



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이학석사 학위논문

구글 검색엔진의 성장에 대한 ANT적 분석

2016년 8월

서울대학교 대학원

과학사 및 과학철학 협동과정

장 중 혁

구글 검색엔진의 성장에 대한 ANT적 분석

지도교수 홍 성 욱

이 논문을 이학석사학위논문으로 제출함

2016년 6월

서울대학교 대학원

과학사 및 과학철학 협동과정

장 중 혁

장중혁의 석사학위논문을 인준함

2016년 7월

위 원 장 장 대 익 (인)

부 위 원 장 홍 성 욱 (인)

위 원 이 준 석 (인)

국문초록

이 논문은 1998년 구글 검색엔진의 탄생부터 약 10년에 걸친 검색엔진 주도의 인터넷 성장기에 일어난 행위자네트워크의 변화과정을 살펴보았다. 이미 야후!라는 포털 사이트가 강하게 자리잡고 있었던 당시 다른 검색엔진들의 실패 속에서 구글이 성공할 수 있었던 원인은 무엇일까? ‘구글의 혁신적 기술이 성공의 원인’이라는 기술결정론의 신화는 얼마나 믿을만한 것인가? 이 논문에서는 이 문제를 행위자네트워크이론의 이론적 도구를 활용하여, 그 기간 동안 빠르게 증식된 행위자들인 하이퍼미디어의 독자, 하이퍼텍스트의 저자, 하이퍼텍스트의 독자라는 세 종류의 하이브리드에 대한 미시적 분석으로부터 접근하여 해명하고자 시도하였다. 먼저, 이 논문은 웹의 성장과정에서 증식된 주요 행위자들을 계보학적으로 추적하여, 이들이 각자 어떤 기원으로부터 등장하여 여러 번의 하이브리드화 과정을 거쳐 웹의 성장에 중심적 행위자가 될 수 있었는가를 살펴보았다. 또한 구글 검색엔진이 성장하기 직전에 웹의 행위자들에 가장 큰 영향력을 행사하고 있었던 야후! 디렉토리가 이들 행위자와 어떻게 동맹을 맺으며 영향력을 확대하였으며, 동시에 어떤 동맹에 실패함으로써 위기가 시작되었는지도 살펴 보았다. 이어서 구글 검색엔진의 핵심기술인 ‘페이지랭크’가 고안되어 ‘검색 기술’로 정착되는 과정에 대한 미시적 분

석을 통해, 이것이 단순한 기술혁신의 결과가 아니라 웹 내의 하이브리드들의 증식에 의해 구성된 결과라는 사실을 보이고, 구글이 이 성공을 지속하고 확대하기 위해 어떻게 검색 품질을 희생하는 사회적 선택을 하였는지도 살펴 보았다. 끝으로 구글이 초기의 기술적 특성 때문에 우연히 채택하였고, 웹 스팸이라는 공격을 막아내는 과정을 계기로 발전시킨 N-gram이라는 기술이 어떻게 ‘페이지랭크’를 넘어 구글 검색엔진의 영향력을 확장시키는 결정적 기술이 되었는가를 제시하였다. 구글은 이러한 역사적 과정을 통해 구축한 영향력으로 인터넷 안에서뿐 아니라 실물 세계에 까지 영향력을 확대하고 있으며, 이는 근대적 세계의 핵심적 행위자인 국가를 넘어서 초국가적 행위자가 될 가능성을 보여주고 있다.

주요어 : 구글, 행위자네트워크이론, 하이브리드, 인터넷, 검색엔진

학 번 : 97321-503

목 차

1. 서론.....	7
2. 이론적 배경과 선행 연구.....	17
2.1 인터넷 역사학의 방법론적 논점.....	17
2.2 행위자네트워크이론의 자원들.....	20
2.2.1 번역(Translation).....	21
2.2.2 두 개의 이질성, 하이브리드(Hybrid)와 이종성(Heterogeneity) ...	24
2.2.3 ANT의 연구 방법론.....	30
3. 구글동맹에 참여한 하이브리드들의 계보와 기원.....	32
3.1 하이퍼링크와 구글 검색엔진의 ‘페이지랭크’(PageRank).....	32
3.2 ‘인간-하이퍼링크’ 하이브리드의 두 형태.....	35
3.3 ‘하이퍼미디어’로의 우회 : ‘인간-하이퍼링크’ 하이브리드의 증식....	40
4. ‘번역’으로서의 배제 : 야후! 위기의 출발점.....	46
4.1 야후는 어떻게 웹을 번역하고자 했는가?.....	46
4.2 하이퍼미디어의 독자를 흉내 내는 인간 vs. 텍스트 저자를 흉내 내는 검색엔진.....	52
4.3 새로운 유형의 검색엔진들과 동맹 대상의 재구성.....	58
4.4 ‘검색기술’은 누구를 대표하여야 하는가?.....	60
5. ‘페이지랭크’는 어떻게 검색엔진이 되었나.....	66
5.1 ‘색인(index)’에 대한 새로운 정의, ‘독자의 문서’를 색인하다.....	67

5.2 기술로서의 '페이지랭크' 탄생의 전사(前史) : 대용량 문서 처리 시스템 COPS.....	72
5.3 '페이지랭크' 행위자네트워크의 탄생	76
5.4 백업 크롤러, '검색엔진'이 되다	83
6. 새로운 하이브리드의 증식 : 구글의 위기와 새로운 번역의 전략.....	92
6.1 N-gram : 블랙박스 안의 행위자.....	93
6.2 구글을 둘러싼 새로운 하이브리드의 증식과 위기.....	99
6.3 구글의 새로운 번역 전략 : N-gram 필터.....	101
6.4 새로운 하이브리드를 위한 대표 선출	104
6.5 보일의 실험실과 진공펌프로서의 인터넷.....	106
7. 결론 : 구글 동맹은 어떻게 공적 영역을 사유화하였는가.....	113
참고문헌.....	119
Abstract.....	131

그림목차

그림 1 Hybrid의 시야.....	26
그림 2 Heterogeneity의 시야	28
그림 3 제너두 프로젝트의 문서 편집기(Hypertext Editing System).....	39
그림 4 월드와이드웹 사이트 증가 통계	50
그림 5 TREC과 실험 평가 대상의 차이.....	64
그림 6 N-gram 청크 개념도	75
그림 7 해쉬 청크 개념도	75
그림 8 초기 페이지랭크의 행위자네트워크 확장	77
그림 9 페이지랭크가 '백업 크롤러'로 결절화된 행위자네트워크	79
그림 10 OPP로서의 백업 시스템.....	82
그림 11 전문 검색엔진(Full Text Search Engine)으로서 구글의 랭크병합.....	89

1. 서론

웹은 1993년 최초의 일반 사용자용 웹 브라우저인 모자이크(Mosaic)가 보급된 이후 20여 년 만에 우리의 일상 속에 완벽하게 자리잡은 거대한 기술시스템이다. 기술사학자 자넷 어배티(Janet Abbate)는 미 국방부의 아르파넷(ARPANET)에서 출발한 인터넷이 확장되어 대학과 군사 연구소의 컴퓨터를 연결했던 과정이 국방부의 이해관계가 관철된 사회적 과정이라는 것을 잘 보여주었다.(Abbate, 1999) 그러나 어배티의 연구를 포함한 인터넷 역사에 대한 많은 연구는 TCP/IP 프로토콜이 정립되고 인터넷이 개방되는 시점까지의 초기 역사를 주로 다루고 있다. 따라서 웹의 등장 이후에 본격적으로 대중화된 인터넷이 어떻게 성장하고 변화하였는가를 알기 위해서는 1990년대 이후의 인터넷의 역사를 연구해야 한다. 하지만 이런 시도는 바로 인터넷의 ‘비가시성’이라는 문제에 직면한다.

투명하고 개방되어 있다는 믿음과 달리 인터넷은 몇 가지 이유에서 비가시적이다. 우선 인터넷은 기술적 요소로 구성되어 있기 때문에, 일반인들이 이해하기 쉽지 않다. 게다가 이런 장벽을 넘어서더라도 인터넷은 가시적이지 않다. 그 이유는 인터넷을 구성하는 요소들은 우리가 종전에 경험해 보지 못한 이종적(heterogeneous) 거버넌스로 얹혀 있기 때문이다.

인터넷의 거버넌스는 기업의 내부 네트워크나 인터넷 뱅킹의 계정 시스템과 같이 극단적으로 사유화된 영역에서부터, 개방되어 있지만 독점적 거버넌스로 운영되는 도메인네임 등록과 같은 위계적(hierarchical) 개방형 거버넌스 영역, 인터넷의 개방성을 상징하는 기술 요소인 IP 네트워크의 표준규격과 같은 비위계적(non-hierarchical) 거버넌스 영역 등에 넓게 걸쳐있다. 인터넷의 탄생 초기에 개방적인 거버넌스 영역이 중심을 이루었던 것과는 달리, 현재의 인터넷은 대부분 폐쇄적 거버넌스에 의해 지배되는 ‘사유화된 영역(privatized domain)’에 속해있다. 인터넷은 많은 사람들의 기대와 달리 매우 제한적 영역에서만 개방된 네트워크이고, 많은 부분이 불투명한 사유화된 거버넌스에 의해 지배되는 네트워크다.

그러나 인터넷 안에서의 사유화된 영역의 성장은 두 가지 점에서 위험하다. 하나는 그것이 사적 영역의 팽창에 의한 공적 영역 파괴로 귀결된 근대의 다른 위기들과 비슷한 양상을 보이고 있다는 점이다. 근대에 대한 비판적 연구자들은 사적 영역이 자연과 공적 영역을 향해서 확장되면서 사회의 위기를 낳고 있다고 지적한다.(Arendt, 1958; 이홍균, 2006) 이러한 경고는 인터넷에 대해서도 유효한데, 미디어 연구자인 시바 바이디야나단은 저서 <모든 것의 구글화(Googlization of Everything)>에서 구글이 전인류적 가치를 사유화된 영역 안으로 끌어들여 독점하는 권력작용을 만들어내고 있다고 경고한다.(Vaidhyathan, 2012) 무엇보다 20세기

후반 이후의 사적 영역 팽창은 사유화라는 메커니즘을 중심으로 한 신자유주의적 팽창이라는 점에서 개인들에 대해서는 재봉건화(refeudalization)와 같은 위험을 수반한다.(Habermas, 1991) 다른 하나의 위험은 실물세계에서와 달리 인터넷 안에서의 사적 영역 팽창은 앞서 기술한 이유로 가시적이지 않다는 것이다. 이는 사적 영역의 팽창으로부터 공적 영역을 방어하려는 노력이 대응할 시기와 대상을 놓치게 만든다. 따라서 현 시점에서 이러한 문제에 대응하는 가장 우선적 조치는, 인터넷 안에서 사적 영역의 팽창이 ‘어떻게’ 이루어지고 있는가를 가시화시키는 것이다.

본 연구는 인터넷이 대중적 성장이 이루어진 기간 중에서도 1997년부터 2006년까지 인터넷을 급격하게 변화시킨 구글 검색의 기술적 요소에 주목한다. 본 연구에서 대상으로 한 기간의 선정이 인터넷의 역사를 기술하는데 일반적으로 적용할 수 있는 시기 구분(periodization)을 시도하는 것은 아니지만, 이 기간은 인터넷의 역사에서 ‘검색엔진이 지배한 10년’(Decade of Search Engine)이라고 불릴 수 있는 시기다. 이 시기에 인터넷은 15배 이상 성장하여 7,000만 명 수준이었던 인터넷 사용자는 10억 8,000만 명 수준으로 급증했다. 본 연구의 핵심적 주장은 이 시기 인터넷의 성장이 ‘글로벌화된 인터넷의 사유화’와 관련되어 있다는 것이다. 물론 그 전에도 인터넷 안에는 ‘야후!’(Yahoo!) 디렉토리와 같은 사유화된 허브(hub)들이 있었지만, 그러한 허브들은 국가 단위로 분산되어 있었

다. 그러나 ‘검색엔진이 지배한 10년’이 지났을 때, 전세계의 웹 노드들은 몇몇 예외적 국가들을 제외하면 구글이라는 하나의 사유화된 허브로 집중된 구조로 연결되었다. 2007년 전세계 웹 검색엔진 시장에서 구글의 점유율은 80%를 넘어서면서,(Stoffe, 2013) 아시아 일부 국가를 제외하고는 사실상 모든 웹 사용자는 구글엔진이 보여주는 검색결과에 의존하지 않고는 웹 내에서 자신이 찾는 문서에 도달할 방법이 없게 되었는데(집중화), 이는 ‘검색 결과로 무엇을 보여줄 것인가’를 공적 통제 없이 결정하는 것이 그 자체로 인터넷 안에서 웹 사용자가 어떤 문서에 도달하게 될 것인가를 ‘결정하는’ 권력을 사유화하는 것을 의미하며, 이는 웹이 인터넷에서 차지하는 비중을 고려하면 사실상 인터넷을 사유화한 것과 같은 의미였다.

본 연구는 인터넷의 이러한 집중화와 성장을 가시화시켜, 그것이 구글 검색엔진 기술을 성공시킨 미시적 사유화의 ‘원인’인 동시에 그러한 사유화의 거시적 확장의 ‘결과’임을 보일 것이다. 이 주장의 실천적 함의는, 인터넷 안에서 이루어지는 사적 영역의 미시적인 기술 혁신이 실물 세계에서 공적 영역과 결합되어 있던 인간이나 사물을 그로부터 이탈시키거나 결합을 느슨하게 만듦으로써 공적 영역의 파괴 내지는 축소를 매개할 수 있다는 것이다. 물론 이것이 인터넷 안에서의 공적 영역의 성장을 배제하거나 실물세계에서의 사적 영역 팽창이라는 신자유주의적 변화

를 부정하는 이분법은 아니다. 그럼에도 인터넷의 성장은 실물세계에서 형성된 사적 영역 팽창에 의한 공적 영역 파괴의 글로벌화된 메커니즘에 대한 저항을 우회하여 이것이 강화될 수 있는 환경을 제공한다는 것이다.

본 연구가 구글을 선택한 이유는 구글이 전세계적으로 가장 지배력이 큰 검색엔진이라는 점도 있지만 다른 이유도 있다. 그 중 하나는 구글은 구글 이전에 인터넷의 성장을 이끌어 왔던 ‘사유화된 허브’인 야후! 디렉토리와 비교 가능하기 때문이다. 이러한 비교는 구글 검색엔진이 주도한 성장이 어떻게 야후!와 같은 다른 종류의 허브가 주도한 성장과 달리 매우 짧은 기간 안에 글로벌화된 인터넷의 사유화를 매개할 수 있었는지를 보여줄 수 있다는 점에서 본 연구의 목적에 잘 부합한다. 또한 구글은 연구 대상으로서, ‘기술적 혁신’에 의해 성공한 대표적 사례로 꼽혀 구글의 검색 기술에 대한 정보는 학술 논문에서 미디어의 인터뷰까지 매우 풍부하다는 것과 구글 스스로가 구글 내에서 진행된 연구결과를 발표한 학술 논문들과 검색엔진의 변경 사항에 대한 정보를 상세하게 제공하고 있기 때문에, 기술적 세부사항을 분석해서 사유화된 영역 내부를 들여다 보기에 적당하다는 장점도 가지고 있다.¹

¹ 본 연구에서 인용되거나 언급되는 구글의 기술연구 내용에 대한 주요 정보는, 구글의 창업자 또는 연구자가 직접 참여하여 발표된 학술 논문과 구글 검색에 관한 공식 블로그 역할을 하고 있는 사이트의 게시물을 대상으로 하였다. 그 정보의 상당 부분은 “Research at Google”이라는 명칭의 구글 공식 사이트(research.google.com)와 구글의 검색 기술 책임자인 매트 컷(Matt Cuts)의 블로그인 mattcutts.com을 통해 획득되었다.

본 연구는 이를 위해 과학기술학의 ‘행위자 연결망 이론’ (Actor-Network Theory, 이하 ANT로 약칭)이 주목하는 인간과 비인간 행위자가 결합되어 새로운 행위능력을 갖게 된 ‘하이브리드’ (hybrid)의 증식에 초점을 맞춘 분석을 수행할 것이다. 이러한 접근법은 인터넷의 성장이 대중적 단계에 들어선 1990년대 이후 인터넷의 변화를 기술하는데 매우 유효한데, 그 이유는 이 시기 인터넷의 성장과 변화를 어떤 주체의 의지나 행위로 기술하거나 네트워크나 연결된 컴퓨터의 숫자로 현상적으로 기술하는 것의 결정론적 단순화로 인한 한계를 극복할 수 있기 때문이다. 인터넷과 같이 개방된 거대 기술 네트워크의 성장과 변화를 기술하는데 있어, 주체를 중심으로 한 기술은 변화 계기의 다양성과 우발성을 포괄하기 어렵고 현상적 기술은 객관적이지만 그것이 ‘어떻게’ 일어난 일인지에 대한 이해를 제공하는데 한계를 가지고 있다. 그런 점에서 웹을 중심으로 ‘어떤’ 유형의 행위자가 성장과 변화의 매개로서 ‘어떻게’ 증식하였는가를 기술하는 것은 이러한 한계를 극복하는 유효한 전략이 된다.

