(: National Cancer Institute Thesaurus)[3] 데이터셋을 사용한다. FMA는 해부학적 정보를 위해 분산 프레임워크에 통합된 정보 리소스이고, SNOMED는 SNOMED 국제 조직에서 관리하는 컴퓨터 처리 가능한 의학 용어 모음이다. 그리고 NCIT는 미국 국립 암 연구소에서 개발한 공개적으로 사용 가능한 용어집이다. 두 개를 결합한 FMA-NCIT와 FMA-SNOMED 데이터셋에서 제공하는 일치 답변 분석에서 일치 항목의 상당 부분이 일대다 유형인 것으로 나타났다. 그러나 기존 매칭 알고리즘은 일대일 매칭만 처리할 수 있어서 이런 알고리즘이 달성할 수 있는 최대 재현율 값은 제한된다. 이러한 제한으로 인해 일대다 매칭을 처리하기 위해 임계값 집합을 사용하는 대체 일치 방법을 탐색하게 되었다.

우리는 FMA-NCIT와 FMA-SNOMED 데이터셋을 사용하여 기존 전통적인 매칭 알고리즘은 물론이고 여러 다양한 알고리즘에 대한 전반적인 성능을 측정하였는데, 표 1은 이러한 알고리즘들에 대한 성능 측정 결과를 보이고 있다. 이 표로부터 예측할 수 있듯이 우리의 접근 방식(SCBOW+BERT+SimRank)은 FMA-SNOMED의 P 성능만 제외하고 다른 모든 방법보다 우수한 것을 알 수 있다. 구체적으로 살펴보면, 전통적인 매칭 알고리즘보다 R, P, 그리고 F 결과가 FMA-NCIT에서 7.1%, 10.2%, 9.5%, 그리고 FMA-SNOMED에서 30.2%, 2.5%, 6%의 성능 향상을 보이고 있다.

VI. 결 론

본 논문에서는 온톨로지 이질성 문제를 해결하기 위해 SCBOW와 BERT 모델을 결합한 온톨로지 매칭 방법을 제안하였다. 또한 온톨로지 매칭 문제에서 발생할 수 있는 일대다 문제를 해결하기 위해 SimRank 알고리즘을 사용하였다. 제안된 온톨로지 정렬 모델에서는 텍스트 유사도 계산과 구조적 유사도 계산의 2단계로 구성되고 이들 결과를 통합하여 최종 일치 결과를 산출한다. 성능평가 방법은 정보검색 분야에서 가장 널리 활용되고 있는 재현율, 정밀도, F-측정값을 사용하였는데, 본 논문에서 제안한 방법이 기존의 전통적인 매칭 알고리즘은 물론이고 SCBOW와 SCBOW+SimRank 방법들보다 성능이 우수함을 보였다.

표 1. 성능 측정 결과
Table 1. Performance measurement results

	FMA-NCIT dataset			FMA-SNOMED dataset		
	R	P	F	R	P	F
Tranditional Matching Algorithm	0.849	0.752	0.789	0.640	0.654	0.731
SCBOW	0.852	0.779	0.802	0.645	0.668	0.689
SCBOW+SimRank	0.891	0.753	0.818	0.715	0.679	0.730
SCBOW+BERT+SimRank	0.914	0.837	0.872	0.917	0.671	0.778

1) https://si.washington.edu/projects/fma
2) https://www.nlm.nih.gov/healthit/snomedct/snomed_overview
3) https://www.cancer.gov/research/resources/resource/197

감사의 글

본 논문은 교육부 및 한국연구재단의 4단계 BK21 사업(경북대학교 컴퓨터학부 지능융합 소프트웨어 교육연구단)으로 지원된 연구임(4120240214871). 본 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1A1B02008553).

References

[1] Y. Lee and Y. Sun, “Entity Matching Method of Knowledge Graphs using Graph Convolutional Network and Embedding Techniques,” *Journal of Korean Institute of Information Technology*, vol. 21, no. 6, June 2023, pp. 09-19.

[2] H. Duan and Y. Lee, “Entity Matching Method Using Semantic Similarity and Graph Convolutional Network Techniques,” *Journal of the Korean Institute of Electronic Communication Sciences*, vol. 17, no. 5, Oct. 2022, pp. 801-808.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *ICLR(International Conference on Learning Representations)*, Arizona, USA, Sept. 2013.

[4] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532-1543.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching Word Vectors with Subword Information,” *Transactions of the Association for Computational Linguistics*, vol. 5, June 2016, pp. 135-146.

[6] T. Kenter, A. Borisov, and M. Rijke, "Siamese CBOW: Optimizing Word Embeddings for Sentence Representations,” *54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, Aug. 2016, pp. 941-951.

[7] J. Devlin, M Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *2019 Conference of the North American Chapter of the Association for Computational Linguistics*, vol. 1, June 2019, pp. 4171-4186.

[8] G. Jeh and J. Widom, "SimRank: A Measure of Structural-Context Similarity,” *8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, July 2002, pp. 538-543.

[9] D. M W Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, Jan. 2011, pp. 37-63.

[10] J. Lee and O. Kwon “Performance Assessment of Machine Learning and Deep Learning in Regional Name Identification and Classification in Scientific Documents,” *Journal of the Korean Institute of Electronic Communication Sciences*, vol. 19, no. 2, Apr. 2024, pp. 389-396.

저자 소개

단홍조우(Hongzhou Duan)

2009년 하얼빈 이공대학교 소프트웨어 공학과 졸업(공학사)
2019년 경북대학교 대학원 컴퓨터학과 졸업(공학석사)
2021년 ~현재 경북대학교 대학원 컴퓨터학과 (박사과정)
※ 관심분야 : 시맨틱 웹, 지식 그래프 임베딩, 딥러닝, 빅 데이터

이용주(Yong-Ju Lee)

1985년 한국과학기술원 정보검색 전공(공학석사)
1997년 한국과학기술원 컴퓨터공학전공(공학박사)
1998년 8월 ~ 현재 경북대학교 IT대학 컴퓨터학부 교수
※ 관심분야 : 링크드 데이터, 시맨틱 웹, 빅데이터, 웹 사이언스