

# 웹 사이언스: 웹을 이해하기 위한 학제간 접근

James Hendler, Nigel Shadbolt, Wendy Hall, Tim Berners-Lee, Daniel Weitzner

한국어 번역: 윤석찬 - translated by Seochan (Channy) Yun

## 서론

수 많은 컴퓨팅 인프라와 기술적 측면에서 웹은 위대한 성공을 이끌어 냈지만 아직 제대로 연구된 적이 없다. 여기서 웹을 전반적인 관점에서 바라볼 때 나타나는 기술적 사회적 도전들을 살펴 보고 웹이 성장함에 따라 도출되는 사회적 영향력을 이해하고자 한다. 우리가 미래 웹을 생각해 본다면 '시스템 생물학'이라는 관점에서 체계적 접근이 필요하다.

컴퓨터 사이언스의 각 분야뿐만 아니라 모든 컴퓨팅 영역에서 웹의 영향이 커지고 있는데도 ACM 분류표에서 컨퍼런스 연구 논문을 낼 때 표시할 수 있는 분류는 고작 "기타 분야(miscellaneous)"이다. 또한, 전 세계 각 대학의 컴퓨터 관련 학과 정규과목을 보면 서비스 코스로서 "웹 디자인"이나 개발 언어를 가르치는 "웹 프로그래밍"이 대다수이다. 웹 구조나 통신 규약을 가르치는 과목은 거의 보기 어렵다. 이건 마치 브라우저 바깥에서는 웹이 존재하지 않는 것과 같다. 대다수 "정보학"을 다루는 학교나 학과에서 웹 애플리케이션이나 웹 2.0 을 다루는 과목은 있지만 웹의 기본 원리와 구조 프로토콜을 다루는 곳은 극히 드물다.

간단히 말해 네트워크라는 주제가 학교에서 매우 오래된 학과목이고 TCP/IP 프로토콜로 정의할 수 있는 인터넷이 컴퓨터 과학에서 중요하게 다루어진 것이 그 원인 중 하나다. 웹이 그 자체의 프로토콜과 알고리즘 구조적 원리를 가지고 있는데 CS 분야에서는 인터넷 상위에 있는 애플리케이션 정도로 취급하면서 그 자체로 중요하게 다루지 않았다.

웹은 컴퓨팅 역사뿐만 아니라 인류의 커뮤니케이션에서 가장 많이 사용될 뿐만 아니라 변화의 주체이기 때문에 더 이상한 일이다. 학계에서 연구하고 가르치고 출판하는 것도 웹을 통해 바뀌었다. 산업계에서는 아예 그 자체로 새로운 영역 (여러 영역일 수도 있고)이 생겼을 뿐 아니라 산업체 전체에서 제공하는 서비스를 묶어 내고 연결시키는 데도 영향을 끼치고 있다. 정계에서도 각 정부들이 시민들과 의사 소통하는 방식이 바뀌고 있을 뿐 아니라 다수가 소통하는 방식에서 정부가 가장 첫 번째로 선택하는 영역이 되고 있다. 미국 대통령 선거에서 후보자들이 유튜브 동영상을 통해 질문을 받았을 정도이다. 인류의 숫자를 10<sup>10</sup>으로 놓고 본다면 웹 문서의 숫자는 그보다 훨씬 많다고 추정<sup>11</sup>되고 있다. 물론 컴퓨팅 자체가 웹에

서 중요한 영역이었다. 우리가 늘 쓰는 웹은 처음 발명하기 전부터 있었던 컴퓨터 과학 영역의 발전 결과에 기초하고 있다. 오늘날 검색 엔진 역시 1960 년대의 정보 변환 기술에 기반을 두고 있다. 1990 년대에<sup>9,23</sup> 여러 혁신으로 인해 만들어진 검색 알고리즘들은 현재 웹 사용에 있어 중추적 역할을 하고 있다. 분산 컴퓨터 클러스터를 통해 데이터를 처리하는 오픈 소스 소프트웨어인 Hadoop 같은 새로운 자원으로 인해 학생들이 이러한 알고리즘을 탐구하고 MapReduce 병렬 알고리즘<sup>11</sup> 같은 대용량 웹 프로그래밍을 실험하는 게 가능해 졌다. 이러한 것은 이전에 대학에서 불가능하던 것들이다.

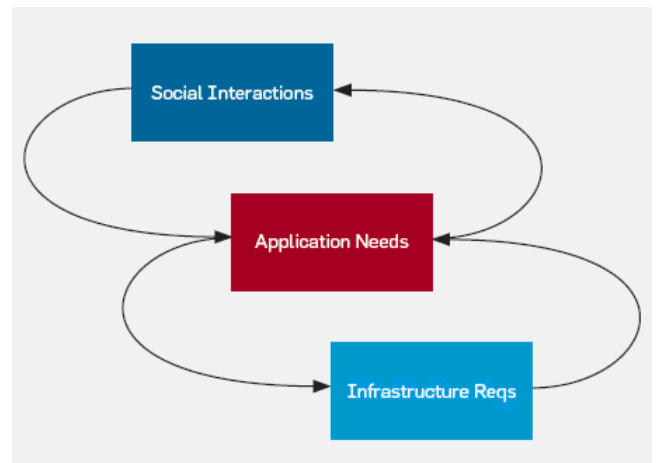


Figure 1. 웹을 통한 사회적 상호작용

웹 상에서 인간 사이의 소통의 측면에 대한 연구 역시 이루어졌다. 소셜 네트워킹, 태깅, 데이터 통합, 정보 검색, 웹 온톨로지 같은 다양한 연구 주제들이 새로운 "소셜 컴퓨팅"이라는 영역으로 정보학문을 다루는 학교에서 연구 중이다. 일반적인 네트워크 특성, 컴퓨팅에서 정치 및 정책적인 측면에서 상호 연관성 체계 및 경제적 측면을 다루는 과목들을 개설하고 있다. 하지만, 이들 과목에서도 웹은 다양한 원칙 중 하나의 개체로만 다룬다. 다른 경우, 웹을 다양한 브라우저 사용자 간의 사회적 상호작용을 지원 하는 단순한 동적 콘텐츠 연동 개념으로만 생각하기도 한다. CS 분야나 정보학 코스 양쪽 모두 웹을 그 자체로서 목적으로 다루지기도 보다는 기술적 혹은 사회적으로 콘텐츠를 전달해 주는 수단으로만 연구한다는 것이다.