본 연구는 웹의 전사(前史)와 발전과정 안에서 두드러진 증식 패턴이 형성되었던 인간-비인간 하이브리드로서 세가지 유형의 ‘인간-하이퍼링크’ 하이브리드에 주목한다. 이들은 각자의 발생 계보로부터 온 것인데, 전통적 ‘텍스트의 저자’에서 내려온 ‘하이퍼텍스트 저자’, ‘텍스트의 독자’에서 비롯된 ‘하이퍼텍스트 독자’, ‘미디어의 독자’가 하이

퍼미디어의 행위자네트워크와 결합되어 탄생한 ‘하이퍼미디어 독자’가 그것이다. 이러한 계보학적 접근은 하이퍼텍스트라는 새로운 의미-권력 네트워크 안의 매개(‘주체’가 아니라)들을 위한 것인데, 이는 ANT가 행위자를 ‘본질’로 환원시키지 않고 구성적으로만 다루려는 입장과 잘 부합한다.(미셸 푸코, 2003) 이 세가지 유형의 ‘인간-하이퍼링크’ 하이브리드들은 웹의 탄생 전에 발생하여 이후 인터넷 성장의 각 단계에서 급속한 증식의 패턴을 형성했던 대표적 ‘인간-하이퍼링크’ 하이브리드들이다. 이들에 초점을 맞추는 이유는, 그것이 흔히 웹의 역사를 몇몇 혁신적 서비스의 성공으로 설명하는 변형된 기술결정론을 넘어, 오히려 하이브리드의 증식이 초래한 웹의 변화가 어떻게 그런 서비스들을 성공시켰는가를 설명함으로써 웹을 탈역사적인 신화가 아닌 역사적 존재로 볼 수 있게 해주기 때문이다.

웹의 초기 야후!는 하이퍼미디어의 이용경험을 가진 독자들과의 동맹을 확대하면서 ‘웹의 하이퍼미디어’로서 인터넷의 성장을 주도했다. 그러나 야후!는 웹의 성장 속에서, 자신에게 위기를 가져온 자신이 연결하지 않는 웹 페이지라는 하이브리드들의 증식을 외면함으로써 웹에서 소수의 링크를 찾는 ‘하이퍼미디어의 독자’들이 받아들일 수 있는 질서를 제공하였지만 동시에 스스로를 웹의 매우 작은 부분만을 대표하는 위기로 몰아 넣었다. 야후!의 대안이 되려 했던 검색엔진 역시 통제되지 않

은 상업적 하이퍼텍스트의 저자와 같은 하이브리드의 증식을 감당해 내지 못했다. 하지만 구글 검색기술에게 웹 내의 이러한 하이브리드의 증식은 중요한 성공의 계기가 되었다. 구글의 핵심 기술인 ‘페이지랭크’의 성공은 그것이 탈역사적인 ‘혁신적 기술’ 이어서가 아니라 당시 빠르게 증식하는 다양한 형태의 하이퍼텍스트의 저자나 독자와 같은 하이브리드들과의 동맹에 성공했기 때문이었다. 그러나 기술의 성공에 힘입어 기술을 매각하려 했던 발명자들의 목표는 실패했다. ‘페이지랭크’는 스스로를 유지하기 위해서는 앞으로 더 나아감으로써 상용 검색엔진이 되어야 했다.

상용 검색엔진이 되기 위해 구글은 이중적 요소들과 하이브리드화되어야 했다. 기술적으로는 본문 색인에 의한 전문검색(全文檢索) 랭킹을 링크 기반의 ‘페이지랭크’와 병합하는 것이었는데, 이는 결과적으로 ‘페이지랭크’의 비용효율성과 검색품질을 희생하는 것이었다. 하지만 이를 통해 구글은 상용 검색엔진과 관련된 행위자들로부터 검색엔진으로 인정되었을 뿐 아니라 향후 ‘페이지랭크’를 대신하여 구글이 맞이한 위기를 해소하고 구글의 사적 권력을 다른 영역으로 확장시키는 새로운 기술적 매개도 획득했다.

구글의 위기는 다른 검색엔진과 마찬가지로 ‘페이지랭크’를 노린 역공학을 동원한 스팸머들의 공격에서 시작되었다. 그러나 구글은 스팸머

들을 동맹에서 제외하여 포털이 되기로 전략을 바꾼 다른 검색엔진들이들에 대한 ‘배제(exclusion)’에 실패했던 것과 달리 블랙박스화된 ‘페이지랭크’의 중심을 새로운 기술로 이동시켜 자신의 질서를 위협하는 스팸머들에 대한 ‘배제’에 성공함으로써 ‘동맹’에 성공했다. 이러한 성공은 구글이 저작권이나 포르노에 관련된 통제, 악성코드의 차단, 정치적 검열과 같이 구글의 질서를 위협하는 웹 내의 다른 권력작용을 향하여 확장하는 출발점이 되었다.

본 연구는 이와 같이 1990년대 이후 인터넷의 역사를 하이브리드들에 대한 ‘동맹’의 성공과 실패를 계기로 한 이들의 증식과 사멸의 역사로 봄으로써, 하이브리드들과의 ‘동맹’에 성공한 사적 영역의 행위자네트워크가 어떻게 성장과 사유화를 통해 실물 세계에 형성되어 재봉건화(refeudalization)를 방어하는 기제로 작동 중인 공적 장치들을 우회하여 재봉건화를 작동시킴으로써 결과적으로 공적 영역의 축소를 초래하는가를 보일 것이다.

이 글은 크게 세 개의 부분으로 구성된다. 그 첫 부분인 2절에서는 인터넷의 역사 연구 방법론에 대한 선행연구를 살펴보고, ANT가 어떻게 인터넷의 역사 연구에 적용될 수 있는지를 살펴볼 것이다. 두 번째로 3절에서 6절에 이르는 부분에서는 ANT를 웹의 역사 연구에 실제로 적용할 것이다. 3절에서는 ANT의 이론적 접근법을 활용하여 웹의 성장을 이끈 행

위자네트워크들이 ‘동맹’의 대상으로 삼았던 존재들의 계보학적 기원을 추적할 것이다. 이들의 계보에 대한 연구는 인터넷의 역사적 구획마다 주도적인 역할을 한 행위자네트워크가 어떻게 각자의 대상들과의 동맹에 성공하고 실패했는가를 연구하는데 중요한 출발점이 된다. 그리고 4절에서는 구글 등장 전에 웹의 성장을 이끌었던 야후!의 성공과 그 성공에 개입했던 행위자들의 역할을 살피고, 야후!가 택한 번역의 전략이 어떻게 성공과 함께 위기를 초래하는가를 살펴볼 것이다. 5절에서는 구글이 어떻게 전통적 의미의 ‘검색엔진’이 아니었음에도 어떻게 주변 행위자들의 목표를 변경시키며 ‘검색엔진’이 될 수 있었는가를 설명하면서, 구글이 어떻게 다른 ‘검색엔진’과는 다른 동맹을 통해 의무통과점(Obligatory Passage Point: 이하 OPP)을 구축하였고 그러한 동맹의 매개로서 ‘페이랭크’라는 기술이 어떻게 작동하였는가를 보일 것이다. 6절에서는 ‘검색엔진’으로서의 OPP 구축에 성공한 구글이 블랙박스화된 ‘페이랭크’안쪽에서 기술의 하이브리드화를 통해 스스로의 목표를 변형시켜 어떻게 ‘구글화’(Googlization)의 확장을 이루었는지를 살펴볼 것이다. 마지막 7절에서는 본론의 논의를 요약하면서 인터넷 안에서의 사적 영역 팽창과 관련된 향후의 연구 과제들을 제시할 것이다.

2. 이론적 배경과 선행 연구

본 연구는 ANT의 이론적 자원들을 이용한 인터넷 역사 연구를 통해, 인터넷의 성장과 변화의 질서에 대한 규명과 그것의 실천적 함의를 살피고자 한다. 이를 위해 인터넷 역사 연구에 대한 이전 논의를 간략하게 살펴봄으로써 인터넷 역사연구의 현황을 파악하고, 이를 통해 ‘기술연구를 통한 인터넷 역사연구’가 해결해야 하는 이론적 장애물에 대한 본 연구의 접근법과 이에 사용하려는 ANT의 이론적 자원들을 살필 것이다.

2.1 인터넷 역사학의 방법론적 논점

이광석은 “인터넷 역사 방법론의 논점” (2015)에서 인터넷의 역사를 사회문화사로 보는 관점과 제도사로 보는 접근의 한계를 지적하면서, 인터넷 역사 서술에서 공백으로 남겨져 있던 ‘기술로서의 인터넷에 대한 존재론적 해명’과 ‘국내 인터넷의 특수한 진화과정에서 나타난 소수 기술 주체들에 대한 기술문화사적 연구’를 대안으로 제시한다. 인터넷에 대한 존재론적 해명을 위해 이광석은 미디어철학자인 Kittler(Friedrich Kittler)의 매체유물론적 관점과 시몽돈(Gilbert Simondon)의 기술철학, ‘기술에 대한 사회구성주의적 연구’를 결합시켜 이전 인터넷 역사서술

이 가진 ‘사회결정론’적 경향을 극복하는 구성주의적 존재론을 시도한다. 그의 연구는 그간 한국의 인터넷 역사 연구가 극히 일부를 제외하고는 정부가 주도한 제도사적 관점에서 서술되었다는 점과 미디어의 사회문화사로서의 인터넷의 역사에 접근할 때 기술의 역사로서의 측면을 놓치고 있다는 비판에서 출발된 것인데, 이것이 매우 적절한 비판임에도 그의 연구가 어떻게 그러한 접근들의 대안이 될 수 있는지를 보여주지 못하고 있다. 이광석이 비판하는 두 접근은 그것이 옳은가의 문제와는 별개로 ‘인터넷의 성장이 어떻게 이루어졌는가?’에 대한 구체적 설명을 제공하는 반면, 그가 말하는 ‘기술계보학’²과 ‘구성주의적 공진화’는 추상적 논의에 머물므로써 자신이 말하는 ‘기술의 우선성’이 어떻게 혼한 ‘기술결정론’을 피하면서 인터넷의 성장에 대한 구체적 설명을 제공할 수 있는지를 보여주지 못하고 있다.

인터넷의 역사학이 기술계보학을 방법적으로 채택한다는 것은 인터넷 안에서 미시적이든 거시적이든 이를 이용하여 권력이 어떻게 탄생하고 투쟁하고 소멸하는가를 해명하기 위한 것이다. 따라서 이광석이 이러한

² 푸코(Michel Foucault)에 있어 ‘계보학’은 전통적인 역사학의 틀을 거부하는 일종의 반역사로서, 주체가 특정한 목적의식을 가지고 역사를 창조하며 조율해 나간다는 식의 전통적 역사철학과 총체성의 이념을 거부한다.(윤평중 1996) 이광석이 ‘소수성’에 주목하는 것 역시 ‘계보학’이 역사 분석에서 일반적으로 취하는 입장인데, 이광석은 이를 시몽동의 논의로부터 가져오고 있다.

방법론적 선택을 정당화하기 위해서는 기술계보학이 인터넷의 역사 연구에 구체적으로 어떻게 적용되는가를 보여야 한다. 인터넷 역사학의 연구 방법으로서의 기술계보학이 정당화될 수 있는 길은 둘 중 하나인데, ‘인터넷의 성장과 변화의 질서’에 대한 담론을 사회문화사와 같은 다른 연구 방법에 맡기고 기술계보학 방법론의 적용 범위를 인터넷에 대한 미시적 기술사로 제한하는 길을 가거나, 기술계보학 연구를 통해 인터넷이 가진 인공물로서의 자율성과 비인간행위자들과 결합된 인간행위자들의 행위능력이 어떻게 매개(mediation)를 통해 새로운 권력작용을 구성하는가를 보임으로써 ‘인터넷이 어떻게 성장하였는가?’에 대해 설명하는 것이다.

본 연구는 후자의 접근법을 택하여 ANT와 결합된 기술계보학적 방법이 어떻게 인터넷의 역사 연구에 적용될 수 있는지를 보이고, 기술계보학적 연구를 통해 인간과 비인간의 결합으로서의 하이브리드를 식별하여 이들의 증식과 그것으로부터 비롯되는 위기를 기술함으로써, ‘인터넷은 어떻게 성장하는가?’라는 질문에 대한 답변이 ‘인터넷의 성장을 주도한 권력은 어떤 방향으로 자신을 확장하고 있는가?’라는 질문에 대한 답을 동시에 내포하는 인터넷 성장에 대한 설명을 제시할 것이다.

2.2 행위자네트워크이론의 자원들

브루노 라투어와 그 동료 연구자들은 인간과 비인간, 사회와 자연, 주체와 객체, 존재와 인식을 분리하여 어떤 대상에 대한 설명이 가능하다고 보지 않는다. ANT는 그러한 경계를 파괴함으로써 성장하는 네트워크에 대한 더 좋은 기술(description)을 추구한다. ANT가 경계들을 파괴하는 이유는 분명하다. 그것은 그러한 경계 자체가 특정한 행위자네트워크가 성장한 원인이자 결과인 질서(order)를 담고 있으며, 그렇기 때문에 새로운 질서와 함께 성장하는 행위자네트워크는 기존의 경계에 의존하여 설명될 수 없기 때문이다.(홍성욱, 2010: p.31) 새로운 행위자네트워크의 성장이란 기존의 행위자네트워크가 구성하였던 집합들과 그것에 의해 만들어진 경계에 대한 ‘문제시(problematization)’에서 시작되는 새로운 질서다. 그런 이유에서 행위자네트워크 이론은 행위자로 보이는 존재가 사실은 미시적 원인으로부터 발생한 질서에 의해 성장하고 유지되고 있는 네트워크임을 보임으로써(Law, 1992) 대상에 대한 더 좋은 기술(description)을 추구한다.

ANT는 일종의 ‘실물의 기호학’ (Material Semiotics)³을 지향하는데, 이는 대상을 ‘검증된 방법’에 의해 텍스트화한 지식과 지식의 ‘대상’이라고 믿어지는 물질 세계 사이에 ‘지시적’이거나 ‘법칙적’인

³ ‘물질의 기호학’ 보다는 ‘실물의 기호학’이 적절한 번역으로 판단된다.

- 즉, ‘왜’ (Why) 질문에 대한 해명인 - 관계를 전제하지 않고, 기호학적 도구를 통해 ‘그 둘 사이에 텍스트화되지 않은’ 것들로부터 새로운 의미작용과 권력작용이 ‘어떻게’ (how) 발생하는가를 기술하는 이론적 도구를 지향한다는 의미다.(라투어, 2010; Law, 2009) ANT는 이를 위해 기호학과 권력이론으로부터 많은 용어들을 차용하여 사용했는데, ‘행위소’ (actant), ‘기입’ (inscription)이나 ‘번역’ (translation) 같은 기호학의 용어들과 ‘동원’ (mobilization), ‘징병’ (enlist)과 같은 ‘권력 이론’ 혹은 ‘정치 이론’의 용어들이 그것이다.(Law, 1992)

이를 위해 사용될 ANT의 개념들 중에서 주요한 것들은 다음과 같은 것들이다.

2.2.1 번역(Translation)

ANT에서 ‘번역’은 행위자네트워크가 새로운 행위자를 네트워크 안으로 끌어들이고 안착시키는 일련의 과정을 뜻한다. 기호학적 측면에서 보자면 ‘번역’은 ‘내러티브’를 구성하는 것이고, 행위자들에 대한 의미작용이다. 권력 이론의 측면에서는 ‘번역’은 새로운 행위자를 ‘훈육’하여 지배를 내재화시킴으로써 ‘동맹’을 늘리고, 동맹 내의 행위자들에게 새로운 행위자를 동맹 안으로 끌어들이는 역할을 부여하여 동맹을 확산시키는 전략적 실행이다. 이러한 동맹 확산에서 행위자네트워크는

끌어들이려는 행위자들을 위한 진입점을 설치하는데, 그 진입점은 각 행위자들에게는 자신들의 실체적 목표를 실현하는데 효과적 경로에 대한 제안인 동시에 행위자네트워크로서는 이를 통해 동맹을 확산시킬 수 있다는 가설이다. 그것이 ANT에서 OPP라고 부르는 것이다. 여기서 중요한 사실은 번역하려는 행위자와 번역 대상이 되는 행위자 사이에는 어떤 종류의 법칙적 관계도 전제되지 않는다는 것이다. 오히려 ANT에서 법칙은 이러한 번역이 성공한 예외적이거나 잠정적 결과이며 원인이 아니다.(라투어, 2010) 따라서 ANT의 번역에서 OPP는 때로는 성공하고 주로 실패하는 가설적 양상이다.

