

시맨틱 웹을 위한 요소기술 - 인공지능

백성복 인공지능기반기술(PG1005) 부의장, KT 융합기술원 책임연구원

1. 머리말

웹3.0을 실현하는 두 가지 핵심 기술로 시맨틱 웹과 인공지능을 들 수 있다. 시맨틱 웹과 웹 3.0을 구분하지 않고 사용하는 경우도 있지만, 보다 정확히 말하자면 시맨틱 웹이 먼저 제시되고 웹3.0이 뒤이어 제안되면서 먼저 나온 시맨틱 웹 개념을 포괄하게 된 것으로 볼 수 있다.

팀 버너스 리는 2001년 논문에서 컴퓨터가 인간 언어의 시맨틱을 처리할 만한 적절한 방법을 가지고 있지 못하다는 점을 지적하면서 '시맨틱 웹'의 필요성을 제기했다. 즉, 인간 언어를 구성하는 단어나 문장이 가지고 있는 실제 문맥을 표현하거나 전달할 방법이 컴퓨터 안에 구현되어 있지 않다는 지적이다. 시맨틱 웹을 통해 기존의 웹페이지들이 담고 있는 의미들을 구조화함으로써 컴퓨터가 사용자들을 위해 더 복잡한 고도의 작업을 처리해 줄 수 있다고 주장한 것이다.

버너스 리의 이러한 주장은 그가 1990년대 월드 와이드 웹 기술을 처음 발표하면서 핵심 개념으로 제시했던 탈중앙화와 상향식(Bottom-up) 설계라는 웹의 근간 개념을 재차 강조한 것으로도 볼 수 있다. 탈중앙화 개념은 웹에 정보를 게재할 때 중앙에 위치한 상부 기관으로부터 허가를 받을 필요가 없도록 정보 체계를 구축하는 방식을 말한다. 중앙화된 제어 노드가 없기 때문에 위험의 집중화 현상을 피할 수 있고 검열과 통제로부터 자유롭다는 장점이 있다. 상향식 설계 개념은 소수 전문가 집단에 의해 작성되는 코드가 인터넷의 전체 정보를 제어하는 대신에 모든 사람들이 공평하게 참가하고 원하는 바를 시도해 볼 수 있는 환경을 구축하자는 개념이다.

시맨틱 웹이 제안될 당시, 인간의 언어를 컴퓨터가 이해하도록 하자는 핵심 개념이 인공지능 진영의 기술 발전 방향과 일치했기 때문에 이 두가지 기술이 함께 묶여서 거론되곤 했다. 초기에는 시맨틱 웹이나 인공지능 기술 모두가 큰 성공을 이루지 못했기 때문에 두 기술 모두에 대해 업계의 반응은 냉담했다. 하지만 시맨틱 웹의 방향성 자체는 의미가 있었고 지속시킬 만한 가치가 있었기 때문에 웹3.0이라는 좀더 포괄적인 개념으로 이어져 왔다.

웹3.0은 버너스 리가 제안했던 시맨틱 웹이라는 본래의 개념보다 훨씬 확장되고 진보된 개념으로 볼 수 있으나, 현재로서는 용어 자체가 어떤 경향에 대한 기술에 가깝고 명확히 실체화된 기술이나 대상을 정의하는 단어라고 보기는 어렵다. 따라서 웹3.0의 구체적 양상을 논하기 위해서는 그 핵심 요소인 시맨틱 웹과 인공지능에 대해 살펴보는 것이 더 바람직하다.

2. 시맨틱 웹 개요

시맨틱 웹은 특정 문서나 애플리케이션을 연결하기보다는 그 내용과 의미 자체를 연결하는 방법을 제시한다. 어떤 구체적인 정보를 참조할 때 그 정보가 포함된 문서나 애플리케이션을 가리키는 것이 아니라 정보 자체를 가리키는 방식이다. 해당 정보가 변경 또는 갱신되더라도 별도로 작업을 추가할 필요가 없다는 장점이 있다.