이제 우리는 새로 시작되는 학제간 학문으로서 웹을 그 자체로 연구 목적으로 삼으려는 웹 사이언스<sup>5,6</sup>에 대해

이야기하려고 한다. 웹 디자인에 의한 사회적 상호 작용, 이를 지원하는 유연하고 공개된 애플리케이션 개발 그리고 대규모 애플리케이션의 데이터를 구조적으로 다뤄야 하는 요구 사항(Figure 1) 등이 서로 연관되어 중요한 연결 고리를 이루고 있다. 하지만, 이들 사이에는 사회학적 애플리케이션과 기반을 이루는 네트워크 연구를 분리시키려는 학문적 경계라는 방해 요소가 있었다. 우리는 이러한 관계들을 인지하고 컴퓨팅에서 웹 관련 연구의 상태를 재조명 해보고 새로 출현하는 도전적인 문제를 웹 과학자로서 연구자들이 탐구해야 할 요소에 집중하고자 한다.

## 웹 사이언스란?

일반적으로 과학이란 관찰된 현상을 설명하거나 이를 통해 나온 법칙을 이끌어 내는 분석적 학문활동을 말한다. 컴퓨터 과학은 예상되는 동작을 지원하기 위해 만들어진 형식과 알고리즘을 종합적으로 다룬다. 웹 사이언스는 이 두 가지 패러다임을 합치는 방법을 신중하게 찾으려고 한다. 웹을 현상으로 이해해야 할 뿐만 아니라 미래 성장 및 역량에 대해 만들어지는 것으로 연구해야 할 필요가 있다.

거시적 관점에서 웹은 인공적인 언어와 프로토콜로 이루어진 약간 공학적인 인프라 구조이다. 하지만, 이는 역시 거시적 관점에서 새로운 영역으로 웹에 의해 만들어진 정보를 인간이 생산하고 링크를 걸고 소비하는 상호 작용을 포함하고 있다. 어떤 것은 공학적으로 만들어지는 게 가능한 것도 있고 아닌 것도 있다. 또한, 웹이 인간 상호 작용의 시스템의 일부라는 점도 기억해야 한다. 이는 사회에 크게 영향을 미치고 있는데 이전보다 더 많은 영역에서 정보를 만들 수 있는 기회와 도전을 만들어 내고 있기 때문이다.

웹을 이해하는 최선의 방법은 알고리즘으로 분석해 온 개별 애플리케이션과 함께 자산으로서 연구해 왔던 규약(프로토콜)의 조합으로 보는 것이다. 하지만, 소프트웨어 공학적인 모범 사례로 여겨져 온 전통적인 CS의 사이클인 요구 사항 정의, 설계, 개발 및 테스트 같은 방식으로 보기 어렵다.

**대용량 시스템으로 인해 사회적 영향이나 거시적 기술을 분석함으로써 예측 못했던 새로운 영역이 나타날 지도 모른다.**

Figure 2는 웹 개발의 새로운 방식을 표시하고 있다. 소프트웨어 애플리케이션은 알고리즘과 설계 같은 기술에 기반하고 있지만 사회적 관점을 포함하고 있다. 즉, 웹 개발을 하나의 사용자와 하나의 기계로 한정해서 만든다는 건 정말로 모순되는 일인 것이다. 일반적으로 웹 서비스는 작은 그룹에서 테스트 하고 (사내 배포 같은) 한정된

사용자들에게 서비스를 시작한다. 이 때, 시스템의 미시적인 부분을 테스트 한다. 몇몇 경우에서 사람들이 점점 미시적 시스템을 받아 들이면 소문의 크기가 점점 커진다. 예를 들어 첫 웹 브라우저였던 Mosaic 이 1992 년 처음 출시되었을 때 사용자 수는 매우 빠르게 성장하여 첫 회에 백만 명이 넘었다. 최근에 사진 공유 서비스인 Flickr 나 동영상 업로드 서비스는 YouTube, 소셜 네트워크 서비스인 mySpac 나 Facebook 도 마찬가지다.

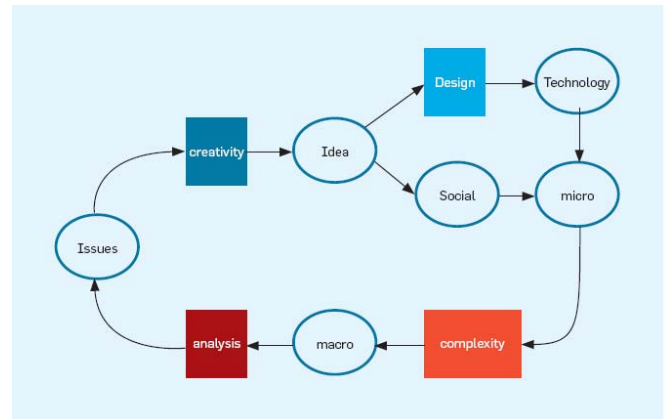


Figure 2. 웹 애플리케이션 개발의 새로운 방식

사용자들끼리 상호 작용하는 미시적 사용이 예상치 못한 방식으로 거시적 시스템에서 다른 방식으로 분석될 수 있다는 점이 흥미롭다 하겠다. 이들 거시 시스템은 미시적인 규모에서는 발생하지 않는 도전을 일으킨다. 예를 들어 Mosaic 을 널리 배포했을 때 성장하는 웹 콘텐츠를 적당하게 찾아야 할 필요성이 생겼다. 검색은 그 후에 산업에서 중요한 애플리케이션이 되었다. 또 다른 경우 대용량 처리 시스템은 미시적 체계 내에서의 사회적 영향에 대한 분석으로는 예측할 수 없는 새로운 자산들이다. 이러한 이슈들을 다루는 것은 다음 세대 기술이 무엇인지를 알 수 있게 해 준다. 예를 들어 검색 엔진의 성공으로 인해 필연적으로 검색 랭킹을 높일 수 있는 예상치 못한 다양한 알고리즘 기법이 나오게 되었는데 반대로 이를 물리칠 수 있는 또 다른 검색 기술이 나오게 된다.

웹이 성공하고 웹 애플리케이션들이 발전시킨 본질적인 것이 우리가 원하는 것을 만들어 내는 시스템을 디자인하는 새로운 방식을 개척해 왔다는 것이다. 우리가 할 수 있는 최선의 것은 미시적인 설계와 구현이다. 거시적인 효과가 나는 바른 기능을 만들 수 있다면 우리는 어떻게 그것을 알 수 있을까? 특히, 웹 기술의 성공과 실패가 사용자들 사이의 사회적 상호 작용과 관련 되어 있기 때문에 나중에 이야기 하겠지만 웹에 대한 이해가 단순히 기술적 이슈의 분석을 넘어서 수백만 사용자의 사회적 동력을 분석이 필요하다.

웹의 폭이 넓어지고 다중 사용자 기반으로 바꿈에 따라 수학, 컴퓨터과학, 인공 지능, 사회학, 심리학, 물리학 및 경제학 같은 분야가 참여하는 학제간 연구가 필연적이 되었다. 우리는 컴퓨터 과학자들이 웹이 광범위하게 채용되고 사회 구조와 정치 시스템, 회사 및 대학이 다양한 영

향을 받고 있기 때문에 각종 도전들이 있음을 널리 알리고 학문으로서 성장시켰으면 한다.