번역은 몇 개의 단계적 과정들을 통해 이루어지는데, 번역을 수행하는 행위자네트워크 자체의 이중성 때문에 그 과정들을 추상화된 개념으로 일반화하기는 어렵다. 다만 이를 은유적으로 포괄할 수 있는데, 그것이 ANT에서 ‘문제제기’ (problematization), ‘관심끌기’ (interessement), ‘등록하기’ (enrollment), ‘동원하기’ (mobilization) 정도로 표현되는 과정들이다. 이것을 한 단계 더 세분화하는 것도 가능한데, ‘문제제기’는 ‘행위자들 간의 상호정의’와 ‘OPP 설정’이라는 두 요소로 쪼개지며, ‘관심끌기’는 ‘관심끌기 장치’나 ‘다른 네트워크로부터의 단절’ 같은 요소로 나뉜다.(Callon, 1986) 라투어는 이러한 어휘들을 이중적 행위자들이 네트워크 안에서 벌이는 행위를 기술하기 위한 ‘인프라 언

어’라고 지칭한다.(라투어, 2010) 이는 이 용어들이 다른 이론들에서와 같은 의미로 행위자들의 행위를 ‘인식론적으로 일반화’한 ‘메타언어’가 아니라는 입장을 드러내기 위한 라투어의 용어 선택이다.⁴ ANT의 술어들은 번역하려는 행위자의 시도와 행위를 기술할 뿐 번역 대상에 대해 작동하는 상태에 인과적 설명을 부여하는 것이 아니다.(Latour, 2004)

많은 연구자들은 행위자네트워크 이론의 핵심적 개념을 ‘번역’이라고 이해한다. 그러나 ANT의 ‘번역’을 ANT의 다른 어휘들로부터 따로 떼어서 핵심적 개념이라고 설명하는 것은 적절치 않다. 즉, ANT의 ‘번역’은 미셸 세르가 사용한 ‘번역’에서 기인한 것이 맞다 하더라도 그것과는 다른 의미로 해석되어야 한다. 특히 세르의 번역이 전제로 하는 이종성(heterogeneity)은 인식론적인 것인 반면(Brown, 2002) ANT의 ‘번역’과 연결된 이종성은 존재론적이기 때문이다. 따라서 ANT의 ‘번역’은 ANT가 이야기하고 있는 ‘이종성’과 함께 이해될 때에만 ANT의 그것이 될 수 있다.

행위자네트워크 이론에서 ‘번역’이라는 용어는 이중적 효과를 가지고 있다. 하나는 번역하려는 쪽이 의도하는 질서를 번역 당하는 쪽에 부

⁴ 라투어의 이러한 어휘 선택도 여전히 한계를 가지고 있다. ANT는 모든 종류의 환원이나 토대주의를 거부하는데, ‘인프라 언어’는 상황에 따라서는 모든 기술(description)을 ‘인프라 언어’로 환원시킬 수 있다는 인상을 주기 때문이다.

여한다는 것이고, 다른 하나는 번역하려는 대상이 이미 다른 언어로 작성된 텍스트와 비슷하다는 의미에서 다른 행위자네트워크의 질서 안에서 의미와 권력을 생산하고 있는 상태에서 번역이 시작된다는 것이다. 따라서 행위자네트워크의 소멸 역시 이러한 이중적 연결이 끊어져 나가는 것을 의미한다. 번역은 필연성에 의존하지 않고 ‘보편성(universality)’ – 또는 충만함(fullness) – 을 지향하는 연결이며, 그런 점에서 이는 전투에서 새로운 점령지를 획득하는 것과 유사하다. 따라서 번역은 방법이나 전략적 측면에서 어떤 것을 활용하건 상관없다.(Callon, 1986: p.204) 물질적 이어서 오래 지속시킬 수 있다면 작동할 가능성이 더 높다는 것 외에는 어떤 선택적 선호도 존재하지 않는다.

2.2.2 두 개의 이질성, 하이브리드(Hybrid)와 이종성(Heterogeneity)

ANT에서 이종성은 1차적으로는 행위자네트워크를 구성하고 있는 행위자들의 이종성을 표현하기 위해 사용된다. 행위자네트워크는 인간과 비인간, 물질과 비물질, 현재와 과거 같은 이중적 행위자들이 결합되어 형성된 결합체들을 번역하여 네트워크의 일부가 되게 함으로써 확장된다. 라투어는 이를 하이브리드라고 부르는데,(라투어, 2010) 이는 인간과 비인

간, 존재와 인식과 같은 정화작용(purification)에 의한 이분법의 경계를 가로질러 매개에 의해 발생하는 이중적 결합이다. 라투어에 따르면 하이브리드는 정화작용과 다른 직교 축에서 - 번역의 축에서 - 발생하면서 번역이 한계에 도달할 때까지 무제한적으로 증식하여 행위자네트워크를 성장시킴과 동시에 위기로 끌고 간다. 이것은 모든 행위자를 기준이 되는 하나의 행위자네트워크의 시야에서 볼 때 기술되는 이중성인데, 이 메타포를 기준으로 ‘번역’을 해석하면 ANT가 염두에 두고 있는 전형적인 ‘번역’은 행위자네트워크가 갖고 있지 않은 방향 성분에 위치하는 하이브리드를 매개하여 권력의 질서 안에 동원하는 것이 된다. 번역이 성공적이라면 행위자들은 하이브리드화되어 OPP의 가설을 받아들이고, OPP의 가설을 따르지 않는 행위자네트워크 주위의 행위자들은 숫자가 줄어들게 된다. 그러나 번역이 더 많은 행위자와의 동맹을 시도할수록 OPP의 가설이 작동하지 않을 가능성은 증가한다. 번역의 질서는 더 희박하게 작동되며 자신의 OPP에 의한 하이브리드화가 아니라 다른 행위자네트워크의 OPP에 의한 하이브리드화가 진행될 확률이 증가한다. 물론 실패한 번역의 대상이 더 이상 증식되지 않는다면 위기는 심화되지 않겠지만, 이들이 빠르게 증식하고 있다면 상황은 달라진다. 따라서 라투어는 이러한 하이브리드를 인정하고 행위자네트워크 안의 성분들의 조합으로 구성될 수 없는 존재임을 인정하여 하이브리드를 독자적인 신분으로 받아들이기 전

까지는 하이브리드가 행위자네트워크에 위기를 초래한다고 말한다.(라투어, 2010)

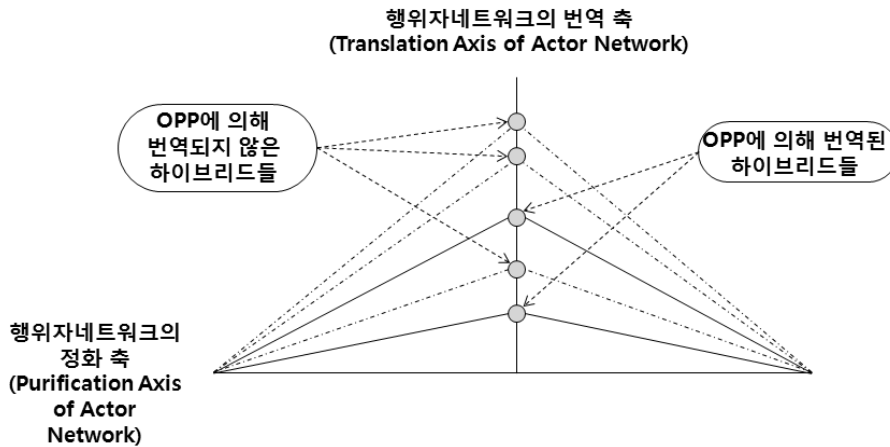


그림 1 Hybrid의 시야

번역의 축은 정화의 축에 대해 직교적(orthogonal)이고 독립적이다. 이는 하이브리드들의 증식이 정화의 축과 무관하다는 것을 뜻한다. 번역의 축 위에서 하이브리드들이 빠르게 증식하고 있다면 행위자네트워크는 더 멀리 있는 행위자 마저 번역해 내야 한다. 이 과정에서 번역되지 않음으로써 행위자네트워크를 위협하는 것이 아니라, 번역이 더 희박하게 작동함으로써 번역의 축 위에 있는 행위자들에 다른 행위자네트워크의 OPP가 작동될 위험이 증폭된다는 것을 의미한다.

때로 행위자네트워크는 행위자들을 ‘배제’ 함으로써 번역하기도 한

다. 그러나 만약 하이브리드들이 이러한 배제에도 불구하고 증식된다면, 번역의 축에서 증가하는 하이브리드들은 행위자네트워크를 위기에 빠뜨린다. ‘배제’라는 번역이 제대로 작동하지 않았기 때문이다. 게다가 그러한 증식을 매개하는 다른 행위자네트워크의 OPP가 작동하고 있다면 행위자네트워크는 또 다른 의미에서 이종적으로 변화한다. 하이브리드의 이러한 이종성은 복수의 ‘권력’의 관계에 대한 것인데, 이것이 ANT의 또 다른 용어인 두 번째의 이종성(heterogeneity)이다.

두 번째 이종성의 한 예가 바로 모든 행위자들이 서로 다른 복수의 행위자네트워크 안에 동시에 존재하기 때문에 생겨나는 이종적 엔지니어링(heterogeneous engineering)의 메커니즘이다.(Law, 1987; Callon, 1987) 이것은 복수의 행위자네트워크가 하나의 행위자에 대해 동시에 결합되어 만들어지는 행위자의 가변적 정체성에 대한 기술이다. 복수의 행위자네트워크의 관계는 ‘포함’이거나 ‘경합’으로 볼 수 있는데, 미셸 칼롱과 존 로는 기술 혁신의 사례를 통해 ‘포함’ 관계에서 나타나는 현상을 이종성(heterogeneity)이라는 용어로 기술한다. 칼롱은 이 메커니즘이 가시적으로 드러난 기술 혁신의 사례를 분석하면서, 그 안에서 활동하고 있는 엔지니어라는 인간행위자에 주목한다. 칼롱은 이러한 엔지니어들은 종종 기술 혁신(포함된 행위자네트워크)에 의해 변화될 사회(포함한 행위자네트워크)의 모습을 그려내는 ‘사회학자’의 역할을 수행하기를 요구 받

으며, 엔지니어로서의 역할과 ‘사회학자’로서의 역할이 뗄 수 없도록 완전히 결합되어 수행된다고 설명한다. 그리고 그러한 엔지니어들을 ‘엔지니어-사회학자’ (Engineer-Sociologist)로 재명명한다.⁵ 이들 ‘이종적 엔지니어’들은 동시에 복수의 행위자네트워크 안에서 활동하며, 기술적 대상을 사회의 구성요소로 번역해 내고 사회적 대상을 기술의 구성요소로 번역해 낸다.

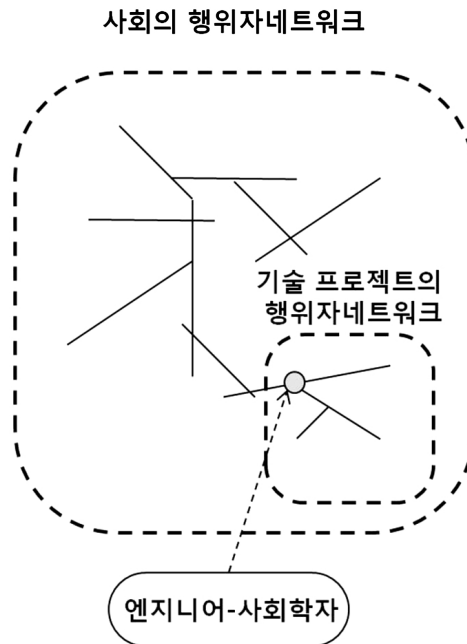


그림 2 Heterogeneity의 시야

⁵ 이는 깔롱이 의도적으로 엔지니어의 ‘사회학자’로서의 역할만을 일부러 부각시켜 분석한 것에서 비롯된 재명명이다.(Callon, 1987) 그러나 실제로는 ‘엔지니어-경제학자’나 ‘역사학자-엔지니어’ 같은 이종적 정체성을 엔지니어에게 부여할 수 있을 것으로 보인다.

ANT는 ‘번역’이라는 행위자네트워크의 존재론을 통해, 기술과 사회가 결합되어 작동할 뿐 아니라 그렇게만 존재한다고 설명함으로써 토대주의와 환원을 거부한다. 이는 다른 이론들이 다양한 형태로 환원을 피하려고 함에도 스스로 이론 안으로 불러들였던 이중성을 방법론적 수준으로 끌어내리거나 폐기함으로써 토대주의로 귀결되는 것과 분명하게 구분된다. ANT는 이중성을 이용하여 ‘사회적(혹은 기술적) 환원주의’(social or technical reductionism) 문제를 피해나가는데(Law, 1987: p.129), 행위자네트워크를 구성하는 행위자들 중 어떤 것은 다른 행위자에 비해 오래 지속되며, 행위자들 중에는 이동할 수 있는 행위자와 그렇지 않은 행위자가 있다는 것과 같은 존재론적 지위를 활용함으로써, 환원하지 않고도 선택을 종결시키는 기제를 만들어 낸다. ANT는 이중성 자체가 변화가 일어나는 계기이자 변화의 분기점을 결정하는 장치이며 선택을 종결시키는 기제라고 주장한다는 점에서 다른 이론과 다르다. ANT에서 이중적인 비인간행위자들이 의미작용을 종결시키는 이유는, 대개의 경우 그것이 다른 것을 환원시킬 정도로 더 근본적이어서가 아니라, 단지 그것이 더 오래 지속되기 때문이다.(라투어, 2010) 심지어 사회가 스스로를 재생산하는 유일한 이유는 그것이 이중적이기 때문이다.(Law, 1992: p.389)

2.2.3 ANT의 연구 방법론

ANT는 일반적으로 방법론을 제공하지 않는다. 그 이유는 ‘방법에 의해 지식을 보장하는’ 메커니즘을 인정하지 않기 때문이다. 따라서 ANT의 방법은 이론적 지식을 만들지 않고 ‘기술’(description)을 생산한다. ANT가 연구자들에게 제공하는 방법론의 전형은 인류학이나 민족지학으로부터 온 ‘이론 없는’ 미시적 연구인데, 그 이유는 ANT에서 행위자네트워크의 성장과 지속을 결정하는 거시적인 이중적 번역의 패턴은 계산의 중심(center of calculation)에 의해 성공적으로 지속되는 미시적 번역 패턴에서 비롯되며, 그것은 오직 현실에서 작동중인 미시적 번역에 대한 추적에 의해서만 발견되기 때문이다. 그렇지만 이것은 전통적 의미의 방법론은 아니며, 예시 정도에 지나지 않는다. 극단적으로는 ANT는 파이어아벤트식으로 방법에 저항한다.(Callon, 1986)

그러나 그 보다 선행되어야 하는 것은 그러한 발견을 수행할 대상을 선정하는 것이다. ANT는 새로운 질서의 발생이 하이브리드들이 빠르게 증식되는 것과 관련되어 있음을 명확하게 제시한다.(Latour, 1993) 따라서 ANT를 활용한 연구에서 대상을 발견하기 위한 최초의 출발점은 행위자들의 양적 증가가 발생하는 지점을 미시적으로 기술하는 것이다. 그것을 통해 그러한 증식의 과정과 행위자의 새로운 하이브리드화, 성장의 패턴이 되어 거시적 수준까지 밀고 올라가게 되는 이중적 번역의 질서를 찾

아내는 것이 ANT가 명시적이지는 않지만 인프라언어와 함께 암묵적으로 제공하는 연구 방법이다. 이러한 연구의 전략에 도움이 된다면, 모든 방법이 허용된다.

본 논문에서는 ANT를 이용한 연구 사례의 방법적 가이드라인을 따라, 인터넷 역사 연구의 대상으로 웹의 성장에서 행위자들의 급격한 증식이 나타난 지점으로서 구글 검색엔진이라는 허브 노드(hub node) 내부에서 작동하는 구글의 최초의 검색 알고리즘인 ‘페이지랭크’라 불리는 기술을 미시적으로 분해하여 그것이 어떤 행위자들과의 동맹을 통해 구축된 OPP를 매개하였는가를 살필 것이다. 그러나 여기서 연구는 두 개의 방향으로 분기되어 각자 다른 기술 방식을 사용해야 한다. 하나는 구글 검색엔진이 동맹을 맺은 행위자들의 시간을 거꾸로 거슬러 올라가는 계보학적 기술이고, 다른 하나는 구글의 검색기술인 ‘페이지랭크’를 성공으로 이끈 동맹이 인터넷의 역사 안에 등장했던 다른 행위자네트워크들의 동맹과 어떻게 달랐고, 최초의 성공을 통해 확보된 ‘검색엔진’이라는 블랙박스 안에서 네트워크의 확장을 위해 어떤 중심이동이 이루어졌는가를 살피는 것이다. 하지만 이 두 방향의 기술(description)은 가능한 ANT의 인프라언어를 활용하여 일관성을 유지하면서 기술되어야 하는데, 이를 위한 표준적 어휘들은 앞서 제시한 ‘번역’과 ‘이중성’을 포함하여 미셀 칼롱이 “번역의 사회학의 몇 가지 요소들”에서 미시적 사례 분석에

적용하였던 어휘들을 사용할 것이다.

본 연구는 인터넷의 역사학에 대한 일반적 메타이론이나 서술방법론을 제공하려는 것은 아니다. 다만 구글 검색엔진 기술 변화에 대한 미시적 기술(description)을 통해 인터넷 역사 연구를 위한 방법으로서 ANT의 이론적 자원을 활용한 ‘사회기술적 역공학’ (socio-technical reverse engineering)의 가능성을 확인하고, 가능하다면 웹 등장 이후 인터넷 역사 서술의 히스토리오그래피를 발굴하여 제공할 수 있게 되기를 희망한다.