보다 근본적으로 시맨틱은 '의미' 또는 '이해'의 문제를 해결하고자 하는 기술이며, 구체적인 세부 기술로는 자연어 처리(Natural Language Processing)나 시맨틱 검색 등이 포함된다. 시맨틱 웹 기술이 다른 데이터 관련 기술(대표적으로 관계형 데이터베이스)과 다른 점은 데이터의 구조가 아니라 데이터의 의미에 집중한다는 것이다. 이러한 근본적인 차이로 인해 정보를 저장하고 질의하고 표현하는 방식이 완전히 달라진다.

대량의 데이터를 다수의 소스로부터 참조해야 하는 애플리케이션을 개발해야 한다면 시맨틱 웹 기술이 매우 유용하다. 반면, 고수준으로 구조화된 트랜잭션 유형의 데이터를 다량으로 취급하는 정보 저장형 애플리케이션의 경우에는 시맨틱 웹 기술이 그다지 강점을 발휘하지 못할 것이다.

시맨틱 웹을 구성하는 대표 표준으로 아래 세가지 표준을 들 수 있다.

- (1) RDF(Resource Description Framework): 시맨틱 웹을 위한 데이터 모델링 언어에 대한 표준이다. 모든 시맨틱 웹 정보는 RDF를 준용하여 기술되고 저장된다.
- (2) SPARQL(SPARQL Protocol and RDF Query Language): 시맨틱 웹의 질의어 작성에 관한 표준이다. 여러 시스템에 나뉘어 저장된 데이터를 질의하기에 용이하도록 만들어졌다.
- (3) WOL(Web Ontology Language): 시맨틱 웹의 스키마를 정의하는 언어에 대한 표준이자 지식 표현을 가능하게 하는 언어 표준이다. WOL을 이용하면 단위 개념들을 정의할 수 있을 뿐 아니라 다른 개념들과 조합해 전혀 새로운 개념도 표현할 수 있게 된다. 다양한 목적으로 만들어진 애플리케이션마다 그에 맞는 독자적인 개념을 필요로 하는데 이때 OWL 기반의 개념 생성 기능은 이를 가능하게 하는 핵심 역할을 담당한다.

시맨틱 웹 기술은 현대적 관점에서 지식 그래프로 해석되기도 하는데 이는 시맨틱 웹의 개념을 명확하게 정의하고 적용하기 위한 노력의 일환이다. 시맨틱 웹의 개념에 대한 불완전한 이해와 오해가 초기의 시맨틱 웹의 발전을 더디게 한 큰 이유이기 때문이다. 완벽한 이해 없이 개발된 애플리케이션들이 이 기술의 복잡성과 구현 규모를 간과하고 성급하게 개발 및 상용화를 추진하도록 부추기면서 초기의 부진을 가져온 면이 있다. 지식 그래프 구조를 사용함으로써 시맨틱 웹 표준을 개발하고 응용하는 개발자들과 의사 결정권자들에게 개발에 필요한 엄밀한 기준과 정확한 틀을 제공할 수 있게 된다.

3. 시맨틱 웹의 목표

시맨틱 웹의 목표는 컴퓨터로 하여금 문서의 내용을 구성하는 엔티티를 이해할 수 있도록 하는 것이다. 시맨틱 웹은 웹 페이지를 구성하는 단순 구조물로 존재하는 데이터에 의미를 부여

하고, 데이터 간에 관계를 형성해 준다.

검색 엔진을 예로 들어 설명해 보자. 사용자가 검색하려는 키워드가 문맥에 따라 다른 의미를 가질 경우, 대상 검색 엔진이 사용자의 의도를 이해하고 있다면 그에 맞는 정확한 결과만 신속하게 산출해 줄 것이다. 그렇지 못한 경우 검색 키워드가 같은 모든 의미를 포괄적으로 적용해야 하기 때문에 관련 없는 검색 결과도 다수 포함될 것이다. 시맨틱 웹은 컴퓨터가 언어의 의미를 파악하도록 구현되기 때문에 검색 과정에 사용되는 키워드에 대해서도 문맥을 이해하고 상황에 맞는 검색 결과를 산출하는 것이 가능해진다.

한 발 더 나아가서, 해당 키워드로 검색하는 의증을 이해할 수 있기 때문에 이와 연관된 또는 이에 이어지는 후속 사건들을 예측하고 적절히 대응할 수 있게 된다. 웹사이트를 통해 항공권을 예약할 경우 연계된 호텔 예약 화면이 연이어 제시되는 앱을 대표적인 예로 들 수 있다.