## 웹 그래프를 넘어서

웹을 이해하는 방법 중 컴퓨터 과학에서 가장 익숙한 것이 웹 페이지(HTML 문서)를 나타내는 노드와 그것을 하이퍼 링크로 연결한 그래프이다. 우리가 이것을 최초 연구에서 웹 그래프<sup>22</sup>라고 불렀다. 웹 그래프를 한 단계 더 들어가 Kleinberg et al.<sup>3</sup>와 Kumar et al.<sup>24</sup>의 연구에 따르면 지수함수의 분포를 보여 준다. 이는 Broder et al.<sup>10</sup>의 연구에서도 그래프에서 최고점의 다각화를 위한 유사한 영향을 보여 주었다. Dill et al.<sup>12</sup>의 중요한 결론은 다양한 방식으로 수집한 대량 웹 데이터 샘플을 통해 볼 때 2005 년에서만 하루에 수백만 페이지가 생성되고 있다는 점이다. 웹 그래프가 성장하고 그 진화를 빨리 알아낼 수 있는 모델을 찾는 다양한 방법을 제안했다. Donato et al.<sup>14</sup>에서는 이러한 모델과 연구에 대해 설명하고 있다.

그래프가 성장하고 분석됨에 따라 그래프의 속성을 파악하기 위한 여러 알고리즘이 제안 되었다. 예를 들어 HITS 알고리즘이나 페이지 랭크<sup>9</sup>는 한 페이지에서 다른 페이지로 하이퍼 링크 하나를 추가하는 것이 "권위"에 대한 추천 행위로 받아 들이고 있으며 웹 페이지를 찾는 검색 기능을 막강하게 해졌다. 현대적 검색 엔진들이 랭킹을 높이기 위한 다양한 어뷰징 시도에 맞서 페이지 기반 랭킹 계산 방식을 벗어나 다양한 문제 해결 방법을 도입하고 있어 웹 그래프 기반 모델은 크롤러의 내부 문제에서 웹 검색을 넘어서는 랭킹 자신을 형성하고 있다.

이러한 웹 그래프에 있는 링크는 전체 웹 중에서 인식자로 제공되는 URI 라는 주소 체계 아래 있는 문서를 HTTP 프로토콜로 GET 요청에서 반환하는 단일 인스턴스를 말한다.

예를 들면 <http://www.acm.org/publications/cacm> 을 일반 웹 브라우저에 입력하고 HTTP 접속을 하면 *Communications of the ACM* 이 라는 HTML 문서를 반환해 준다. 하지만, 그 문서에는 내용 그 자체뿐만 아니라 이미지나 아이콘 파일 같은 다른 URI 들도 포함하고 있음을 유의해야 하고, 문서의 포맷을 위한 CSS 나 DTD 파일에 대한 정보도 포함하고 있다. 따라서 *Communication* 이라는 단일 링크를 클릭해서 웹 페이지를 본다는 것은 여러 개의 서버에 여러 개의 호출을 의미하는 것이다. 이 문서를 쓰는 지금도 이 페이지에는 적어도 20 여 개 이상의 다른 HTTP-GET 호출을 하고 있다. 검색 엔진 크롤러 역시 이들 링크를 인지하고 웹 그래프를 만들면서 하나의 스냅샷을 뜨고 있을 것이다.

하지만, 웹 그래프는 이들 기능을 기반으로 하는 호출과 처리의 한 부분으로서 웹을 추상화 한 것이다. 웹 그래프가 크기에 제한이 없다는 사실이 중요하므로 우리가 웹이라고 부를 수 있는 규약과 서비스가 만들어지고 있는 것이다. 웹은 *The Architecture of The World Wide Web*,

*Volume 1*<sup>21</sup>에서 기술 하듯이 "자원의 위치와 상태의 표현, 에이전트와 그 자원 사이의 상호 작용을 지원하는 프로토콜 같은 핵심 디자인 항목"들로 구성되어 있었다.

웹 의 이런 특징은 호출되는 종류에 따라 서로 다른 결과를 제공하게 된다. 예를 들어, HTML 요청을 처리할 때 서버 설정에 따라 어느 백-엔드 서버 중 하나 같은 클라이언트 요청처럼 숨어서 결과를 보여주는 조건이 생길 수도 있다. 상태를 표시하는 쿠키가 사용될 수도 있기 때문에 사용자의 이전 행동에 따라 서로 다른 결과물이 보이기도 한다. 이 말은 웹 그래프에서 사람에 따라 다른 링크가 생긴다는 것을 의미한다. 사용자에게 따라 다른 링크 상태는 지금까지 우리가 이해하고 있는 웹 그래프 모델과는 완전히 다른 것이다. 특히 애플리케이션으로 웹은 링크 기반의 하이퍼텍스트 페이지의 준정적(quasi-static) 모델에서는 간단히 분석하기 어렵다. 예를 들어, 많은 웹 사이트들이 폼 양식을 이용해서 많은 정보를 서버로 보내고 있으며 여기서는 보이지 않는 이른바 "깊은 웹"이 존재하고 있다. 특히나 링크가 명시적이지 않은 많은 종류의 HTTP-POST 가 웹 그래프에서 중요한 HTTP-GET 대신 사용되고 있다. 또 다른 사이트들은 상태만을 전달하는데 GET 호출을 사용해서 복잡한 URI 들을 만들어 내서 실제 자원의 위치를 파악하기 어렵게 하고 있다.

웹 애플리케이션에서 주로 사용되는 상태를 표시하는 URI 들은 분석하기 매우 어렵다. 2007 년 6 월에 구글의 Udi Manber 는 왜 웹 검색이 어려운가 하는 이슈를 소개 하면서 하루 평균 20~25%의 웹 페이지가 이전에 있지 않았던 URI 들이고 이는 서버 기반의 정보를 이용해서 생성되고 있는 것이라고 한다. 웹 그래프 모델은 오로지 문서를 호출하는 데 유용하다. 많이 알려진 대로 구글이 하루에 1 억 건 이상의 검색 질의를 받는다면 그 중 20%인 2 천만 개의 URI 이 새로운 URI 라는 것이고 이는 웹 그래프가 계속 바뀌고 있고 초당 200 건 이상이 변하고 있다는 것을 의미한다. 이런 링크들을 거둬 제공하는 법칙으로 따라갈 수 있을까? 동일 성장 모델이 이런 행동을 설명할 수 있을까? 간단히 말해 모른다.