3. 구글동맹에 참여한 하이브리드들의 계보와 기원

3.1 하이퍼링크와 구글 검색엔진의 ‘페이지랭크’ (PageRank)

‘페이지랭크’는 구글이 세계 최대의 인터넷 검색 서비스 회사가 될 수 있게 만든 핵심 기술이다. ‘페이지랭크’는 인터넷 상에 있는 웹 문서 중에서 사용자가 입력한 ‘검색 키워드’와 관련된 웹 문서 중에서 ‘중요도’가 가장 높은 문서를 찾아내는 기술이며, 이는 구글 이전의 검색엔진들이 제시했던 검색 결과와는 비교되지 않을 정도로 높은 만족도

를 제공한 검색 기술이다.(Page, 1999) 이 기술의 고안자인 래리 페이지(Lawrence Page)는 1999년의 ‘페이지랭크’를 설명하는 논문의 제목에서 ‘페이지랭크’가 웹에 질서(order)를 부여했다고 표현했는데, 만약 우리가 “검색엔진이 주도한 10년”이라고 부르려는 기간 동안의 인터넷 성장이 실체적인 것이라면 인터넷의 성장은 바로 이 ‘웹의 질서(order of Web)’로서의 ‘페이지랭크’와 연관되어 있다는 뜻이기도 하다. 우리는 구글이 ‘페이지랭크’라는 기술을 통해 만들어낸 미시적 질서가 어떻게 패턴이 되어 웹의 전 영역으로 확산되었는가를 살펴보기 위해 ‘페이지랭크’라는 기술이 식별하고 정의하는 주변의 행위자들은 무엇이고 그 행위자들이 어떻게 ‘페이지랭크’를 상호정의하면서 ‘페이지랭크’의 행위자네트워크 안으로 번역되어 들어갔는지를 살펴보아야 한다.(Callon, 1986)

‘페이지랭크’가 ‘웹에 질서를 부여했다’는 표현에 내포된 기술적(technical) 함의 중 하나는 구글의 검색 기술인 ‘페이지랭크’가 웹 문서가 아닌 일반적인 텍스트 검색에는 적용되지 않는다는 것이다. 그 이유는 ‘페이지랭크’가 일반적인 텍스트 안에는 존재하지 않는 요소인 ‘하이퍼링크(hyperlink)’ 정보를 검색에 사용하고 있기 때문이다. ‘하이퍼링크’란 웹과 같이 우리가 일상에서 경험하고 있는 ‘클릭 가능한’ 전자문서인 ‘하이퍼텍스트’의 구성 요소로서, 이미 많은 컴퓨터

이용자들이 익숙하게 사용하고 있는 ‘마우스 클릭에 의한 다른 문서로의 이동’을 구현하는 전자문서의 구문론적 요소다. 만일 ‘페이지랭크’가 다른 검색기술과 결정적으로 다른 점이 웹에 질서를 부여한 것이라는 설명을 받아들인다면, 웹에 부여했다는 질서는 어떤 형태로든 ‘하이퍼링크’와 관련되어 있을 것이다. 하지만 이는 ‘페이지랭크’가 ‘하이퍼링크’라는 이름으로 불리는 모든 대상과 맺은 관련성은 아니다. ‘하이퍼링크’는 웹에 적용되기까지 문서작성 소프트웨어나 게임, 컴퓨터 운영체제 등에서 다양한 형태로 존재하여 온 기술들을 총칭하는 용어이고, ‘페이지랭크’가 관련을 맺고 있는 대상은 오직 ‘웹의 하이퍼링크’일 뿐이기 때문이다.

그렇다면 ‘페이지랭크’는 왜 1997년의 웹에서 최초로 작동된 것일까? 이에 대한 가장 단순한 답변은 “발명자들이 그때가 되어서야 발명했기 때문”이겠지만, 이 질문을 약간 바꾸면 그렇게 쉽게 답할 수 있는 질문이 아니라는 것이 드러난다. ‘페이지랭크’ 기술은 팀 버너스 리(Tim Berners-Lee)가 웹을 탄생시킨 직후인 1990년대 초반의 웹 환경에서도 성공할 수 있었을까? ‘페이지랭크’ 기술은 1997년에 성공했던 것과 같은 방식으로 2016년에도 성공하고 있는 것일까? 이를 더 분명하게 보기 위해서는 먼저 ‘하이퍼링크’라는 블랙박스 안에 축적된 행위자들과 시간들을 풀어 헤쳐 볼 필요가 있다.

3.2 ‘인간-하이퍼링크’ 하이브리드의 두 형태

2차 세계대전이 막바지에 다다른 1945년 7월, 당시 트루먼 대통령의 과학자문관이었던 바니바 부시(Vannevar Bush)는 *Atlantic Monthly*에 “우리가 생각하는 대로” (As We May Think)라는 제목의 글을 발표했다.(Bush, 1945) 그 글에서 그는 개인들이 지식활동을 하는 과정에서 도움이 될 상상 속 기계를 하나 제안했다. 이 기계의 기본적 형태는 개인이 쉽게 자신이 작성한 자료를 마이크로필름으로 만들 수 있는 탁상형(desktop) 기계였다. 이것은 당시 미국에서 마이크로 필름 산업이 주도하여 확산시킨 ‘부식 위험이 있는 종이를 대체하는 지식과 정보의 보존’ 방법으로서의 마이크로 필름화와 관련된 제안이기도 하지만(Baker, 2002), 부시의 제안에는 그것을 넘어서는 다른 내용이 포함되어 있었다. 그것은 자료에 있는 키워드와 연관된 다른 마이크로 필름화된 문서나 자료들을 연결하는 색인을 만들어 개인들이 ‘연상하였던 생각의 흐름’을 자료들을 연결하는 기록으로 남겨두는 기능에 대한 고안이었다. 자료를 다시 보는 사람이 마이크로필름으로 영사기로 투사된 문서의 내용에서 특정 단어를 가리키는 것 만으로 그 단어에 연결된 다른 자료를 기계의 작동에 의해 신속하게 확인할 수 있게 하겠다는 것이다. 부시는 이러한 기계를 만든다면 개인과 인류는 지금까지의 이론 것 보다 훨씬 대단한 지적 성

과를 이룰 수 있을 것이라는 미래상을 제시했다.(Bush, 1945) 그는 이 장치의 이름을 ‘기억확장장치’ (MemEx : Memory Extender)라고 불렀는데, 이 기기가 개인의 기억을 논리적으로 정돈된 지식의 영역을 넘어 연상의 영역으로까지 확장할 뿐 아니라 특정 개인의 연상을 다른 사람에게까지 전달하는 ‘기억의 확장’을 가능하게 할 것이라고 생각한 것이다.

당시 메멕스(MemEx)는 구현되지 않았다. 그러나 그로부터 20여 년이 지난 1965년 테드 넬슨(Theodor H. Nelson)은 부시의 고안을 전혀 다른 기술적 기반 위에서 재정의하고, 그 안에서 전자문서의 형식인 ‘하이퍼텍스트’와 ‘하이퍼링크’의 개념을 제시했다.(Nelson, 1965) 넬슨은 스스로 자신의 기획이 비록 광학기계가 아닌 컴퓨터를 이용하여 구현된 것이지만 근본적 아이디어에 있어서는 부시 고안을 계승하였다고 설명했다. 그리고 사람들은 그것을 순순히 받아들였다. 아마도 사람들이 그러한 주장을 받아들인 이유는 넬슨의 고안이 제공하는 ‘효용’이 부시가 메멕스를 통해 사람들에게 주려고 했던 ‘효용’과 거의 일치한다는 판단 때문이었을 것이다. 그러나 넬슨의 하이퍼텍스트를 사용하는 인간과 메멕스의 사용자가 목표로 하는 것이 무엇인가를 자세히 보면, 넬슨이 단순히 메멕스의 광학기계를 컴퓨터로 바꾼 것이 아님을 알 수 있다. 넬슨은 부시가 메멕스 고안에서 제시한 사용자와는 전혀 다른 목표를 가진 하이퍼텍스트 사용자를 제시하고 있었다.

물론 두 고안의 사용자는 가장 중요한 것을 공유하고 있는 것처럼 보였다. 그것은 바로 ‘연결’을 만드는 것이었다. 부시는 텍스트의 선형적 구조 안에 논리적으로 기록할 수 있는 것을 넘어서는 ‘연상’의 기록을 남기기 위해 비선형적 연결들을 남기는 사용자에게 주목했는데(Bush, 1945), 이 사용자는 전통적인 의미의 텍스트 저자나 독자와는 달랐다. 메멕스가 가정했던 가장 중요한 사용자는 전통적인 ‘텍스트의 독자’와 메멕스라는 광학기계가 결합되어 새로운 목표를 갖게 된 ‘독자 하이브리드’였다. 하지만 넬슨의 고안에 등장한 ‘연결’을 만드는 사용자는 넬슨이 1965년 컴퓨팅기계협회(Association for Computing Machinery, ACM) 컨퍼런스에서 자신의 ‘하이퍼텍스트’가 부시를 계승한 것이라고 밝힌 최초의 순간에도 메멕스의 사용자와는 달랐다. 그는 자신이 고안한 시스템을 PRIDE(Personal Retrieval Indexing and Documentary Evolution)라고 명명했는데, 이 이름은 메멕스 보다 확장된 내포를 갖고 있었다.

하드웨어는 준비되었습니다...(중략)... 이제 이걸로 또 무엇을 할 수 있을지 생각해 봅시다. 이 발표에서 제가 가장 주목하고 있는 것은 이것이 소설이나 철학, 설교, 뉴스, 테크니컬 저술과 같은 모든 형식의 글쓰기에 적용하려는 것입니다.(Nelson, 1965)

넬슨은 컴퓨터에 의해 구현된 시스템을 통해 부시가 상상했던 메멕스의 사용자와 마찬가지로 연상의 흔적을 기록하여 ‘개인적으로 검색 가능하게 정보를 색인하는’ 사용자일 뿐 아니라, 본문 자체를 비선형적으로 쓰는(non-sequential writing) 사용자로 확장할 수 있다고 보았다. 그는 이것이 부시의 고안을 일반화시킨 것이라고 생각했는데, 이러한 일반화가 ‘문서의 진화’ (Documentary Evolution)를 의미하며, 이것이 저술과 발간, 참조, 색인, 검색, 쓰기와 읽기 등으로 연결된 전통적 문서 시스템에 내재된 문제들을 해결할 것이라고 예상했다. 물론 이것은 부시의 메멕스로부터 나올 수 있는 진화 경로 중 하나였다. 그러나 넬슨이 끌어들이려 한 ‘인간-소프트웨어’ 하이브리드는, ‘연결’을 만들고 그것을 읽는 독자에 가까웠던 메멕스의 하이브리드와는 달리 ‘새로운 유형의 저자’ 즉 ‘텍스트 시스템’ 안에서 발생한 ‘텍스트의 저자’가 ‘하이퍼텍스트’의 행위자네트워크로 결합되어 탄생한 ‘저자 하이브리드’를 향해 있었다.

넬슨이 웹의 탄생에 가장 결정적 기여자라는 것은 이견의 여지가 없다. 그 이유는 그가 메멕스의 기술적 기반을 컴퓨터로 옮김으로써, ‘하이퍼링크’를 컴퓨터를 이용한 다른 종류의 고안으로 이식 가능하게 만들었기 때문이다. 하지만 넬슨 자신이 기대했던 것은 그와는 다른 것이었다. 그는 텍스트와 멀티미디어 모든 영역의 저자들이 자신이 정의한 구조

를 가진 파일의 형태로 저작하는 하나의 OPP를 만들어내고자 하였는데, 그것은 하이퍼텍스트 작성 시스템(Hypertext Editing System: 이하 HES)이었다. 저자들이 더 이상 텍스트를 쓰지 않고 HES 앞에서 구조화된 파일을 만들게 된다면, 즉 ‘좋은 텍스트를 생산’ 하려는 자신의 목표를 ‘HES를 이용해 작성’ 하는 것으로 변경하는 것을 받아들인다면 HES는 ‘저자 하이브리드’ 들의 OPP가 될 수 있었다.

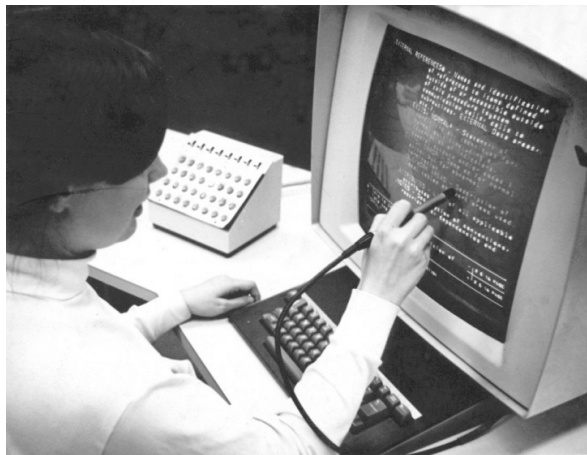


그림 3 제너두 프로젝트의 문서 편집기(Hypertext Editing System)

출처 : Wikipedia

그런 점에서 보면, HES를 단순한 메멕스의 계승으로 본 하이퍼텍스트에 대한 많은 역사적 기술들은 단지 부분적으로는 맞는 기술인데, (Bolter, 1994; Landow, 1991) 그 이유는 메멕스의 ‘하이퍼링크’가 글을 읽는 방식이었다면 HES의 ‘하이퍼링크’는 글을 쓰는 방식이었기 때문이다.

HES가 원한 행위자네트워크의 확장의 출발점은 글을 읽고 하이퍼링크를 남기거나 하이퍼링크를 따라 읽는 ‘독자 하이브리드’의 증식이 아니라 새로운 방식 - 하이퍼텍스트의 문법을 따라 혹은 과거의 비선형적 글쓰기를 꿈꾸었던 저자들이 원했던 방식 - 으로 글을 쓰는 ‘저자 하이브리드’들의 증식이었다.

3.3 ‘하이퍼미디어’로의 우회 : ‘인간-하이퍼링크’ 하이브리드의 증식

부시의 고안을 계승한 것은 넬슨만은 아니었는데, 넬슨이 ‘하이퍼텍스트’라는 개념을 제시하기 2년 전인 1963년 서덜랜드(Ivan Edward Sutherland)는 부시의 고안에서 인간과 기계의 상호작용에 대한 부분에 주목하여 이를 컴퓨터 기반 도면설계 장치인 ‘스케치패드’로 구현했다. 이때 사용된 것은 CRT(Cathode-Ray Tube) 모니터와 라이팅펜이었는데, 사용자는 펜으로 종이에 그림을 그리듯 모니터 위에 그림을 그릴 수 있었다.(Nelson, 1974: DM23-30)⁶ 하지만 ‘스케치패드’에는 부시의 고안에

⁶ 넬슨의 Computer Lib/Dream Machine은 Computer Lib라는 책과 Dream Machine이라는 두 책을 Back to Back으로 붙인 책이어서, 인용 페이지가 Computer Lib인 경우 CL로 Dream Machine인 경우 DM으로 표기하였다. 이는 넬슨 스스로가 책 내에서 사용한 참조 표기법이다.

서와 같이 ‘연상’에 의해 다른 저자의 자료와 연결을 기록하는 독자는 없었다. 스케치패드에는 도면 설계자가 소프트웨어 템플릿을 이용하여 도면에 표시하기 어려운 부가적 정보들을 화면 밖에 별도로 기록하고 이를 불러올 수 있도록 만듦으로써, 다중 차원을 가진 문서를 편집할 수 있고 읽을 수 있도록 만든 기기였다. 넬슨이 부시의 메멕스 기계로부터 ‘하이퍼링크’라는 기술을 분리해 명명함으로써 메멕스를 ‘저자 하이브리드’를 증식시키는 HES로 연결시킨 부시의 계승자라면, 서덜랜드는 부시의 메멕스를 포인팅 디바이스를 손에 든 ‘독자 하이브리드’로 연결시킨 메멕스의 또 다른 계승자였다.

이렇게 전자문서를 활용한 응용 컴퓨팅 기기가 등장하자, 넬슨의 하이퍼텍스트와 서덜랜드의 포인팅 디바이스를 이용한 사용자 인터페이스를 통합하는 새로운 혁신이 시도되었다. 1968년 더글러스 엥겔바트(Douglas C. Engelbart)는 서덜랜드의 고안이 도면작성 기기 뿐 아니라 넬슨의 하이퍼텍스트의 고안을 통합하는 더욱 범용화된 사용자 인터페이스로 활용될 수 있음을 간파했고, 이것을 위해 범용화된 그래픽사용자인터페이스(GUI : Graphic User Interface)에 더 적합한 포인팅 디바이스를 발명하고 이를 ‘마우스(Mouse)’라고 명명했다.(Nelson, 1974) 하지만 사용자들에게 GUI의 경험을 전파시킨 것은 마우스를 장착한 개인용 컴퓨터가 아니었다. GUI를 가장 대중적으로 확산시킨 계기는 조이스틱이나 TV 리모컨

과 같은 마우스 이전 단계의 포인팅 기기들을 장착한 ‘하이퍼미디어’들의 활발한 보급이었는데, 1980년대에는 퍼스널컴퓨터나 게임, 양방향 TV, 레이저디스크영상장치와 같은 미디어 기기들에 의해 제한적인 GUI를 활용한 다양한 시도가 이루어지기 시작했다.