여기서 한 가지 고려해야 할 사항은 바로 여러 정보 주체들 간의 데이터 공유나 협업 이슈이다. 현재의 웹 구조에서 이런 유형의 연계 서비스를 만들기 위해서는 사전에 정보를 주고받는 주체들 간에 별도의 계약을 맺어야 한다. 시맨틱 웹이 완성되면 모든 정보를 누구나 무료로 자유롭게 사용할 수 있기 때문에 협약이나 계약이 불필요해진다.

4. 인공지능과 시맨틱 웹

시맨틱 웹 초창기에는 논리와 추론 중심의 지식 표현이 궁극적인 구현 목표로 자리잡았다. 지식 표현에는 데이터, 특히 방대한 양의 시맨틱 데이터가 필요했지만 소수의 전문가들만 참여하는 시맨틱 데이터 생성의 문제가 걸림돌로 작용했다.

웹 영역이 점차 확장되고 웹 기술도 급격히 발달하면서 시맨틱 정보를 기술해야 할 대상 영역도 폭발적으로 증가했다. 웹의 모든 문서나 내용에 대해 사용자들이 자발적으로 참여하여 시맨틱 데이터를 지속적으로 생산해 달라고 요청하는 것은 초기에 소수 전문가들을 움직여서 시맨틱 정보를 추가하려던 것보다 현실적으로 더 불가능해 보인다.

가능성 여부의 문제를 잠깐 접어두고 구현 방안에 대해 생각해보자. 시맨틱 웹을 실현하기 위한 실천 전략으로는 상향식(Bottom-up) 접근 방식과 하향식(Top-down) 접근 방식이 주로 거론된다.

상향식 접근법은 인터넷에 존재하는 기존의 모든 문서에 정보를 추가하여 기계가 이해할 수 있는 형태로 변환하는 전략이다. 이때 추가되는 정보라 함은 웹사이트를 구성하는 엔티티들과 그들 간의 관계를 의미한다. 이와 같은 작업은 주로 RDF나 마이크로포맷을 통해 이루어진다. 상향식 접근법의 경우 온톨로지 구조가 너무 복잡해지거나, 일관성을 잃기 쉽다. 온톨로지를 구축하는 작업은 ER 모델에 기반한 DB 스키마 설계보다 훨씬 어렵고 복잡한 과정을 거쳐서 만들어지기 때문이다.

하향식 접근법은 지금 현재 상태의 웹사이트에 담겨있는 데이터를 더 잘 이해할 수 있는 애플리케이션을 개발하는 방식을 택한다. 잘 정리된 온톨로지 구조를 확보할 수 있다는 장점은 있지만, 작성자의 경험에 의존해야 하고 적용 범위가 제한적일 수밖에 없다는 단점을 가진다. 하향식 접근법을 선호하는 사람들은 상향식 접근법이 웹3.0 환경에 전혀 들어맞지 않는다는 점을 지적한다. 인터넷에 존재하는 모든 웹사이트나 텍스트에 정보를 추가하여 기계가 이해할 수

있도록 설명을 더하는 일은 물리적으로 가능하지 않다는 입장이다.

또 다른 방식으로 생산자(Producer)와 소비자(Consumer)의 관점에서 시맨틱 웹 구현 문제에 접근하는 방안을 생각해 볼 수 있다.

먼저 생산자 입장에서는 시맨틱 태그를 자동으로 달아주는 자동 주석기(Semantic Annotator) 형태의 접근 방법이 있다. 이 기술은 자연어 처리 기술에 크게 의존하기 때문에 초기의 결과는 만족스럽지 않았다. 최근 이 분야의 기술은 비약적 발전을 이루어 시맨틱 웹 구현을 위해 여러 측면에서 활용 가능한 수준에 다다랐다. 실제로 GPT-3 등 인간의 언어 생성 능력에 필적할 수준으로 문장을 만들어 내는 딥러닝 모델들이 개발되고 있으므로 이를 활용한 자동 태깅 기능의 실현 가능성은 매우 높다.