---

**오늘날 상호 교환 애플리케이션은 서로 많이 분리되어 있다는 것을 보여주는 제한적인 초기 단계의 소셜 머신에 속한다.**

---

웹을 그래프로서 분석하는 것은 짧은 시간 단위에서 이들의 동적인 특성을 무시하는 것이다. 웹 사용자들이 여러 번 같은 링크를 클릭해서 서버 부하가 짊어져서 서비스 거부 공격 같은 효과가 나는 현상들은 웹 그래프 모델로 설명하기 어렵고 그래프 기반 분석에서 광범위하게 쓰는 용어들로도 표현하기 힘들다. 네트워크 수준에서 이들을 표현하지만 프로토콜 수준뿐만 아니라 매초당 수천 대의 서버가 받는 수백만의 요청으로부터 나타나는 현상들 뿐만

아니라 웹의 핵심 측면이 무시되고 있다. 웹의 이런 역동성이 10 여 년 전에 분석된 적 <sup>20</sup>이 있지만 (i) 웹 콘텐츠 양의 기하급수적 성장, (ii) 웹 서버와 애플리케이션의 다양성과 숫자의 변화, (iii) 전 세계 여러 곳의 다양한 사용자들의 증가 같은 조합에서 새로운 모델이 나와서 유효성 여부를 확인하지 않고서는 비슷한 분석이 거의 불가능하다. 이러한 모델은 변화된 웹의 세부 구조를 주목하고 그곳에서 일어나는 다양한 사용자의 인터랙션과 복잡성을 이해했을 때론 가능하다.

게다가 오늘날 정교한 웹 사이트들은 웹 브라우저 내의 스크립트를 이용해서 강력한 사용자 인터페이스 기능을 제공하고 있다. 이들 애플리케이션은 오픈 API 를 통해 원격 데이터 모델에 뒷단에서 접근하고 있다. 이들 애플리케이션 구조는 사용자들과 기업들이 기존 웹 서버의 강력한 저장 공간과 사용자 컴퓨터와의 강력한 조합에 의해 짧은 시간 안에 다양한 형태의 웹 시스템을 만들어 낼 수 있도록 해 주고 있다. 그러한 시스템은 거시적인 규모의 웹에서 매우 흥미롭지만 아직까지 잘 이해하지 못하고 있다. 그러한 시스템이 실제로 안정한가? 또는 공정한가? 새로운 흐름을 효과적으로 만들어 낼 것인가? 만약 그게 규제되어야 한다면?

비슷하게 사용자 생산 콘텐츠(UGC) 역시 친구라는 제한된 접근을 제공하는 간단한 시스템으로 개인적인 정보를 저장해 오고 있다. 이들 정보들은 아직 대규모 분석이 된 적이 없다. 또 어떤 사이트는 사용자의 친구인 척 하는 것만으로도 접근을 허가하고 있다. 이를 위해 여러 서드 파티 인증 수단이 제공 되고 있기도 하다. 복잡한 시스템도 조각 조각 만들어지고 있고 내 상사가 이 사진을 못 볼 거라는 확신이 깨질 수도 있다.

이러한 토론을 하는 목적은 웹 프로토콜의 상세나 웹 모델링적 접근의 상대적 장점에 빠지지 말고 현재 웹의 상태에 대해 비판적으로 계속해서 바라볼 것을 요청하는 것이다. 프로토콜 및 이슈를 이해하는 것은 웹의 기술적 구조를 이해하는 것과 웹의 동적 특징과 모델을 이해하는데도 중요하다. 웹 시스템을 분석하려는 시도는 이들의 동적인 특성을 이해해야 가능하다. 이러한 분석과 모델링은 컴퓨터 과학자에게도 중요한 도전이 될 것이다.

## 거듭제곱의 법칙에서 사람들까지

수학을 기반으로 웹을 분석하는 데는 여러 번의 실패가 있었다. 다양한 웹사이트를 이용하는 구조가 (수학적으로) 흥미로운 주제임에 반해 시간에 따라 사이트의 동작이 변화하는 데는 대개 유용하지가 않다. 2 백만 개의 영문 항목과 6 백만 개의 언어별 페이지를 가진 온라인 백과사전인 위키백과(www.wikipedia.org)를 예를 들어 보자. 이들 내부의 문서간에는 매우 논리적인 링크들이 있으며 이들은 웹의 기본 구조와도 비슷하나 관리가 되는 텍스트 구성이기 때문에 다른 부분을 가지고 있지 않다. 여러 가지 답변이 가능할 것 같은데 Figure 3 에서는 이들 중 하

나의 결과를 보여 준다. 이 경우 RDF 를 이용해서 위키백과어의 링크를 모아 둔 DBpedia (dbpedia.org)에서는 링크 라벨의 관점에서 분석을 했다. 웹 기반 태깅 시스템에서 태그의 사용에서 보여지는 특징과 유사한 효과를 추측 <sup>29</sup> 하게 하는 증거 <sup>16</sup> 들이 있다. 현재는 이들 결과들이 웹 그래프의 규모와 상관없는 특징을 해석하는데 사용된 특별한 방식 <sup>3</sup>의 모델과 분리되어 있는지 여부를 조사하는데 집중되어 있다.

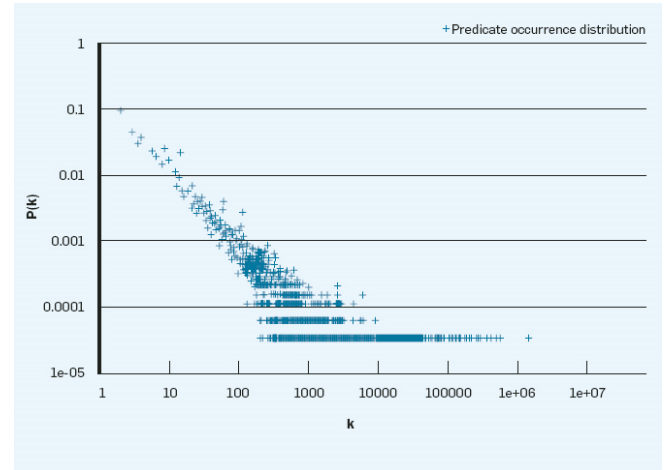


Figure 3. 위키백과어의 링크 구조 분석 결과

안타깝게도 이런 효과를 설명을 해도 위키백과어를 다른 측면은 또 이들 모델로 설명되지 않고 이들 특성을 따르지 않는다. 위키백과어는 MediaWiki 라는 소프트웨어 패키지로 만들어지고 있고 이는 자유롭게 배포 가능해서 위키백과어가 아닌 다른 사이트에서 이용할 수 있다.