‘하이퍼미디어’는 넬슨이 최초로 사용한 용어인데, 넬슨은 이것을 ‘하이퍼텍스트’의 확장된 형태로 보았고, 그런 이유로 넬슨에게 ‘하이퍼미디어’의 중심적 행위자는 ‘저자’였다. 그러나 현실의 ‘하이퍼미디어’는 넬슨의 기대와는 달리 ‘저자 하이브리드’와의 동맹을 중심으로 확산되지 않았다. 오히려 하이퍼텍스트 기술은 미디어 시스템이라는 행위자네트워크와 연결되자 미디어 ‘수용자’라는 인간 행위자와 결합하여 ‘하이퍼링크를 따라 가며 보는’ 목표를 가진 새로운 하이브리드를 만들어냈다. ‘하이퍼미디어 수용자’들은 빠르게 증식하여 네트워크를 성장시켰다. 그런데 이것은 ‘하이퍼미디어’ 행위자네트워크에 위기를 초래했는데, 그것은 ‘미디어’와 ‘수용자’라는 이분법을 더 이상 받아들이지 않는 하이브리드들이 PC 주변에서 증식되고 있었기 때문이다.

1989년이 되자 유럽입자물리연구소(CERN : Conseil Européenne pour la Recherche Nucléaire)의 팀 버너스 리는 당시 초기 성장을 하고 있던

인터넷을 통해 전송되는 문서 형식인 하이퍼텍스트마크업언어(HTML : Hyper Text Markup Language)라는 기술 규격과 함께 이 문서를 인터넷을 통해 전송해 주는 웹 서버와 클라이언트 소프트웨어로서의 가장 단순한 형태의 브라우저를 구현하였다.(Berners-Lee, 1989) 버너스 리의 하이퍼텍스트는 당시의 컴퓨터 사용자들에게는 너무 친숙한 것이었는데, 하이퍼텍스트는 이미 PC에서 전자문서를 읽으면서 활용하고 있는 하이퍼링크나 하이퍼미디어 기기에서 경험한 ‘하이퍼링크’의 경험과 크게 다르지 않았기 때문이었다. 웹은 이러한 인간-소프트웨어 하이브리드들과의 동맹을 성사시키기 좋은 상태에 있었다. 그러나 인터넷이 등장하기 전에 이미 ‘PC 통신’이나 문서 저작툴과 파일 전달 매체들이 이들과의 동맹을 시도하고 있었고, 인터넷이 등장한 후에는 고퍼(Gopher)나 뉴스그룹(Newsgroup) 같은 서비스들은 인터넷 안에서 많은 인간행위자들을 동맹 안으로 끌어들이었을 뿐 아니라 텍스트들을 실어 나르고 있었고, 심지어 고퍼는 위계적인 메뉴구조 안에서 제한적이거나 ‘하이퍼링크’ 기술을 활용하고 있었기 때문에 웹의 성공은 불확실했다.

버너스 리의 웹과 연결된 ‘하이퍼링크’라는 새로운 문제제기(problematization)가 성공적으로 작동하려면, 잘나가던 고퍼 서비스의 위기가 찾아와야 하고, 고퍼나 뉴스그룹 사용자들은 마우스 클릭 정도의 조작만으로 문서의 한 지점에서 다른 문서로의 ‘연상적인’ 이동을 원하

고 있어야 하고, 텍스트들은 고퍼의 위계적 메뉴구조나 뉴스그룹의 시계열적 구조를 넘어서 더 많은 사용자들의 화면 위로 달려나가길 원해야 했다. 이러한 모든 가설적 상황은 일반 사용자를 위한 브라우저인 모자이크(Mosaic)가 배포되기 시작한지 불과 1년여 만에 명백하게 해소되었다. 고퍼를 개발했던 미네소타 대학은 라이선스 사용료를 받겠다는 발표를 통해 고퍼를 스스로 위기에 빠뜨려 주었고(Patry, 2009: p.181), 웹의 하이퍼링크를 만들거나 클릭하는 인터넷 사용자는 고퍼나 뉴스그룹 사용자에 비해 엄청나게 빠른 속도로 증식했고(Frana, 2004: p.29) 문서들은 빠른 속도로 하이퍼링크를 담은 하이퍼텍스트로 재매개화(remediation)되어 배포되었다.(Bolter, 2001) 1995년이 되자 웹은 FTP나 고퍼, 텔넷과 같은 다른 모든 인터넷 서비스들을 압도함으로써 웹의 ‘하이퍼링크’는 가설적 단계를 넘어 인터넷에 연결된 인간행위자들과 비인간행위자들에게 OPP가 되었다. ‘하이퍼링크’가 지시하지 않는 웹 문서는 사용자들의 접근이 허락되지 않았기 때문에, 웹 문서로 존재하기 위해서는 ‘하이퍼링크’와 연결되어야만 했다. 1995년에 웹 사용자의 상당수는 웹 브라우저의 주소창에 www로 시작되는 유알엘(URL)을 입력하는 방식이 아닌 ‘하이퍼링크’를 클릭하는 방식으로 새로운 웹 페이지에 도달하기를 원하고 있었다.

웹의 인간-소프트웨어 하이브리드는 세 개의 서로 다른 원천으로부터

유래된 것인데, 하나는 ‘하이퍼링크를 만드는 독자’로서 메멕스로부터 온 것이고, 다른 하나는 ‘하이퍼텍스트’로 글을 쓰는 저자로서 넬슨의 하이퍼텍스트로부터, 마지막 하나는 ‘하이퍼링크를 클릭하는’ 하이퍼미디어의 독자로부터 온 것이었다. 이들이 자기들 간에 또는 웹의 다른 요소와 어떻게 결합되어 새로운 목표를 만들고 누가 더 많이 증식하게 될 것인가가 웹이 어떻게 될 것인지를 결정하는 요소였다.

‘하이퍼미디어’의 독자인 하이브리드들은 웹과 결합하자마자 가치 있는 정보를 ‘발견’하려는 새로운 목표를 향해 빠르게 증식했고 메멕스의 ‘하이퍼링크’를 만드는 독자 하이브리드와 넬슨의 ‘하이퍼텍스트’저자 하이브리드는 상대적으로 천천히 증식되었다. 하이퍼미디어의 독자들은 웹의 ‘하이퍼링크’를 더 많이 소비하기를 원했지만 그 증가 속도는 느렸고 밸런스가 이루어지지 않았다. 양질의 ‘하이퍼링크’를 공급할 수 있는 새로운 작동 방식이 웹의 성장을 위해 필수적이었다. 그 시기에 대거 웹으로 뛰어든 것은 광고 페이지로 가득 찬 닷컴(.com) 기업들이었다.

‘페이지랭크’가 마주했던 ‘하이퍼링크’는 바로 이 시기의 ‘하이퍼링크’였다. ‘하이퍼링크’를 소비하려는 사용자들에게는 하이퍼미디어에서 경험했던 것과 같은 양질의 ‘하이퍼링크’는 상대적으로 적고 한편으로는 나쁜 ‘하이퍼링크’들은 빠르게 증식하고 있었다. 메타태그

와 같은 웹 문서 작성의 가이드라인을 제대로 따르지 않는 웹의 저자들도 대거 늘어나고 있었다. 구글 말고도 야후! 디렉토리나 알타비스타 검색엔진과 같은 행위자네트워크들이 이렇게 제멋대로 증식하고 있는 웹에 질서를 부여하기 위해 ‘힘겨루기’를 하고 있었는데, 그 최초의 성공은 ‘야후!’가 만들어 냈다.

4. ‘번역’으로서의 배제 : 야후! 위기의 출발점

4.1 야후는 어떻게 웹을 번역하고자 했는가?

웹의 사용자들은 웹을 정보의 바다라고 불렀다. 이 비유는 매우 적절했는데, 그 이유는 웹은 독자들이 육지에서 출간된 책에서 거의 보지 못했던 정보들을 보여주었다는 것과 원하는 것을 찾기에는 너무 넓다는 점에서였다. 1995년 웹이 인터넷 서비스의 주류로 부상하자, 개인들의 웹사이트뿐 아니라 닷컴 기업들의 웹 사이트와 그 기업들로 몰려든 투자와 닷컴기업 간의 제휴에 의한 링크 연결이 기하급수적으로 증가했다.⁷ 따라서 사용자들은 자신에게 유용한 좋은 웹 문서의 링크를 찾아내는 것은

⁷ 2003년의 마이크로소프트의 연구에서, 웹페이지 당 하이퍼링크의 평균 개수가 62개를 넘는데 median 즉 빈도 측면에서 가장 많은 것은 23개를 조금 넘는 수준인 것으로 드러났다. (Broder et. al. 2003) 이는 페이지당 하이퍼링크 개수가 매우 많은 일부 닷컴기업의 웹페이지들이 평균을 크게 끌어 올린 것으로 해석된다. 참고로 1999년의 AT&T Labs의 연구는 평균 하이퍼링크 개수를 22.6개 수준인 것으로 보고한 바 있다.(Duchamp 1999)

더 어려워지고 있었다.(Brin and Page, 1998) 이러한 어려움을 해결하기 위해 웹 문서 독자들이 최초로 찾아낸 방법은 유용한 링크를 모아둔 ‘야후!’ (Yahoo!)와 같은 웹에 대한 ‘길잡이’ (directory) 혹은 ‘관문’ (portal) 사이트, 즉 웹에 대한 ‘하이퍼미디어’를 이용하는 것이었다.

야후! 디렉토리는 길잡이 사이트의 운영자가 직접 방문하여 살펴 본 웹 사이트로 연결되는 링크를 제공한다는 점에서 한 사람의 연상 경험이 담긴 문서를 다른 사람들에게 전달하는 메멕스의 고안과 매우 유사했다. 하지만 그 링크들은 ‘길잡이’ 페이지를 만드는 작업자들이 자기 자신의 연상이 아니라 임의의 타인에게서 나타날 연상을 추정하여 구성해낸 것이라는 점에서 메멕스의 연결과는 달랐다. 야후!는 ‘하이퍼링크를 만드는 독자’를 고용하여 분류학적 기준과 사이트의 콘텐츠 품질에 대한 평가에 따라 다른 웹 문서에 대한 하이퍼링크를 배열한 문서를 출간하는 (publishing) 일종의 ‘하이퍼미디어’였다. 그렇기 때문에 야후!에게는 모아 놓은 웹사이트 목록이 사용자들에게 유용한 링크인 것이 중요했고, 이를 위해 야후!는 미디어의 ‘편집자’로서 자신의 디렉토리에 등록을 원하는 사이트들에게 다음과 같은 조건을 제시했다.

아래의 내용을 포함 하는 경우에는 등록이 보류, 거절될 수 있으며, 등록된 이후라도 통보 없이 삭제 처리될 수 있습니다.

- * 상업용 사이트인 경우, 사이트 내에서 공식 상호를 확인할 수 없는 경우;
- * 사이트의 운영목적, 제품 및 서비스 등을 명확히 구분할 수 없는 메뉴로 구성된 경우;
- * 1개의 디렉토리에 등록을 하는 것을 원칙으로 하고 있습니다;
- * 사이트가 특정 지역을 위한 것인 경우, 방문자가 사이트 내에서 오프라인 주소를 확인할 수 없는 사이트;
- * 일부 내용이라도 완성되지 않은 공사중인 홈페이지;
- * 하나의 링크라도 연결이 안 되는 홈페이지;
- * 영어로 되어 있지 않은 홈페이지 (또는 영어 버전이 없는 홈페이지);
- * 특정 브라우저에서만 사용이 가능한 사이트 (예, 자바로만 구성된 사이트);
- * 주 7일, 하루 24시간 접속 가능하여야 함;
- * 사이트를 직접 운영하지 않고 다른 사이트를 주소만 포워딩한 경우;
- * 사용자가 보게 되는 내용과 다른 내용을 검색엔진에 제공하는 사이트

(출처 : Yahoo Directory Submit Terms Of Service)⁸

⁸ Yahoo Directory Submit Terms Of Service, 2016.01.13 Retrieved, <https://info.yahoo.com/legal/us/yahoo/directorysubmit/directorysubmit-159.html>

야후!는 자신들이 등록한 하이퍼링크에 사용자들을 속이거나 기대하지 않은 내용을 보게 만드는 사이트가 연결되는 것을 경계하고 있었다. 이중 일부는 야후!의 법률 자문이 각국의 상법이나 판례에서 디렉토리 서비스의 책임에 대한 사안을 참조하여 만든 것이었다. 성인물이나 약품의 판매와 같이 국가별로 내용적 기준이 달라지는 것도 중요한 고려사항이었다. 야후!에서 이러한 기준을 실행하는 것은 사람이었는데, 이들을 ‘서퍼(Surfer)’라고 불렀다. 이들은 그것 외에도 야후!의 내용적 등록 기준을 이해하고 등록 신청된 사이트가 그 조건에 부합하는지를 판단했다. 따라서 그들은 이런 기준을 스스로 만들고 이 기준에 따라 웹사이트의 등록 여부를 판단하며, 그 지역의 정서나 법률, 형편을 알고 있으면서, 다른 웹 사용자가 가장 자연스럽게 연상할 수 있는 방식으로 ‘분류 체계를 구성하는’ 사람들이었다.

그런 점에서 서퍼들은 매우 다중적인 존재였는데, 이들은 각 국가의 법률의 해석자이자 문화적 감수성의 대변자였고 이를 기초로 정보를 분류하는 분류학자이면서 웹에 대한 사용자 경험의 질적 평가기준을 체화시킨 웹 사용자의 대표였다. 이들은 마치 개가식 도서관의 사서들처럼, 신착 사이트를 어떤 분류에 넣어야 하는지를 판단하고, 새로운 카테고리를 신설할 필요가 있는지를 토론했고, ‘야후!’ 디렉토리에 등록된 사이트의 링크를 클릭하였을 때 그 링크가 깨져있다면 그것으로 인해 ‘야

후!’ 서비스의 품질에 불만을 갖게 되는 ‘사용자’들을 대변하기도 했다.⁹

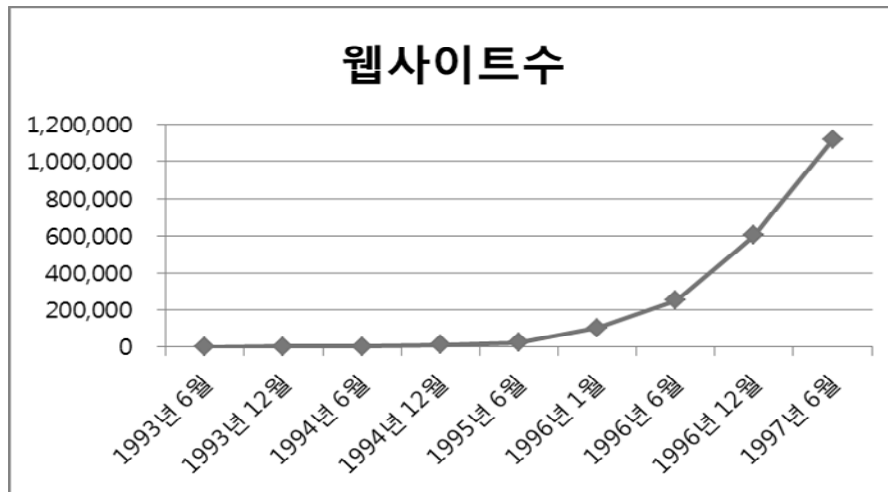


그림 4 월드와이드웹 사이트 증가 통계¹⁰

그러나 1997년이 되자 웹은 ‘길잡이 사이트’들이 인터넷 안에 있는 모든 웹 문서에 대해 다른 사용자를 대신하여 경험하고 이를 하나의 하이퍼미디어로 만드는 것이 불가능할 정도로 커져 있었다. 1999년에는 모든 웹페이지의 40%가 매주 갱신되었고 그 중 23%는 매일 갱신되었

⁹ [나는프로] ‘야후코리아’ 김경희 서평팀장, *이코노미21*, 2000, 2014.02.27 Retrieved, <http://www.economy21.co.kr/news/articleView.html?idxno=469>

¹⁰ Internet growth, *netvalley.com*, <http://www.netvalley.com/archives/mirrors/davemarsh-shats-1.htm> 내용을 이용하여 그래프로 재편집

다.(Cho and Garcia-Molina, 2000) 양적으로도 그 많은 사이트를 모두 디렉토리에 담을 수 없었고, 어제 확인한 사이트가 여전히 디렉토리에 소개된 내용을 제공하고 있을지도 알 수 없었다. 이러한 상황에서 ‘길잡이 사이트’ 들은 둘 중 하나를 택해야 했다. 하나는 제공하는 정보의 범위를 좁혀 뉴스나 날씨, 쇼핑, 구인/구직, 음악과 같이 가장 빈번하게 요구되는 정보를 높은 품질로 직접 제공함으로써 하이퍼미디어로서의 성격을 더욱 강화하는 것이었고, 다른 하나는 웹 문서를 읽고 평가하여 ‘하이퍼링크’를 만드는 일을 사람이 아니라 검색엔진과 같은 기계에게 맡겨 가능한 많은 웹 사이트에 대한 정보를 제공하면서도 비용 효율성을 달성하는 것이었다. 하지만 문제는 사람에 의해 선택된 링크와 같은 수준으로 ‘기준’을 적용하여 웹 사이트의 링크를 골라낼 수 있는 검색엔진이 없다는 것이었다.