다음으로 소비자 입장에서 접근하는 방안이 있다. 소비자 입장에서는 현재 웹이 완전히 시맨틱 웹으로 전환된 것이 아니기 때문에 다양한 수준의 웹 페이지들(완전히 시맨틱 웹으로 구현된 웹과, 시맨틱 개념이 전혀 반영되지 않은 웹, 그리고 이 둘 사이에 존재하는 웹 등)을 시맨틱 웹의 형식과 문맥으로 받아들이고 이해할 수 있는 기능이 필요하다. 구글의 트랜스포머 계열 모델들을 필두로 자연어 처리 분야에 괄목할 만한 발전이 있었고, 이러한 기술 트렌드는 다양한 수준의 웹을 시맨틱 문맥에서 처리할 수 있는 애플리케이션들의 출현을 가능케 한다.

한 발 더 나아가, 자연어 처리 기술의 발전 속도에 비추어 보면 과연 시맨틱 태깅 작업이 따로 필요한지, 즉 현재의 월드 와이드 웹을 구성하는 콘텐츠들의 형식을 굳이 시맨틱 웹의 형식에 맞추어 변형할 필요가 있는지에 대해서도 다시 생각해 볼 필요가 있다.

5. 시맨틱 웹과 개인화

시맨틱 웹의 개인화 측면을 담당하는 표준으로 APML(Attention Profile Markup Language)이라는 마크업 언어 표준이 있다. APML은 사용자 프로파일을 기술할 때 사용하는 언어이다. APML로 기술된 사용자 프로파일에는 개인의 관심사나 기호가 기록되며 이 정보는 사용자가 접속하는 웹 사이트와 그 안의 데이터를 필터링하는데 사용된다.

웹 사용자들은 자신의 취향에 맞는 개인화된 정보를 원하고 기업들은 이를 지원하기 위해 사용자 정보를 수집한다. 현재의 인터넷 서비스 기업들은 이러한 목적으로 개인 사용자가 설정한 선호 정보보다 훨씬 많은 정보를 수집하여 개인별 정보 필터링에 적용하고 있다. 구글, 페이스북(메타), 아마존, 야후 등은 일찍부터 웹사이트 방문자의 행동 양태를 기록하고 이용하고 있다.

오래 전부터 기업들의 개인 정보 유출 사고와 잘못된 기업 간 공유로 개인 정보보호 관리에 어려움이 있다. 이런 형태의 정보 공유 및 이용 구조를 개선하고 개인 정보 유출 문제에 대한 예방책이 제시될 때 시맨틱 웹의 구축이 속도를 낼 수 있을 것이다.

6. 시맨틱 웹과 클라이언트 기반 딥러닝

웹3.0, 그리고 시맨틱 웹의 대표 플랫폼은 웹 브라우저이다. 브라우저 중에서도 데스크탑용 보다는 이동단말용 브라우저의 사용율이 높다. 인공지능 애플리케이션 진영에서도 클라이언트 기

반 딥러닝 기술에 관한 연구개발 사례가 늘어 나는 추세이기 때문에 시맨틱 웹의 주력 형상은 이동단말용 브라우저가 될 가능성이 높다.

이동단말용 브라우저에 시맨틱 웹을 구현하기 위해서는 이동단말의 한정된 자원을 효율적으로 이용할 수 있는 인공지능 프레임워크를 사용해야 한다. 현재 텐서플로우 라이트(TFLite)와 Tensorflow.js가 각광받는 프레임워크이다. TFLite는 이동단말 자체에 설치되고 작동되는 애플리케이션을 구현하는 방식이고, Tensorflow.js는 웹 사이트에 모델을 적재하여 브라우저에서 구현하는 방식이다.

TFLite 기반의 애플리케이션은 이동단말에 장착된 장치에 직접 접근하는 방식이므로 사용감이 좋고 자원도 효율적으로 이용할 수 있다는 장점이 있다. 반면, Tensorflow.js 기반의 웹 페이지는 이미 펼쳐진 플랫폼 구조를 활용할 수 있다는 장점을 갖는다. 이동단말의 특정 장치를 구동하는 기능을 구현할 경우 브라우저를 경유해야 하므로 아무래도 부자연스럽다는 점과, 매번 모델 파라미터 전체가 브라우저로 다운로드 되어야 하므로 모델 전송을 위한 일정량의 대역폭이 추가로 필요하다는 점을 단점으로 들 수 있다. 각 사용자 별로 모델을 학습시키기고 유지시키기 어렵기 때문에 연합형(Federated) 모델의 구현도 쉽지 않다.