이들 중 일부는 성공했고 대부분은 실패했다. 솔직히 "기술적" 설명으로는 위키를 이해하기 어렵다. 하지만 같은 소프트웨어이면서도 위키백과어만이 성공한 이유를 그것을 만들었던 사용자들을 이해한다면 가능하다. 문서를 만들고 편집하고 추적하는 기능은 기반 기술이 제공하고 있다. 소셜 모델은 기술이 설명하기 힘든 사람의 상호 작용으로 설명이 가능하다. 모든 "소셜 머신"의 동적인 특성에 대해서는 다양한 학문 분야에서 고도로 복잡한 수십 개의 학술 논문이 발표되었다. 위키백과어에도 그 자체로 이들 정보를 추적하는 페이지를 가지고 있을 정도다. en.wikipedia.org/wiki/Wikipedia:Wikipedia\_in\_academic\_studies

소셜 머신에 대한 아이디어는 *Weaving the Web*,<sup>8</sup> 에서 소개되었는데 웹의 구조적 모양 자체가 개발자 뿐만 아니라 사용자에게도 컴퓨터 기술을 이용해서 온라인에서 실현될 수 있는 소셜 시스템을 만들 수 있는 기능을 제공하고 있다는 가설을 세웠다. 소셜 머신은 위키백과어의 경우 미디어위키 같은 기반 기술은 포함하고 있고, 또한 기술을 관리하는데 사용되는 조직 구성과 정책, 규칙 등도 가지고 있다. 그러한 사례는 현재 웹에서 매우 많이 찾아볼 수 있다. LiveJournal 이나 WordPress 같은 블로그 호스팅 서비스의 경우 애플리케이션뿐만 아니라 관련 블로그, 퍼머링크, 트랙백 같은 블로그스피어라는 것을 이끌어 주는 다양한 소셜 메카니즘을 제공하고 있다. 유사하게 마이스

페이스나 페이스북에서 소셜 네트워킹 사이트에서 흔히 사용되는 규칙처럼 사이트의 흥망을 좌우하는 것이 사용자 커뮤니티를 지원하는 규칙이나 정책이라 할 수 있다. 웹 기술의 성공과 실패는 점점 이들 소셜 기능이 얼마나 잘 구현하느냐에 달려 있는 것처럼 보인다. 성공적인 애플리케이션을 만드는 엔지니어의 능력은 이들 시스템의 사회적 측면 [b](#)의 기능을 얼마나 잘 이해하고 만드느냐에 달려 있다.

오늘날 상호 교환 애플리케이션은 서로 많이 분리되어 있다는 것을 보여주는 제한적인 초기 단계의 소셜 머신에 속한다. 우리는 (i) 언젠가 오늘날 우리가 가지고 있는 것보다 더 효과적이고 중요한 소셜 머신이 있을 것이고, (ii) 이는 전혀 다른 사회적 프로세스로 상호 연결되면 바로 웹에서 상호 연결되며, (iii) 단일 프로젝트나 사이트에서가 아니라 사용자 커뮤니티가 직접 만들고, 공유하고 소셜 머신을 적용할 기술이 필요하다. 이러한 소셜 머신의 새로운 상호 작용이 만들어지고 진화되기 전에 많은 연구 주제와 질문들을 해결해야 한다.

소셜 머신의 근본적인 이론적 특성은 무엇인가? 이를 만들기 위해 어떤 알고리즘이 필요할 것인가? 소셜 소프트웨어의 웹 인프라를 구성하는 기술적 효율성을 디자인 하기 위해 어떤 기반 구조와 원리가 필요할 것인가? 정보 공유라는 사회적 특성을 만들고 이것이 사회적 규범과도 일치하는 그러한 메커니즘을 제공하기 위해 어떻게 현재 웹 인프라를 확장할 것인가?

문화적 차이가 웹 상의 소셜 메커니즘의 사용 및 개발에 어떻게 영향을 끼치는가? 웹은 전 세계에 퍼져 있고 하나의 문화에서 보여지는 특성은 다른 곳에서는 기대하지 않는 결과를 초래할 수 있다. 웹 인프라가 문화적 차이를 극복하게 할 수 있을지 또는 문화 상호적 이해를 증진시킬 수 있을 것인가?

덧붙여 정보와 인간 상호 작용의 중요한 측면은 개인 정보나 저작권, 법적 규범뿐만 아니라 정보의 사용에 대한 묵시적 기대와 신뢰성 같은 특성에 대한 추론과 표현에 대한 능력이다. 이러한 정보 중 일부는 오늘날 웹에서 사용되고 있지만 이들을 계산하고 형식적으로 표현하는 데는 부족하다. 전통적인 암호 보안 연구와 잘 알려진 접근 통제 프레임워크 같은 경우 온라인 환경에서 이러한 도전에 실패를 경험해 왔다. 게다가 미래 소셜 머신의 형성에도 충분치 않다. 최근 기존의 암호적 접근법으로 개인 정보 보호를 하는 것은 공개 웹 환경에서 실패한다는 것을 보여준 개인 정보 보호에 대한 형식 모델 [b](#)이 제안되었다. 상업적 학술적 저작권 강화 역시 비슷한 문제에 봉착 [27](#)했다. 개인적인 신뢰성을 사회적이고 법적인 규범이 지배하는 정보 사용으로 증가시키려는 기술적이고 사회적인 전통으로 웹을 엮매려는 구조를 되돌아보면서 학제적 연구로 이러한 문제를 풀어보려는 것이 우리가 달성하려는 모범적인 웹 사이언스 연구 분야의 궁극적 목표이다. 정책을 규정하려는 모델을 만드는 시도가 계속 실패하는 것

은 정치 과학 사회적 정보를 상호 교환 가능한 가장 중간 재적인 웹의 능력을 방해하기 때문이다.

더욱이 우리의 온라인 정보 생활에서는 진실과 신뢰성을 판단할 수 있는 사회적 기관이 빠져있는 상태에서 다양한 협업 방식으로 웹 상의 정보가 만들어지고 제공되고 있다. 미래 웹을 설명하기 위해서는 컴퓨팅 구조를 이해하는 것뿐만 아니라 사용자들끼리 상호 작용을 어떻게 지원할 것인가 하는 점도 고려해야 한다. 웹의 사회적 영향을 조사하는 연구의 중요한 측면은 동적인 인간 상호 작용을 지원해 주는 웹 구조를 이용하는 온라인 소사이어티에도 관련된다. trout.cpsr.org 같은 작업 등등은 어떻게 웹이 정치적 공간에서 더 많은 사람들이 참여하도록 독려할 것인가 하는 노력이다. 새로 출현하는 웹에 대한 연구와 기술의 진화, 사회적 니즈등을 종합하는 것은 미래 웹을 디자인 하는데 중요한 초점이다. [30](#)

## 데이터 웹

웹 2.0 이라고 알려진 기술 중 태깅에 대한 광범위한 사용 역시 새로 출현하는 연구 분야 중 하나이다. 뉴스 기사, 블로그, 사진, 동영상 등 수 많은 웹 자원들에 사용자가 지정하는 키워드나 태그라는 주석이 달리고 있다. 이들은 나중에 다른 자원을 검색하거나 브라우징하는데 도움이 된다. 태그를 통한 분류 즉, "폭소노미"라고 부르는 것들이 어떻게 그 물건의 내용을 설명하는데 도움을 주는 메타 데이터로서 사용될 수 있느냐 하는 점을 연구해 왔다.