결국 ‘길잡이 사이트’의 대표주자였던 야후!는 자신들이 충족시키려는 대상을 제한하는 방향으로 선회하여, 사용자들이 가장 빈번하게 찾는 정보들을 직접 제공하기 시작했다. 야후!는 결국 자신의 등록 기준을 모두 충족시킬 수 있는 소수의 사이트 밖의 웹 문서들을 자신의 네트워크 안으로 끌어들이기를 포기했는데 그 이유는 명백했다. 당시 웹 내에서 가장 빠르게 증식하는 행위자는 인터넷 안에 어딘가 있을지도 모를 임의의 웹페이지를 찾는 것이 아니라 자신이 원하는 내용이 담긴 ‘텍스트’를

향하는 독자 하이브리드들이었고, ‘길잡이 사이트’ 들로서는 독자들이 원하는 ‘텍스트’ 를 제시할 효과적인 방법이 있다면 굳이 어떤 위험이 도사리고 있을지 모르는 웹 페이지에 의존할 이유가 없었기 때문이다. 이러한 웹 페이지들에 대한 ‘배제(exclusion)’ 는 야후!가 선택한 이들에 대한 ‘번역’ 의 방식이었는데, 야후!의 배제는 시스템 수준에서는 잘 작동하였지만 그것이 번역의 축 위에서 증가하는 하이브리드들의 증가를 막지는 못하였다. 1996년 야후!는 웹 검색 서비스에 알타비스타(AltaVista) 검색엔진을 채택하면서 임의의 웹 페이지에 대한 동맹으로부터 멀어지기 시작했다. 하이브리드들은 야후!의 의도와는 다른 경로로 야후!의 네트워크 주위에서 증식되고 있었다.

4.2 하이퍼미디어의 독자를 흉내 내는 인간 vs. 텍스트 독자를 흉내 내는 검색엔진

최초의 웹 검색엔진이 등장한 것은 1993년이였다. 하지만 앞서 서술한 바와 같이 인터넷의 하이퍼텍스트 내용을 자동으로 ‘검색 가능한’ (searchable) 데이터베이스로 만들어주는 검색엔진들은 ‘야후!’의 경쟁 상대가 되기는 어려웠다. 검색엔진의 검색결과로 나온 링크를 클릭하여 나타나는 웹 문서들은 ‘야후!’ 디렉토리가 모아둔 웹 사이트 목록에 있

는 잘 만들어진 사이트들에 비해 정보의 품질이 일반적으로 낮았다.(Brin and Page, 1998: 107)

검색엔진들이 만들어낸 링크가 ‘길잡이 사이트’ 들에 비해 품질이 낮았던 이유는 다양하다. 초기의 검색엔진들은 웹 페이지의 규격 안에 들어 있는 ‘제목’ 이나 ‘메타 태그’ 만을 인덱싱 하여 검색 결과를 제공했다. 웹이 발달하기 시작한 초기에는 웹 문서의 작성자들은 웹 문서의 작성 규칙을 충분히 이해하고 이를 따르는 작성자들이어서, 웹 문서의 형식적 요소인 ‘제목’ 이나 ‘메타 태그’ 등에 해당 페이지가 담고 있는 내용에 대한 정보를 충실하게 입력했다. 하지만, 1994년 이후 웹이 대중화되었을 때 웹 문서 작성자들은 이미 통제되지 않을 만큼 다양한 사람들이었다. 이들은 제목이나 메타태그가 없는 웹 문서들을 만드는 경우도 많았고, 심지어 검색엔진들을 속여서 더 많은 방문자를 끌어오기 위해 ‘가짜 메타태그’ 나 키워드를 잔뜩 집어넣은 웹 문서를 기계적으로 생성하는(Spamdexing, Web Spamming) 상업적 서비스를 제공하는 기업들도 나타났다.(Metaxas and DeStefano, 2005) 검색엔진은 자동으로 웹 페이지를 수집하여 독자를 대신하여 웹 페이지의 텍스트를 읽고 그 경험을 다른 독자들을 위한 평균적 언어를 통해 하이퍼텍스트를 만들어야 했다. 하지만 이렇게 혼탁해진 웹 환경에서 검색엔진이 독자들을 대신하여 웹 페이지의 의미를 읽어내어 다른 독자들에게 ‘하이퍼링크’ 를 제공하는 것

은 너무 어려운 일이었다.

검색엔진 개발자들은 이런 문제들을 해결하기 위해 웹 페이지의 제목이나 메타태그를 넘어 본문 전체(Full Text) 정보를 수집하여 의미를 추출하는 자연어 검색 기술을 도입했다. 이는 단순히 제목과 메타 태그만을 수집하는 초기의 웹 로봇에 비해 훨씬 많은 컴퓨터와 저장장치를 요구했다. 특히 전문검색(全文檢索, Full Text Search)은 언어별로 정교한 분석 능력을 갖춘 색인기(indexer)를 필요로 했다. 한국어와 같이 어미의 변형이 심한 언어들에서 소프트웨어가 형태소 분석(形態素 分析, stemming)을 통해 ‘의미’를 담고 있는 어근을 추출하여 검색을 위한 비교 연산을 하는 것은 더 많은 컴퓨팅 자원을 필요로 했다. 알타비스타와 같은 선도적 검색엔진들은 자신들이 다른 검색엔진과 달리 메타태그나 제목 만이 아닌, 얼마나 많은 웹 문서를 전문색인(full text indexing)하고 있는가를 홍보했는데, 경쟁의 초기에 실제로 더 많은 웹 문서를 색인한 전문 검색 엔진들이 더 만족스러운 검색 결과를 내기도 했다.¹¹

초기 전통적 정보검색 기술을 활용한 웹 검색엔진의 기술적 지향점은 ‘텍스트의 저자’를 흉내 내는 것이었다. 만약 검색엔진이 검색 대상 웹

¹¹ 이러한 대표적 사례는 컴퓨터 하드웨어 기업인 DEC가 자신들이 만든 64비트 Alpha 칩의 우수성을 선전하기 위해 만들었던, ‘Altavista’라는 검색 엔진이 있다. “Altavista”. *Wikipedia*. 2015 [updated 14:42, 16 September 2015; cited 22 September 2015]. <https://en.wikipedia.org/wiki/AltaVista>

페이지 안에서 사용된 키워드를 색인화하고 검색 사용자가 입력한 키워드에 따라 적절한 웹 페이지의 링크를 제공할 수 있다면, 야후! 디렉토리에 등록되지 않은 웹 사이트 내의 페이지를 찾는 독자 하이브리드들은 검색엔진 주위로 모여들 것이기 때문이다. 그러나 그러한 가설에 의한 OPP로서의 검색엔진은 좀처럼 작동되지 않았다. 그 이유는 두 가지로 볼 수 있는데, 하나는 웹 페이지의 저자와 독자는 텍스트의 저자와 독자와는 달랐다는 것이고, 다른 하나는 웹은 전통적 인쇄매체와는 달리 독자의 시선을 끄는 다른 하이퍼링크와의 심각한 경쟁이 벌어지는 공간이라는 것이다.

‘텍스트의 저자’를 흉내 내는 정보 검색기술은 해당 텍스트 안에서 사용한 어휘를 정확하게 검색어로 입력하여야 대상 문서를 찾아낼 수 있었기 때문에, 검색자가 해당 분야에서 흔히 사용되는 키워드(topical query)와 같은 지식을 보유하고 있는 ‘저자를 닮은 독자’, 즉 ‘텍스트의 저자’가 염두에 두었던 바로 그 독자인 경우에만 작동했다.(Singhal and Kaszkiel, 2001) 하지만 웹의 검색자들은 저자가 사용한 어휘에 대한 지식이 없는 경우가 훨씬 많았다. 저자를 전혀 닮지 않은 독자에게 이러한 검색엔진의 유용성은 급격히 저하되었다. 또한 정보 검색기술 기반의 웹 검색엔진들은 하이퍼텍스트의 구문론을 잘 따르면서 하이퍼텍스트 형식으로 글을 쓰는 저자 하이브리드들이 검색엔진이라는 OPP를 통해 빠

르게 증식되기를 원했지만, 그것을 통해 증식된 것은 좋은 글을 쓰는 인간 행위자만은 아니었다. 검색엔진을 통해 빠르게 증식된 또 다른 존재는 검색결과에 자신들이 원하는 링크를 노출시키기를 원하는 스팸머들이 자동화된 웹 스팸 생성엔진을 통해 쏟아낸 웹 페이지들이었다. 검색엔진들은 ‘좋은 하이퍼텍스트’의 저자들 보다 스팸머들을 더 많이 대표하기 시작했다. 스팸머들은 자신의 ‘독자’인 텍스트 기반 검색엔진을 잘 이해하는 저자였기 때문이다.¹² 검색엔진은 자신을 ‘역공학(reverse engineering)으로 번역한 스팸머들에게 길들여지고 있었다. (Metaxas and DeStefano, 2005) 검색엔진들은 스팸머들을 ‘배제’하는 방식으로 번역하기를 원했지만, 오히려 검색엔진의 행위자네트워크가 스팸머들에 의해 번역된 것이다.

하지만 야후! 디렉토리가 고용한 서퍼들은 일반적인 웹 사용자가 해당 분야의 사전 지식이 없는 것을 전제로 디렉토리의 카테고리명을 결정하고 직관적인 연상에 의해서 원하는 정보로 찾아갈 수 있도록 만드는 것을 목표로 했다. 특정 지역을 위한 길잡이 사이트는 그 지역에 정통하거나 해당 지역의 문화적 배경을 대표하는 사람의 ‘시각’에서 독자들이 이해할 수 있는 용어로 필요한 정보를 제공하는 것이 중요하다. 야후!의

¹² 이 기술은 현재 SEO(Search Engine Optimization)이라고 불리는데, 이는 검색엔진이라는 독자에게 최적의 방법으로 메시지를 전달한다는 점에서 매우 적절한 명칭이다.

서퍼들은 하이퍼미디어의 독자들을 대표하여(represent) 정보를 읽고 독자들을 위해 이를 독자들의 언어로 번역해 주는 저자들이었다. 따라서 야후!의 분류 체계에는 다양한 배경 지식이 번역되어 들어 있었다.

결과적으로 ‘텍스트의 저자’를 훔내 내어 동맹으로 끌어들이려 한 검색엔진들은 ‘하이퍼미디어의 독자’를 훔내 내어 하이퍼링크를 생산하는 인간 행위자인 야후! 디렉토리가 고용한 서퍼들과의 경쟁에서 패배했다. 하지만 야후!는 이 승리의 결과로 새로운 위기로 접어들었다. 자신의 디렉토리에 등록되지 않은 웹 페이지를 만드는 저자와 이를 찾는 독자 하이브리드를 배제함으로써, 웹을 야후! 디렉토리에 등록된 가치 있는 사이트와 그렇지 않은 가치 없는 사이트로 정화시키고자 했다. 그러나 그러한 이분법은 작동하지 않았고 야후!가 배제해 버린 디렉토리에 등록되지 않은 사이트에 게재된 웹 문서의 저자들과 이를 찾으려는 독자들은 웹의 성장과 함께 빠르게 증식하고 있었다. 하이브리드의 증식이 언제나 그렇듯이 이들은 야후!가 구축한 웹의 질서에 위기를 심화시키고 있었다. 야후!에 의해 배제된 하이브리드들이 자신의 목표를 달성시켜줄 새로운 동맹을 찾아내지 못한다면 스스로의 목표를 변경함으로써 야후!의 행위자 네트워크에 의해 제시된 OPP를 통해 번역되었겠지만, 이 하이브리드들은 여전히 누군가 새로운 동맹의 대상을 찾고 있었기 때문이다.

4.3 새로운 유형의 검색엔진들과 동맹 대상의 재구성

‘텍스트의 저자’를 흉내 내려는 웹 검색엔진의 기술적 한계가 드러나자 검색기술의 연구자들은 다른 접근법을 찾아 나섰다. 그것은 바로 ‘텍스트’가 아닌 웹의 인간 행위자를 흉내 내는 것이었다. 만약 웹의 인간 행위자들 역시 더 좋은 텍스트를 찾거나 읽기 위한 행동을 하고 있다면, 그들의 행동을 흉내 냄으로써 그들과 동맹 맺기를 원하는 웹의 독자 하이브리드들을 끌어들이 수 있을지도 모르기 때문이다.

검색엔진 연구자들이 포착한 대표적인 웹의 인간 행위자 모델은 두 개였다. 그 중 하나는 웹의 인간 행위자 중에는 더 신뢰할만한 행위자가 있으며, 그런 행위자들은 자신이 관심을 가진 웹 문서에 대한 하이퍼링크를 ‘자신이 작성하는 웹 페이지’에 끌어 모으는 행위자라는 모델이었다. 만약 이것이 이상적인 웹의 인간 행위자의 행동양식이라면 이들이 만든 웹 페이지(hub, 허브)들을 찾아내고 그 페이지 내의 특정한 앵커텍스트¹³를 가진 링크들이 공통적으로 지시하는 웹 페이지를 찾는 것이 그 키워드에 대해 가장 권위 있는 텍스트를 찾는 길이 될 것이다. 이는 마치

¹³ 앵커텍스트(anchor text)는 하이퍼링크를 만들 때 사용되는 문구인데, 하이퍼링크를 읽는 독자들이 이를 읽고 하이퍼링크에 연결된 문서의 내용을 짐작하도록 하이퍼링크의 작성자가 입력한 문구를 의미한다. 예를 들면, 어떤 하이퍼링크에 ‘서울대학교’라는 텍스트가 표시되고 이를 누르면 www.snu.ac.kr로 연결되도록 한다면, 이 하이퍼링크의 앵커텍스트는 ‘서울대학교’가 된다.

특정 키워드에 대한 ‘좋은 길잡이 사이트’를 찾는 것과 유사했는데, 특정 단어에 대해 좋은 길잡이 사이트들에서 공통적으로 포함하고 있는 링크로 많이 연결된 페이지가 그 키워드에 대해 더 ‘권위 있는’ 웹 페이지일 것이라는 아이디어였다.¹⁴ 다른 하나는 자신의 웹 페이지를 구성하면서 그것을 자신과 웹 페이지를 방문하는 제3자에게 더욱 유용하게 만들려는 목적으로 관련된 웹 페이지에 대한 링크를 그 웹 페이지를 특징짓는 앵커텍스트를 포함한 하이퍼링크로 연결해 놓는 웹의 사용자 모델이었다.¹⁵ 만일 검색엔진이 그런 사용자들의 페이지를 최대한 많이 모으고 그로부터 그 사용자들이 특정 앵커텍스트 키워드로 묘사하여 하이퍼링크로 연결한 웹 페이지 중에서 더 많은 사용자들이 ‘기억해 둘 필요가 있다’고 생각한 웹 페이지를 찾아낼 수 있다면, 키워드에 대해 더 좋은 텍스트를 찾을 가능성은 매우 높아진다. 하지만 문제는 어떻게 어디에 있을지 모르는 그 문서를 링크하고 있는 인터넷 상의 다른 문서에 대한 정보 – 이를 ‘백링크’(backlink)라고 부른다 – 를 어떻게 최대한 많이 모으고 그것으로부터 문서와 특정 키워드 사이의 관련성을 계산할 것인가

¹⁴ 이러한 접근을 대표하는 기술은 HITS(Hyperlink-Induced Topic Search)인데, 인터넷 안에서 특정 키워드에 대한 Authority가 가장 높은 웹 페이지를 링크 구조를 계산하여 찾아 내려는 접근을 하고 있다. (Kleinberg 1999)

¹⁵ 이러한 접근을 대표하는 기술은 위에서 자세히 보게 될 리엔홍의 HVV(Hyperlink Vector Voting) 기술과 구글의 ‘페이지랭크’였다.

가였다.

검색 기술 전문가들은 이렇게 자신이 발견한 웹 페이지를 ‘하이퍼링크’로 남기거나 그것을 평가하여 검색 키워드에 적합한 웹 페이지를 찾아내는 사용자를 훔쳐내는 접근법을 링크 기반 검색(Link based Search)이라고 불렀다. 물론 여전히 하이퍼미디어의 독자 하이브리드로부터 온 행위자들에 대해서는 야후!와 경쟁할 수는 없었지만, 링크 기반 검색엔진들은 전통적인 ‘텍스트의 저자’를 훔쳐내는 키워드 기반 검색엔진(keyword based search engine)과의 동맹 경쟁에서 앞서 나가고 있었다. 하지만 일반 검색엔진 사용자들이 이들 사이의 질적 차이를 쉽게 판단을 내릴 수 있었던 것과 달리 검색기술 전문가 집단은 쉽게 합의에 도달하지 못했다. 왜 그런지는 조금 더 자세히 살펴 볼 필요가 있다.

4.4 ‘검색기술’은 누구를 대표하여야 하는가?

웹이 등장하기 전에 이미 ‘정보검색 기술’은 다양한 시도를 하고 있었다. 이미 다양한 형태로 데이터베이스화된 문서들의 검색을 위한 기술들이 수십 년 동안 발전해 왔고, 기업이나 대학, 공공부문의 문서들은 빠르게 전자문서로 대체되어 갔기 때문에, 그 안에 들어 있는 정보를 신속하게 찾아내는 기술에 많은 소프트웨어 기업들이 투자를 하고 있었다. 그 중에서도 뉴스나 도서 같은 비정형적인 텍스트 정보로부터 신속하게

원하는 데이터를 찾아내는 기술은 1980년대 이후로 기술 기업들의 중요한 도전의 장이 되었다. 상용화 가능한 비정형 문서 검색 기술이 나오기만 한다면, 전자화된 문서 정보들에 대한 다양한 응용이 가능할 것이라는 것이 기술 기업과 검색 전문가들의 기대였다.