보완해야 할 몇 가지 단점에도 불구하고 브라우저 기반의 딥러닝 모델 구현 방식은 시맨틱 웹으로 가기 위해 반드시 건너야 할 가장 중요한 길목에 위치한 것으로 보인다. 현대인의 삶은 이동단말 없이 영위될 수 없는 수준에 도달했으며, 미래에도 이동단말이 삶에 미치는 영향력은 점점 커질 것이기 때문이다. 대부분의 개인화 관련 정보는 사용자 단말을 통해 수집된다는 점도 단말의 스크린이 시맨틱 웹과 사용자가 만나는 최대 접점이 될 것이라는 전망을 뒷받침한다.

7. 맺음말

웹3.0은 소셜 미디어, 스트리밍, 온라인 쇼핑 등으로 대표되는 웹2.0의 성공을 계승하고 이를 능가하는 기술로서 잠재력을 인정받고 있다. 새로운 애플리케이션의 출시를 촉진하고, 사용자와 서비스 간 상호작용의 구조와 방식을 향상시키기 위해 시맨틱 웹, 인공지능, 기계학습 등 웹3.0의 핵심 구성 요소들을 적극 융합하고 활용해야 할 필요가 있다.

웹3.0 환경에서는 시맨틱 웹 기술과 자연어 처리 기술을 통해 컴퓨터가 인간과 유사한 방식으로 데이터와 정보를 이해할 수 있게 될 것으로 기대된다. 또한 웹 공간 전체가 시맨틱 웹으로 전환되는 때가 오면 이를 인간 지식을 총체적으로 저장하고 있는 지식 데이터 베이스로 삼아 거대한 궁극의 검색 시스템을 구축하는 모습도 상상해 볼 수 있다.

[참고문헌]

- [1] Wikipedia, "Semantic Web", https://en.wikipedia.org/wiki/Semantic_Web(2022년9월22일)
- [2] Poimaweb, "시맨틱 웹(Semantic Web)", <https://poimaweb.com/html5-semantic-web>(2022년9월22일)
- [3] Joyhong 블로그, "시맨틱 웹과 온톨로지", <https://joyhong.tistory.com/152>(2022년9월22일)
- [4] 요즘IT, "시맨틱 웹에 관하여", <https://yozm.wishket.com/magazine/detail/1377/>(2022년9월22일)

- [5] 조부현, 이복주, "시맨틱 웹을 이용한 웹 변경 탐지 시스템의 설계", 한국컴퓨터종합학술대회 2005 논문집 Vol.32, No.1(B), 2005.
- [6] TRENDING WEB STORIES, "WEB 3.0 THE POWER OF SEMANTIC WEB & AI",
<https://ahmadaluminium.com/web-3-0-the-power-of-semantic-web-ai/>(2022년9월21일)
- [7] Investopedia, "Web 3.0 Explained, Plus the History of Web 1.0 and 2.0",
<https://www.investopedia.com/web-2-0-web-3-0-5208698>(2022년9월21일)
- [8] Cambridge Semantics, "Getting Started with Semantic Technologies",
<https://cambridgesemantics.com/blog/semantic-university/intro-semantic-web/>(2022년9월21일)
- [9] Cambridge Semantics, " Semantic Web Misconceptions",
<https://cambridgesemantics.com/blog/semantic-university/intro-semantic-web/semantic-web-misconceptions/>(2022년9월21일)
- [10] ENTERRA SOLUTIONS, "Artificial Intelligence and the Semantic Web",
<https://enterrasolutions.com/artificial-intelligence-and-the-semantic-web/>(2022년9월21일)
- [11] Gamco, "Artificial intelligence applied to the semantic web",
<https://gamco.es/en/artificial-artificial-intelligence-applied-web-semantics/>(2022년9월21일)

※ 출처: TTA 저널 제203호