오늘날 태깅이 관심을 일으키는 까닭은 태깅에서 "사회적 맥락(social context)"[26](#)을 찾을 수 있기 때문이다. 수 많은 태그들은 문맥에서 충분히 모호하고 사용될 수 있는 용어들이다. 예를 들어 사람의 성은 검색에 도움이 되지 않는데도 Flickr 에서 자주 쓰는 태그이다. 한편 한 명의 개인 사진 같은 특정 소셜 컨텍스트에서 같은 태그는 개인의 특성과 관련되어 있기 때문에 의미가 있다. 메타 데이터로서 태그를 사용하는 것은 그런 사회적 맥락에 의존하고 있고, 이들 도시에서 "네트워크 효과"가 사회적으로 잘 조직화 되어 있다. [19](#)

최근의 시맨틱 웹 기술 응용 [7](#)에서 보듯이 메타 데이터를 이용하는 방법을 볼 수 있고 새로 나오는 웹 기술의 요소들이 중요한 패러다임 변화를 대변하고 있다. 시맨틱 웹은 인터넷과 웹의 초창기부터 기반 네트워크에 대한 추상적 수준의 개념을 표현한 것이다. 인터넷에서는 프로그래머들이 커뮤니케이션 흐름에 따라야 하는 네트워크 기능에 대한 고려 없는 프로그램을 만들 수도 있다. 웹에서는 프로그래머와 사용자들이 컴퓨터에 저장하고 상호 교환하는 부분의 걱정 없이도 상호 연결된 문서를 작업할 수 있다.

시맨틱 웹에서는 프로그래머와 사용자들이 웹 상의 기반 문서가 사람, 화합물, 동의어, 별 같은 실 세계의 물건을 잘 표현하는지에 대한 고려 없이도 이들을 참조할 수 있을 것이다. 기초적인 시맨틱 웹 기술을 정의하고 널리 퍼

트리는 동안 웹을 사용하는 사람으로부터 이러한 연결이 가능하다는 것을 보여주는 일련의 작업이 있었다.<sup>28</sup>

시맨틱 웹 분야는 활동에 대한 두 가지 연계된 원칙을 반영한다. 그 하나는 웹의 데이터에 관계된 것이고 나머지 하나는 의미를 가진 도메인에 대한 것이다. 먼저 데이터가 연결된 애플리케이션에서 큰 혁신이 일어나고 있어서 의미를 전달하는 웹 애플리케이션에 초점을 맞추어 개발할 뿐만 아니라 웹의 기초가 되는 URI 를 이용해 데이터 엔티티를 연결하는 강력한 메커니즘을 제공하고 있다. RDF 를 이용하면 이들 애플리케이션은 SPARQL 언어를 통해 질의 그래프 기반의 트리플 저장 데이터베이스에 초점을 맞출 수 있다. 이는 REST 기반의 모델을 사용하고 있고 이미 존재하는 다양한 스키마의 다중 소스와 데이터를 통합하려는 인터넷 포털이나 웹 애플리케이션을 만드는 데 도움을 줄 수 있다. 두 번째는 웹 온톨로지 언어나 OWL 같은 것을 기반으로 애플리케이션 도메인의 의미적 설명을 표현하거나 지식 기반에서 필요한 웹 또는 비 웹 애플리케이션 사이의 상호 참조를 제공해 줄 수도 있다.

## 웹을 관찰하는 가장 많은 지식을 가진 연구자들보다 웹 그 자체가 더 크게 변화하고 있다.

현재 연구는 어떻게 시맨틱 웹 데이터베이스가 전통적인 DB 기반 접근과 관련하여 대규모 저장 처리<sup>1</sup> 를 할 것인가 하는 문제에 집중되고 있다. 모델링 관점에서 보면 성능을 희생하지 않으면서도 대규모 지식 기반에서 상호 참조 속도를 높이는 도구를 개발하는 것이 주요 목표이다. 즉 웹 규모<sup>13</sup> 의 성능을 제공하면서 추론과 표현력 사이의 트레이드오프를 할 필요가 없는 방식을 말한다. 웹 온톨로지 구축으로 진행되어온 "하향식" 기술과 데이터로부터 의미를 찾으려는 "상향식" 도구들로 시장이 생겨나고 있다. 시맨틱웹의 백엔드를 개발하는 것은 상향식 방법으로 바뀌고 있는데 기존 웹 서버와 잘 결합하는 오픈 소스 기술을 이용하여 웹 애플리케이션을 잘 연결하고 있기 때문이다. 동시에 하향식 방법을 지원하는 다양한 온톨로지 개발 도구들이 생겨 났고 수천 개의 OWL 온톨로지가 각 분야를 모델링 하기 시작했다. 게다가 웹을 수정하는 규칙 기반 추론을 사용하는 것 역시 주목<sup>4</sup> 을 받고 있다. 미래 웹을 만드는 것은 이렇듯 새로 출현하는 기술의 디자인과 이용 그리고 전통적인 DB 접근법과의 차이점에 따라 시맨틱 웹의 백-엔드를 만드는 방법과 온톨로지 기반 애플리케이션을 위한 도구들을 만드는 것을 포함한.

시맨틱 웹은 핵심 신생 웹 기술 이다. 하지만 무엇이 최선의 방식이고 거시적 효과를 내려면 무엇이 중요한가 하는 다양한 의견이 존재한다. 어떻게 웹 시스템이 발전하는가를 더 잘 이해하기 어려운 것이 규모에서 발생하는 기술적 효과를 알기 어렵게 하고 있다. 데이터베이스에서 감춰져 있는 정보를 공유하고 공적으로 노출 시키는 것로부터 어떤 사회적 중요성이 있는가? 웹 시스템이 미시

적에서 거시적 관점으로 옮겨 가면서 어떻게 바뀌고 있는지 사회적 효과의 가능성은 어떻게 될지 이해해야 한다.

## 결론

웹은 우리가 이전에 해 왔던 것과 크게 달라졌다. 웹의 변화 속도가 웹을 관찰하는 가장 많은 지식을 가진 연구자들 보다 더 크게 변화하고 있다. 미래 인간 사회가 미래의 웹과 필수 불가결하게 연결될 것이라는 것은 피할 수 없는 사실이다. 따라서 우리는 미래 웹 개발이 세상을 더 나은 곳으로 만들도록 할 의무를 가지고 있다. 회사는 웹 상에서 만드는 그들의 제품과 서비스가 사회를 저해시키는 부가 효과를 만들어 내면 안되며 정부와 정책 입안자들은 그들이 제정하고 시행하려는 정책과 법률의 중요성을 충분히 고려해야 한다.