주요 선진국들은 이러한 기술의 발전을 장려하기 위해, 도서관이나 뉴스 라이브러리를 대상으로 한 검색 품질 평가 프로그램과 연계하여 검색 품질을 높이기 위한 학술대회를 개최하고 있었다.(Langville and Meyer, 2011: p.8) 미국의 TREC(Text REtrieval Conference)이나 일본의 NTCIR(NII Test Collection for Information Retrieval system), 유럽의 CLEF(Conference and Labs of the Evaluation Forum)와 같은 테스트 콜렉션들은 적게는 수천 건에서 수만 건까지 다양한 언어로 된 검색기술 평가를 위한 테스트 문서 세트를 제공하였다. 주목할 사실은 1999년 개최된 TREC-8과 2000년의 TREC-9에서 모두 공식 결과 보고서는 ‘페이지 랭크’를 포함하여 링크 기반(link-based) 알고리즘들의 검색 품질이 텍스트의 내용만을 이용한 키워드 기반(keyword-based) 정보검색 기술과 비교했을 때 유의한 차이를 보이지 않았다고 결론을 냈다는 사실이다.(Hawking et. al., 1999; Hawking 2000)

이러한 결과는 정보검색 기술 커뮤니티 내의 엔지니어들에게 ‘직관적으로 이상하다’는 반응을 불러 일으켰다. 당시 AT&T에 있었던 검색

기술 전문가인 아미트 싱할(Amit Singhal)은 2001년에 키워드 기반 검색기술과 링크 기반 검색기술을 비교하는 방법에 대한 논문을 발표했는데, 그의 결론은 TREC의 평가 방법이 잘못되었고, 링크 기반 기술이 상당한 차이를 보이며 우월하다는 것이었다.¹⁶ 싱할은 TREC이 웹 검색 트랙에서 사용한 평가방법의 문제를 지적하면서, 문서 테스트 세트와 알고리즘에 포함되어 있을 실제 웹 환경과 다른 요소와 휴리스틱(heuristic)한 요소를 배제시키기 위해 테스트 세트를 실제 웹으로 재구성하고, 시험 질의어(query)를 선정함에 있어서도 TREC에서 사용된 ‘주제별 질의어’(topical query)를 사용하지 않고 ‘실제 사용자 질의어’(real user query)를 사용하는 방식으로 변경하였는데, 이는 실제로 두 기술의 비교 평가에 상당한 영향을 주는 것으로 드러났다.¹⁷

싱할은 자신의 연구가, ‘웹 검색’에서는 전통적인 문서 검색에서 사용되는 키워드 기반 검색기술 보다 페이지 간 링크 정보를 활용한 검색기술이 확연하게 우월하다는 것을 입증한 것이라고 생각했다. 하지만 싱할은 자신이 한 ‘변경된 시험’의 진정한 의미는 알지 못했다. 당연히

¹⁶ “우리의 연구 동기는 TREC의 web track [5, 6, 7]에서 나온 결과와 웹 검색 커뮤니티 내의 일반적인 믿음 사이에 나타난 모순적 상황이었다.” (Singhal and Kaszkiel, 2001)

¹⁷ 싱할이 이 연구를 시작한 것은 AT&T에서였지만, 2001년 5월에 홍콩에서 논문을 발표하는 때는 이미 구글에 입사를 한 후였다. 그리고 그는 2014년 현재 Mr. Search로서 구글의 검색을 총괄하고 있다.

TREC의 평가 기준으로 구글과 비슷한 성능인 것으로 평가를 받은 다른 검색엔진들이 싱할의 시험 기준은 통과하지 못했다는 사실이 무엇을 의미하는지도 알지 못했다. 다음 그림은 이 두 평가를 도식화하여 비교한 것인데, TREC이 검색 대상으로 삼은 웹 문서와 색인들은 ‘텍스트의 저자’가 만들어낸 문서와 색인들이지만, 싱할이 검색 대상으로 삼은 ‘실제의 웹 페이지’는 다른 웹 문서에 대한 ‘하이퍼링크’를 남긴 독자 하이브리드들의 키워드가 포함된 대상이다. 또한 TREC의 평가 기준은 검색엔진들이 어떻게 ‘텍스트의 저자’를 잘 대표하여 독자들의 요청에 적합한 검색 결과를 제공하는가를 평가하는 것인 반면, 싱할이 링크 기반 검색엔진들을 위해 채택한 평가 기준은 검색엔진이 검색 사용자 보다 텍스트를 먼저 읽고 ‘하이퍼링크’를 남긴 독자 하이브리드들을 얼마나 잘 대표하여 검색 사용자에게 검색 결과를 제시하고 있는지를 평가하고 있었던 것이다. 결과적으로 싱할은 TREC의 웹 검색 시험이 ‘방법적으로’ 잘못되었음을 주장한 것이 아니라 링크 기반 검색엔진의 행위자네트워크가 TREC이 염두에 둔 ‘텍스트의 저자’를 대표하는 것보다 더 많은 ‘독자 하이브리드’들과 동맹을 맺는 것이 더 많은 검색 사용자를 동맹으로 끌어들이 수 있음을 발견한 것이다. 그는 웹을 제한적으로 모사한 데이터 세트에서 실제 웹으로 검색 대상을 변경하였을 뿐 아니라 (Singhal and Kaszkiel, 2001) 검색엔진이 대표하려는 행위자를 변경한 것

이다.

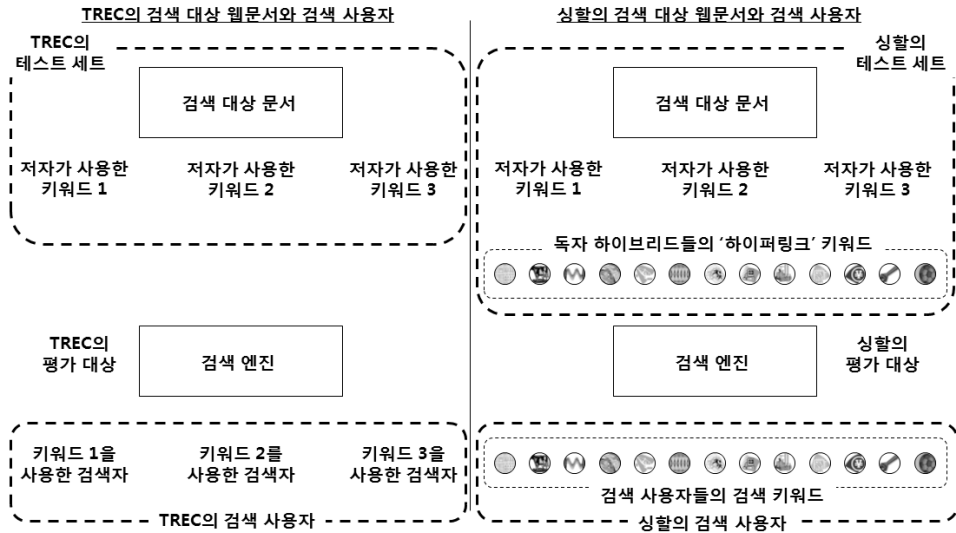


그림 5 TREC과 Simhal 평가 대상의 차이

키워드 기반 검색엔진과 링크 기반 검색엔진은 서로 다른 대상을 동맹으로 끌어들이기를 원했고, 따라서 행위자네트워크의 확장을 위한 ‘문제제기’가 달랐다. TREC의 평가대상인 검색엔진의 OPP가 가설적으로 포착한 검색 사용자는 ‘텍스트의 저자’를 흉내 내어 검색 키워드를 선택할 수 있어야 했지만, Simhal의 평가대상인 검색엔진이 포착한 검색 사용자는 찾아내려는 웹 문서로 향한 ‘하이퍼링크’에 포함된 많은 키워드 중에 자신이 연상한 검색 키워드가 있기만 하면 되었다. Simhal의 평가 방

법론이 TREC과 다른 결론에 이른 이유는 평가 대상인 검색엔진들의 알고리즘 차이가 아니라 검색 대상과 검색 사용자를 다르게 정의한 서로 다른 기술 시스템에 대한 평가였기 때문이다. 싱할이 링크 기반 검색엔진이 일반 웹 검색에 있어 키워드 기반 검색엔진 보다 더 좋은 결과를 낸다는 믿음을 확인할 수 있었던 이유는, 그 시기에 형성된 웹 검색을 둘러싼 행위자들의 분포, 즉 웹 검색의 사회-기술적 아장스망(socio-technical agencement) 때문이었다. 과속방지턱이 만들어 낸 ‘경찰관’의 의미 작용은(Latour, 1994) 운전자가 아닌 인간에게는 나타나지 않는다. 파리의 대학 구내에서 자동차를 운전하는 운전자가 증식되지 않았다면 과속방지턱은 아무런 의미도 만들어 내지 못했을 것이다. ‘경찰관’은 경찰관의 근무 시간 중에 과속범칙금을 받는 것을 꺼리는 운전자만을 번역할 수 있었지만 과속방지턱은 경찰관이 자는 동안에 자동차의 서스펜션이 고장나는 것을 꺼리는 운전자까지 번역해 내는 확장성을 가진 아장스망을 만들어내고 있었다. 링크 기반 검색엔진이 더 좋은 검색 결과를 만들기 시작했다는 것은 단순히 링크 기반 검색기술이 웹에서 더 훌륭한 검색 기술이라는 뜻이 아니라, 링크 기반 검색기술이 더 좋은 결과를 만들 수 있도록 하는 하이브리드들이 증식하고 있음을 의미했고, 이것은 곧 야후!의 ‘배제’라는 번역에 의해 시작된 야후!의 위기이면서 동시에 구글 성공의 출발점이었다.

5. ‘페이지랭크’는 어떻게 검색엔진이 되었나

‘페이지랭크’는 전통적인 정보검색엔진과는 많이 달랐지만, 자신이 다른 웹 검색엔진과 마찬가지로 정보를 ‘검색(Search)’하고 있다고 주장했다. 그러나 ‘페이지랭크’를 포함한 ‘링크 기반 검색엔진’들이 말하는 ‘검색’은 전통적인 정보 검색엔진들의 ‘검색’과는 다른 것이었다. 그럼에도 구글은 ‘페이지랭크’를 앞세워 ‘검색’의 의미를 변경하고, 웹 페이지와 검색 사용자, 검색 광고를 구매하는 광고주를 포함한 하이버리드들에게 OPP로 채택시키는데 성공했다. 본 장에서는 구글의 이러한 성공이 ‘더 좋은 기술’인 ‘페이지랭크’의 당연한 귀결이 아니라 는 것을 살펴볼 것인데, 이는 구글이 가치와 사회에 대해 중립적인 탈역사적 기술혁신으로써 ‘가장 좋은 검색엔진’이 된 것이 아니라 스스로 다른 행위자들과 결합하여 그들이 가지고 있던 ‘목표’들을 변형시킴으로써 ‘검색엔진’이 되었음을 의미한다.

5.1 ‘색인(index)’에 대한 새로운 정의, ‘독자의 문서’를 색인하다

구글의 페이지랭크 외에도 ‘하이퍼링크’를 이용한 접근법을 사용한 검색엔진은 이미 있었다. 그것은 바로 다우존스의 검색 엔지니어였던 리엔홍이 개발한 랭크덱스(Rankdex)라는 검색엔진이었다.¹⁸ ‘페이지랭크’는 랭크덱스의 랭킹 계산 방법을 일부 개선하여 만들어진 것인데, 그 차이는 랭크덱스가 ‘더 많은 웹 문서에 ‘백링크’가 걸린 웹 문서가 더 중요하다’는 논리만을 적용한 반면, ‘페이지랭크’는 ‘더 중요한 웹 문서가 ‘백링크’를 건 웹 문서가 더 중요하다’는 논리를 추가로 적용한 것뿐이었다.(Page, et. al., 1999) 이 두 검색 알고리즘은 모두 인터넷 안의 가능한 많은 문서들을 수집하여 ‘하이퍼링크’가 걸려 있는 문구인 앵커텍스트에 포함되어 있는 키워드들을 색인하였다. 그러나 이들은 ‘색인’의 의미를 전통적 텍스트 검색엔진과는 다른 의미로 사용하고 있었다.

검색은 전통적으로 ‘저자의 문서’에 대한 색인(index)을 통해 구현

¹⁸ Rankdex의 개발자인 리엔홍은 나중에 중국으로 건너와 중국 최대의 검색엔진이 ‘바이두’를 창업하였다. Rankdex의 알고리즘은 HVV(Hyperlink Vector Voting)이라고 명명되었는데, 이는 ‘백링크’의 정보를 이용한 최초의 랭킹 알고리즘이었다. 그런 점에서 ‘바이두’는 구글의 카피캣이 아니라 오히려 반대로 구글이 ‘바이두’의 카피캣이었던 셈이다.

되어 왔다. 색인은 특정한 키워드가 문서 내에 있는 문장들 속 어느 곳에서 사용되고 있는지를 지시하는 목록을 의미한다. 이러한 색인은 대개 책의 마지막에 모아져 있고, 동일한 키워드가 책의 다양한 위치에서 사용되고 있다면 이 목록의 정렬 순서는 페이지 순서에 따른다. 그런데 만일 누군가 하나의 일련번호로 연결되지 않은 많은 문서 더미에 대한 하나의 색인을 만들려고 한다면 어떻게 해야 할까? 각 문장에서 키워드를 추출하는 것은 동일하겠지만, 특정 키워드가 등장한 문서들이 복수로 존재한다면 그 문서의 목록을 어떤 순서에 따라 나열할 것인가? 아마도 한 가지 방법은 그 문서들의 제목에 대한 사전식 정렬일 것이다. 이것이 전통적인 정보검색 시스템의 출발점이었다. 이러한 시스템에서는 검색자가 특정 키워드를 검색어로 입력하면 그 키워드를 포함하고 있는 문서의 목록을 제목의 사전식 정렬로 보여준다. 이 과정은 논리적으로 필연적인데 그 이유는 어떤 문서의 제목이 검색결과 목록에 올라와 있다면, 그것은 바로 그 문서가 입력된 키워드를 ‘어딘가에’ 포함하고 있기 때문이다.

하지만 ‘페이지랭크’가 이용하는 ‘하이퍼링크’의 ‘앵커텍스트’ 색인은 ‘연상적 색인(associative index)’이라는 점에서, 문서 안에 존재하는 키워드 색인과는 다르다. 극단적으로 보면 ‘연상적 색인’에서 어떤 문서는 자기 자신 안에 전혀 포함되어 있지 않은 키워드를 색인 키워

드로 가질 수도 있고, 이는 심지어 다른 언어 간에도 가능하다.¹⁹ 이는 색인과 문서 사이의 관계가 전과는 다르게 정의되었음을 뜻하는데, 리옌홍은 자신이 발표한 1997년 논문에서 자신의 아이디어 안에서 ‘색인’과 ‘문서’의 관계를 다음과 같이 설명하고 있다.

“(키워드 기반 검색에서 사용되는 - 인용자)벡터 공간 모델에서, 문서는 하나의 벡터로 대표(represent)된다. 이 벡터의 각 차원은 단어 나 문서의 내용으로부터 추출된 개념들로 구성된다. (반면에 - 인용자) HVV(Hyperlink Vector Voting)²⁰에서 문서는 0개 이상의 링크 벡터로 대표된다. 각 링크 벡터는 그 문서로 향하는 링크를 대표한다. 하이퍼링크에 대한 기술(description)인 앵커텍스트는 하이퍼링크의 내용으로 취급된다. 그리고 그 벡터는 링크를 대표한다.

¹⁹ 예를 들어, ‘Deep Learning’이라는 새로운 기계학습 기술에 대해 인터넷 안에서 해당 분야 전문가들에게 가장 좋은 평가를 받는 영어로 작성된 문서가 있다면, 한국어를 사용하는 전문가들이 자신의 한국어 웹 페이지에 이를 소개하면서 ‘딥 러닝’이라는 새로운 어휘들을 포함한 앵커텍스트를 사용하여 영어로 된 문서에 ‘하이퍼링크’를 걸 수 있다. 이때 ‘하이퍼링크’의 구문론적 요소인 ‘앵커텍스트’에서 추출된 ‘딥 러닝’이라는 키워드는 영어로 된 ‘Deep Learning’ 문서의 ‘연상적 색인’이 될 수 있다는 뜻이다.

²⁰ HVV(Hyperlink Vector Voting)는 앞서 언급한 Rankdex의 랭킹 알고리즘으로, ‘페이지 랭크’를 설명할 때 흔히 사용되는 “논문의 인용 지표를 응용한”이라는 아이디어를 웹 검색에 최초로 도입한 알고리즘이다.