우리가 웹 상에서 학문 상호간의 협력으로 이런 복잡하고 동적인 활동을 더 잘 이해할 수 있을 때까지는 이러한 목표를 성취할 수 없고 이것이 바로 웹 사이언스의 목적이다. 지구 기후에 대한 인간의 행동에 따른 효과에 대한 이론이 증명이 되든 안되든 그 증거를 분석하고 모으는 것이 기후 변화에 대한 과학자들의 임무이다. 웹 과학자들은 어떻게 인간의 행동이 시스템에 대해 엄청난 속도로 성장하는데 영향을 주는지에 대해 다양한 방법을 찾아보고 그 증거를 수집하고 새로운 방법론을 연구하고 있다. 또한 우리는 현재 웹 개발에 영향을 주는 의사 결정이 향후 미래 웹에 어떤 영향을 줄 것인가 주요 기업과 정부들이 인식하도록 하고 웹의 일부나 전체를 제한하면 어떤 일이 일어나는지 고려해야 한다.

컴퓨팅은 웹 사이언스에 중요한 역할을 수행하고 있다. 우리가 웹에 대해 아는 것 중 많은 것이 이러한 컴퓨팅 이론에 배경을 두고 있는 것이 많다. 하지만, 우리가 여기서 보았듯이 미래 성공적인 웹 애플리케이션을 만들 수 있는 중요한 연구가 진행되어야 하는데, 웹의 동적이고 변화 지향적 특징과 함께 사람들과 웹 기술 기반의 "거시적" 상호 교환으로부터 출현하는 새로운 변화를 바라보아야 한다. 따라서 우리는 웹 애플리케이션의 성공과 실패를 좌우하는 중요한 차이인 "소셜 머신"을 이해하고 그들 사이에 상호 연결 및 공유하는 방법을 배워야 할 것이다.

## 감사의 글

Figure 2 는 Tim Berners-Lee 가 2007 에서 발표한 ([www.w3.org/2007/Talks/1018-websci-mit-tbl/Overview.html](http://www.w3.org/2007/Talks/1018-websci-mit-tbl/Overview.html)) 발표 자료에서 발췌했다. 또한 우리는 WSRI Scientific Council ([webscience.org/about/people/](http://webscience.org/about/people/))의 다른 동료들이 웹 사이언스의 목표와 웹의 상호 작용 및 컴퓨터 정보 과학간의 관계에 대한 상호 작용에 대해 이야기해준데 감사한다. 특히 우리는 Konstantin Mertsalov(Rensselaer Polytechnic Institute)의 DBpedia 분석 결과에 많은 영감을 받았다.

## References

1. Abadi, D., Marcus, A., Madden, S., and Hollenbach, K. Scalable semantic Web data management using vertical partitioning. In *Proceedings of the 33rd International Conference on Very Large Data Bases (Vienna, Austria, Sept. 23–27)*. VLDB Endowment, Heidelberg, 2007.
2. Backstron, L., Dwork, C., and Kleinberg, J. Wherefore art thou R3579X?

Anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of the 16<sup>th</sup> International World Wide Web Conference* (Banff, Alberta, Canada, May 8–12). ACM Press, New York, 2007.

3. Barabasi, A. and Albert, A. Emergence of scaling in random networks. *Science* 286 (1999).

4. Berners-Lee, T., Connolly, D., Kagal, L., Scharf, Y., and Hendler, J. N3Logic: A logical framework for the World Wide Web. *Theory and Practice of Logic Programming* (2008).

5. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., and Wietzner, D. Creating a science of the Web. *Science* 311 (2006).

6. Berners-Lee, T., Hall, W., Hendler, J., O'Hara, K., Shadbolt, N., and Weitzner, D. A framework for Web science. *Foundations and Trends in Web Science* 1, 1 (Sept. 2006).

7. Berners-Lee, T., Hendler, J., and Lassila, O. The semantic Web. *Scientific American* (May 2001).

8. Berners-Lee, T. and Fischetti, M. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Harper Collins, New York, 1999.

9. Brin, S. and Page, L. The anatomy of large-scale hypertextual Web search engine. Presented at the Sixth International World Wide Web Conference (Santa Clara, CA, Apr. 7–11, 1997).

10. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. Graph structure in the Web. In *Proceedings of the Ninth International World Wide Web Conference* (Amsterdam, The Netherlands, May 15–19). Elsevier, Amsterdam, The Netherlands, 2000.

11. Dean, J. and Ghemawat, S. MapReduce: Simplified data processing on large clusters. In *Proceedings of the Sixth Symposium on Operating System Design and Implementation* (San Francisco, Dec. 6–8). USENIX Association, Berkeley, CA, 2004.

12. Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D. and Tomkins, A. Self-similarity in the Web. In *Proceedings of the 27<sup>th</sup> International Conference on Very Large Data Bases* (Rome, Italy, Sept. 11–14). Morgan Kaufmann Publishers, Inc., San Francisco, 2001.

13. Domingos, P., Golbeck, J., Mika, P., and Nowak, A. Social networks and intelligent systems. *IEEE Intelligent Systems, Trends & Controversies* 20, 1 (Jan./Feb. 2005).

14. Donato, D., Laura, L., Leonardi, S., and Millozzi, S. The Web as a graph: How far we are. *ACM Transactions on Internet Technology* 7, 1 (Feb. 2007).

15. Fokoue, A., Kershenbaum, A., Ma, L., Schonberg, E., and Srinivas, K. The Summary Abox: Cutting ontologies down to size. In *Proceedings of the International Semantic Web Conference* (Athens, GA, Nov. 5–9). Springer Berlin, Heidelberg, 2006.

16. Golder, S. and Huberman, B. *The Structure of Collaborative Tagging Systems* (2005); arxiv.org/abs/cs/0508082.

17. Gulli, A. and Signorini, A. The indexable Web is more than 11.5 billion pages. In the special-interest tracks and posters of the 14<sup>th</sup> International World Wide Web Conference (Chiba, Japan, May 10–14). ACM Press, New York, 2005.

18. Hendler, J. Web 3.0: Semantic Web chicken farms. *IEEE Computer* 41, 1 (Jan. 2008).

19. Hendler, J. and Golbeck, J. Metcalfe's Law, Web 2.0, and the semantic Web. *Journal of Web Semantics* 6, 1 (Feb. 2008).

20. Huberman, B. and Lukose, R. Social dilemmas and Internet congestion. *Science* 277, 5325 (July 1997).

21. Jacobs, I. and Walsh, N. *Architecture of the World Wide Web, Vol. One*. W3C Recommendation, Dec. 15, 2004; www.w3.org/TR/webarch/.

22. Kleinberg, J., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. The Web as a graph: Measurements, models, and methods. In *Proceedings of the Fifth*

*Annual International Conference on Computing and Combinatorics* (Tokyo, July 26–28). Springer, New York, 1999.

23. Kleinberg, J. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46, 5 (Sept. 1997).

24. Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. Trawling the Web for emerging cyber communities. In *Proceedings of the Eighth International World Wide Web Conference* (Toronto, May 11–14). Elsevier North-Holland, Inc., New York, 1999.

25. Manber, U. *Why Search Is a Hard Problem*. Presentation at Supernova 2007 (San Francisco, June 16–18, 2008); www.readwriteweb.com/archives/udi\_manber\_search\_is\_a\_hard\_problem.php

26. Marcus, A. and Perez, A. m-YouTube mobile UI: Video selection based on social influence. In *Proceedings of the 12<sup>th</sup> International HCI Conference* (Beijing, July 22–27). Springer, 2007.

27. Samuelson, P. Copyright's fair use doctrine and digital data. *Commun. ACM* 37, 1 (Jan. 1994), 21–27.

28. Shadbolt, N., Hall, W., and Berners-Lee, T. The semantic Web revisited. *IEEE Intelligent Systems* 21, 3 (May/June 2006).

29. Shirky, C. *Power Laws, Weblogs, and Inequality* In Clay Shirky's blog (2003); www.shirky.com/writings/powerlaw\_weblog.html.

30. Shneiderman, B. Web science: A provocative invitation to computer science. *Commun. ACM* 50, 6 (June 2007), 25–27.

31. Weitzner, D., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J., and Sussman, G. Information accountability. *Commun. ACM* 51, 6 (June 2008).

32. Weitzner, D., Hendler, J., Berners-Lee, T., and Connolly, D. Creating a policy-aware Web: Discretionary, rule-based access for the World Wide Web. In *Web and Information Security*, E. Ferrari and B. Thuraisingham, Eds. IRM Press, Hershey, PA, 2006.

## 주의 사항

이 논문은 Web science: an interdisciplinary approach to understanding the web 의 한국어 번역판으로 영어 원본에는 없는 번역 상 오류가 들어 있을 수 있습니다. 오직 영어 원본만이 효력을 지님에 유의하시기 바랍니다. 번역 상 오류에 대한 지적이나 제안하실 내용이 있으면 연락해주시고요. (번역자 연락처: 윤석찬, channy@snu.ac.kr)

## 영문 논문의 온라인 공개 위치

[http://portal.acm.org/ft\\_gateway.cfm?id=1364798&type=html&coll=ACM&dl=ACM&CFID=76510112&CFTOKEN=19304047](http://portal.acm.org/ft_gateway.cfm?id=1364798&type=html&coll=ACM&dl=ACM&CFID=76510112&CFTOKEN=19304047)

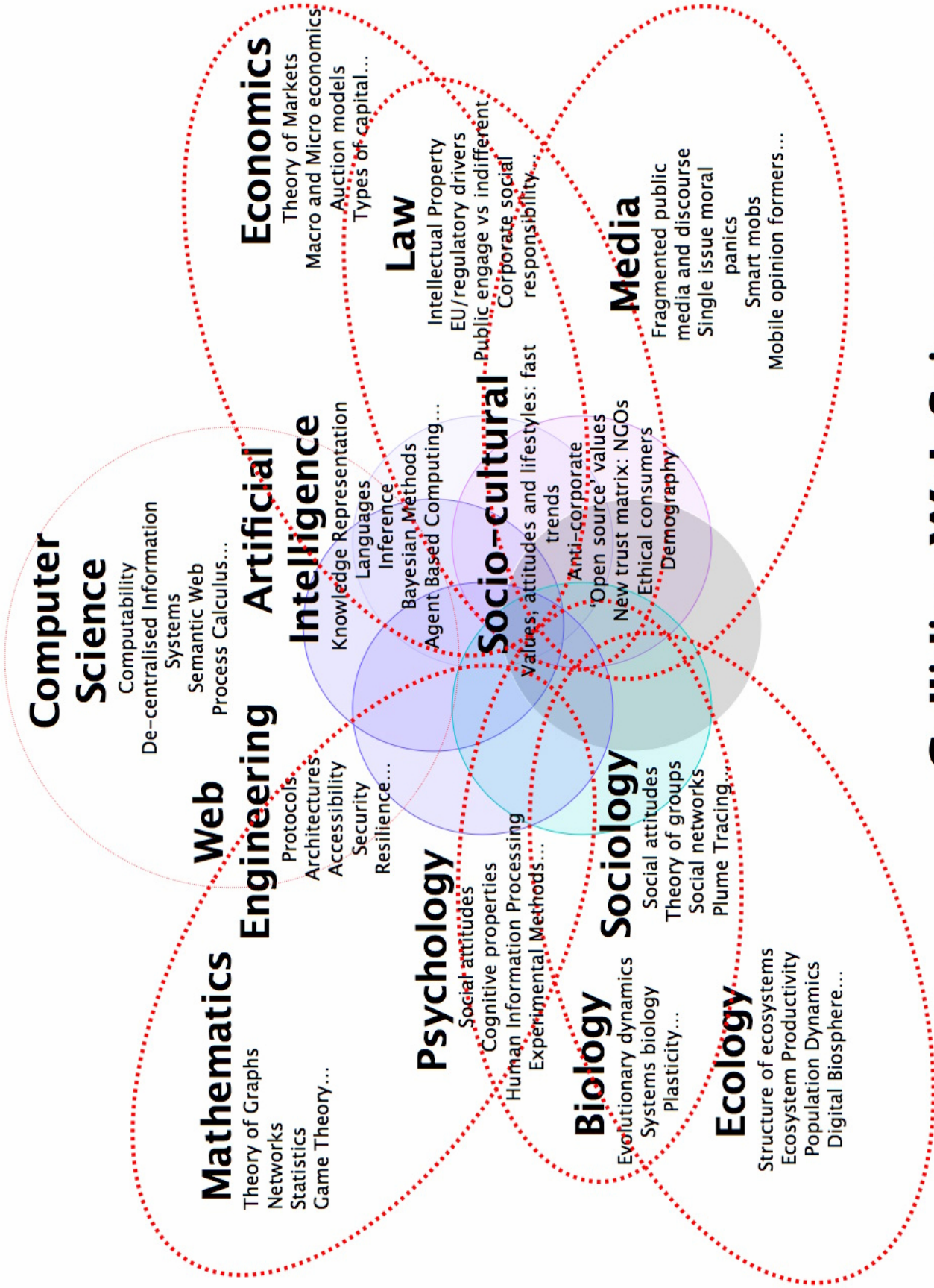
## 서울대학교 의생명 지식공학 연구실

Biomedical Knowledge Engineering Laboratory

<http://bike.snu.ac.kr>

### □ 주요 연구 분야

- 의료 정보 분야 시맨틱 웹
- 온톨로지 공학
- 지식 기반 데이터 마이닝
- 감성 컴퓨팅 및 웹 사이언스



# Colliding Web Sciences