...(중략) 예를 들면, “<http://www.javasoft.com/tutorial>“라는 웹 문서를 가리키는 189개의 링크가 있다면, 이 문서는 189개의 링크를 대표하는 것이다.” (Li and Rafsky, 1997)

그러나 리엔홍은 ‘대표’라는 관계를 혼동하였는데, 문서가 자신을 가리키는 189개의 링크들을 대표하는 것이라는 그의 착각은 HVV에서 ‘링크 벡터’가 ‘문서’가 아닌 ‘링크’를 대표한다는 식의 형식적 절충으로 표현된다. 키워드 기반 검색에서 문서를 하나의 벡터가 대표하는 것처럼 보인 이유는 - 지역구 선거에서 해당 지역 내에 주소지를 둔 후보를 선출하는 것처럼 - 벡터의 각 차원이 문서 안에서만 추출되기 때문인데, 실제로는 문서는 벡터 차원만큼의 서로 다른 색인에 의해 각자 다르게 대표되고 있다. HVV에서는 벡터의 각 차원이 문서 밖에서 우발적으로 발생하기 때문에 문서를 대표하는 것이 하나의 확정적 벡터가 아니었지만, 리엔홍은 그런 대의제를 받아들일 수 없었기 때문에 문서가 자신을 가리키는 링크들을 대표하고 링크 정보로 구성된 벡터가 다시 문서를 대표한다고 혼동한 것이다. 리엔홍의 이러한 혼동은 단순한 논리적 혼동이 아니라 HVV가 찾아낸 새로운 하이브리드를 키워드 기반 검색엔진의 행위자네트워크 구성 성분의 ‘혼합물’로 번역해내려고 시도한 결과였다. 리엔홍은 자신이 고안한 HVV가 ‘키워드 기반 검색기술’의 한 형태로 보이게 만

둡으로써, HVV를 ‘키워드 기반 검색기술’의 행위자네트워크를 확장하는 하이브리드로 받아들여지게 하려는 ‘이종적 엔지니어’(heterogeneous engineer)로 행동한 것이다.

링크 기반 검색기술에서 ‘색인’은 더 이상 문서 내의 키워드가 아니지만 여전히 문서를 대표한다. 또한 어떤 키워드가 문서를 대표하도록 할 것인가도 문서의 저자가 아닌 그 문서를 읽고 그에 대한 ‘하이퍼링크’를 만든 독자가 결정한다. 따라서 링크 기반 검색기술은 문서나 문서의 저자가 아닌 그 문서를 지시하는 ‘하이퍼링크’와 그 생성자, 즉 저자가 아닌 독자가 문서를 대표하게 하는 새로운 네트워크 확장 전략을 실행하고 있었던 것이다. 이들은 웹을 ‘저자의 세계’가 아닌 ‘독자의 세계’로 해석하고 색인하고 있었다. 하지만 이들은 여전히 키워드 기반 검색기술에 의해 이미 만들어져 있던 ‘검색엔진의 행위자네트워크’에 번역되기를 원하는 하이브리드이면서 키워드 기반 검색기술과 경쟁함으로써 이를 위협하는 새로운 행위자네트워크였다.

5.2 기술로서의 ‘페이지랭크’ 탄생의 전사(前史) : 대용량 문서 처리 시스템 COPS

구글이라는 기업의 탄생 배경은 이미 많은 서적과 기사를 통해 알려져 있다. 특히 두 창업자의 스토리는 첫만남부터 매우 세세하게 기록되어 있는데, 러시아 이민자인 세르게이 브린(Sergey Brin)과 미시건 출신 래리 페이지(Larry Page)가 매우 똑똑한 스탠포드대학의 박사과정 학생이었던 것과 ‘페이지랭크’라는 구글의 검색 알고리즘이 학술논문의 인용지수를 하이퍼링크에 응용한 것이라는 사실을 공통적으로 들려준다. 그러나 실제로 구글의 기원을 알기 위해서는 래리 페이지가 페이지랭크 알고리즘을 고안한 1997년으로부터 2년을 더 거슬러 올라갈 필요가 있다.

1995년 세르게이 브린은 헥토르 가르시아 몰리나(Hector Garcia-Molina) 교수의 지도 하에 스탠포드 디지털 도서관 프로젝트에 관련된 논문을 발표했다.(Brin et al., 1995) 브린은 1994년에 스탠포드 대학의 디지털 도서관 프로젝트(Paepcke et al., 1996)의 일환으로 ‘저작권’ 보호를 위해 텍스트의 복제를 탐지하는 “저작권보호시스템”(COpyright Protection System: 이하 COPS로 약칭)이라는 프로젝트에 참여했다. 그 프로젝트 팀은 원 저작자의 텍스트 저작물을 인덱싱하여 다른 저작물 등에서 이를 공정 이용(Fair Use)의 범위를 벗어나서 사용하는 것을 찾아내

는 것을 목표로 했고, 그 출발점은 대용량의 텍스트 저작물 데이터베이스에 대해 임의의 문서가 그러한 데이터베이스 내의 문서를 부분적으로라도 복제 혹은 변형 복제하였는지를 고속으로 판정하는 복제탐지 시스템을 만드는 것이었다.(Peter, 1995)

이 시스템의 최초 참조 모델은 당시에 이미 많이 연구가 축적되어 있던 정보검색 시스템이었다. 그러나 사용자가 입력하는 검색어를 문서 데이터베이스 안에서 찾아내는 정보검색 시스템과 복제 여부를 판정하는 시스템 사이에는 크고 작은 차이점이 있었다. 무엇보다 사용자의 검색어는 짧은 몇 개의 단어이기 때문에 검색 연산의 부하가 상대적으로 적은 것에 비해, 복제 탐지는 짧게는 수십 개의 단어에서 많게는 수천 개 혹은 수십만 개 이상의 단어들로 구성된 문서를 질의(query)로 하여 데이터베이스를 검색해야 했기 때문이다.(Brin et al., 1995) 따라서 스탠포드 대학의 연구팀이 직면한 문제는, 문서의 복제 탐지를 위해 어떻게 정확할 뿐만 아니라 비용 효율적인 시스템을 구성할 수 있을 것인가의 문제였다.(Shivakumar and Garcia-Molina, 1995) 정확하더라도 비용이 높다면 의미가 없고, 비용을 낮출 때 정확도가 심각하게 훼손된다면 그 또한 의미가 없었다. 그런 이유에서 연구팀은 실제의 탐색 알고리즘은 비용이 낮고

고속처리가 가능한 ‘이진탐색’²¹이어야 한다는 것을 전제로 두고, 이를 위해 문서를 복제 판단의 기준이 될 수 있는 ‘의미 단위’로 분할하고자 했다. 연구팀은 이것을 청크(chunk)라고 불렀는데, 이 청크의 후보로 검토된 것은 4가지였다.

그 중에서도 COPS 시스템에 사용할 수 있는 기본적 조건인 ‘탐지 성능’을 어느 정도 이상 충족시키는 후보는 두 가지였다. 하나는, 예를 들어 “서울대학교”라는 텍스트를 “서울”, “울대”, “대학”, “학교”와 같이, 문서의 시작점부터 일정한 글자수로 문서를 잘라서 하나의 청크로 사용하고 그것과 같은 길이의 청크를 우측으로 한 글자씩 이동시켜 가면서 추출하여 그대로 인덱스로 사용하는 N-gram - 한 청크의 길이에 따라 bigram, trigram과 같이 N의 값이 달라지는 - 방식이었고, 다른 하나는 텍스트의 형식적 요소인 마침표와 같은 문장부호를 이용하여 하나의 청크를 중첩되지 않게 추출하고 이를 일정한 길이의 데이터로 만들기 위해 해시함수(Hash function)²²를 사용하여 인덱스를 만드는 해시

²¹ ‘이진탐색’은 동일한 길이의 이진 데이터를 비트(bit) 단위로 비교하여 일치하는지를 찾는 연산으로, 폰 노이만(John von Neumann) 컴퓨터의 구조에서 가장 빠른 데이터 비교 연산이다.

²² “해시 함수”는 임의의 길이의 데이터를 고정된 길이의 데이터로 매핑하는 알고리즘이다. 즉 원래의 데이터 값이 다르면 해시값도 달라진다. 그러나 서로 다른 데이터의 해시값이 같을 확률인 해시 충돌 확률은 0이 아니다. 이 확률이 작은 해시 함수가 더 질 높은 해시 함수로 평가된다.(참조 : 해시 함수, *위키피디아*, https://ko.wikipedia.org/wiki/해시_함수, 2016.05.16 Retrieved)

인덱스 방식이었다.²³

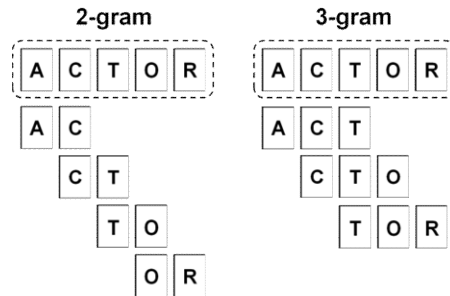


그림 6 N-gram 청크 개념도

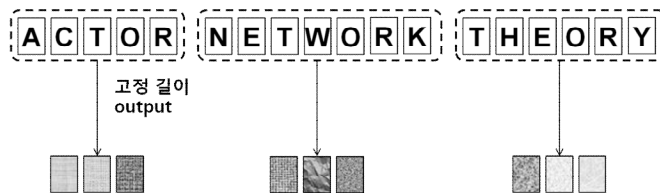


그림 7 해쉬 청크 개념도

결과적으로 COPS와 그것을 발전시킨 SCAM(Stanford Copy Analysis Mechanism)이라는 복제탐지 시스템에는 해시 청크가 선택되었다.²⁴ 하지

²³ 나머지 두 개의 청크는, 개별 알파벳을 각각 하나의 청크로 보는 것과 고정 길이로 만들기 위한 해시를 쓰지 않고 앞의 청크와 뒤의 청크가 공유하는 철자가 없도록(exclusive) 고정 길이로 문자열을 추출하는 - “ACTO”, “R NE”, “TWOR”, “K TH”, “EOR Y”와 같이 - 것이었다. (Brin et al. 1995)

²⁴ N-gram은 해시 청크와 마찬가지로 고정 길이 이진탐색이 가능한 연산적 장점이 있었지만, 해시 청크에 비해 처리해야 할 데이터량이 커지고 역공학(reverse engineering) 공격에 취약한 보안 상의 문제를 안고 있었기 때문이다.

만 COPS에서 배제되었던 N-gram은 구글이 텍스트를 처리하는 기술로 사용되었는데, 그 이유는 구글의 ‘페이지랭크’가 검색 대상 문서의 본문 검색하는 기술이 아니었기 때문이다. 만약 ‘페이지랭크’가 본문 검색기술로 태어났었다면 당시 텍스트 검색 기술로서는 기술적 한계가 널리 공감되어 있던 N-gram은 절대 구글의 텍스트 처리 기술이 될 수 없었겠지만, 본문 검색기술이 아닌 ‘페이지랭크’에서는 텍스트 처리가 검색랭킹을 결정하는데 미치는 영향이 매우 작았기 때문에, N-gram은 COPS 프로젝트의 멤버들이 주축이 되었던 ‘페이지랭크’의 최초 구현에서 앵커텍스트의 처리 알고리즘으로 선택될 수 있었다. 그리고 그들의 이러한 선택은 구글의 미래를 바꾸는 결정적 분기점을 제공했다.

5.3 ‘페이지랭크’ 행위자네트워크의 탄생

‘페이지랭크’를 처음으로 고안한 래리 페이지가 처음으로 제기한 문제는 ‘임의의 웹 문서를 지시하는 백링크 정보를 모두 수집하여, 앵커 텍스트 내에 특정한 키워드를 포함한 백링크가 해당 문서를 지시하는 양을, 각 백링크를 포함한 웹 문서의 중요도만큼 가중치를 주어 계산하면, 특정 키워드에 대해 인터넷 안에서 가장 중요한 웹 문서가 무엇인지 알 수 있지 않을까?’라는 것이었다. 하지만 래리 페이지는 이 문제에 대해 답을 내릴 수 없었다. 이 문제들에 대한 답변을 얻을 방법은 오직 ‘실제

로 만들어 보는 것' 밖에 없었다.

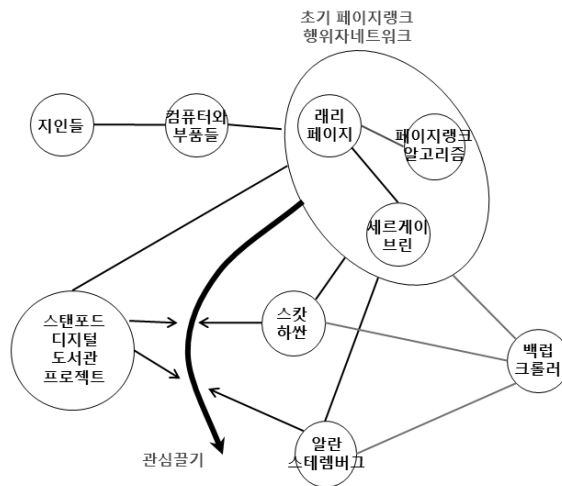


그림 8 초기 페이지랭크의 행위자네트워크 확장

래리 페이지는 동료인 세르게이 브린에게 이 아이디어를 설명하고, 이것을 계산하는 알고리즘을 만들 수 있을 것인지를 물었다. 세르게이 브린은 COPS 프로젝트를 통해 대용량 문서 데이터를 인덱싱하고 추출된 청크를 이용하여 고속으로 유사도를 계산하는 수학적 모델을 연구한 적이 있었기 때문에, 래리 페이지에게 세르게이 브린은 최적의 조언자이자 이 문제에 답을 내기 위해 끌어들여야 하는 최선의 인간행위자였다. 그렇지만 브린만으로는 이 문제들에 대한 해답을 구할 수 없었다. 브린은 이 문제에 대한 수학적 방정식을 만들 수는 있었지만, 이를 시스템으로 구현하

기 위해서는 성능이 좋은 컴퓨터와 이 컴퓨터를 시스템으로 작동시키기 위해 시스템 환경을 구성해 줄 수 있는 시스템 엔지니어와 브린의 수학적 실용 가능한 소프트웨어로 구현해 줄 수 있는 소프트웨어 엔지니어가 필요했다. 페이지와 브린은 이를 위해 COPS 프로젝트에 참여했던 동료들 중에서 이를 구현하는데 도움을 줄 수 있는 시스템 엔지니어인 스캇 하싼(Scott W. Hassan)과 소프트웨어 엔지니어인 알란 스테렘버그(Alan Steremberg)를 참여시켰다.²⁵ 또한 지인들의 컴퓨터 부품들을 끌어모아서 가능한 충분한 스토리지를 갖춘 하드웨어도 갖추었다. 이러한 행위자들은 ‘페이지랭크’의 주변으로 모여들어 각자의 역할을 부여 받았고, 그것을 통해 만들어진 것이 ‘백럽 크롤러’(Backrub Crawler)라고 불리는 최초의 ‘페이지랭크’ 머신이었다.

²⁵ 이들이 COPS 프로젝트의 일원이었음은 브린의 논문 외에도 Röscheisen(1997)이나 Paepcke et al. (1996) 등의 사료를 통해 확인된다.

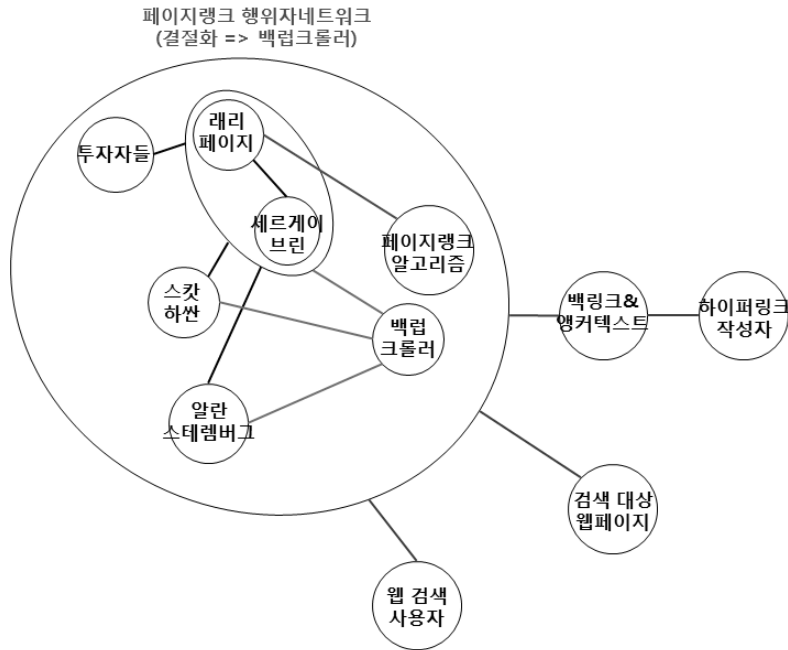


그림 9 페이지랭크가 ‘백랍 크롤러’로 결절화된 행위자네트워크

‘백랍 크롤러’는 인터넷에 접속된 컴퓨터 내에서 작동하면서, 웹 문서에 포함된 링크들을 수집하고 이를 따라 다른 웹 문서들을 발견하는 동작을 반복적으로 수행하는 ‘웹 로봇’과 수집된 웹 문서들로부터 ‘앵커 텍스트’와 ‘링크’를 쌍으로 추출하여 저장하면서, ‘앵커 텍스트’를 N-gram으로 분절하여 인덱싱하는 ‘인덱서’, 그리고 이렇게 수집된 웹 문서 URL과 N-gram 인덱스, 백링크 정보를 이용하여 각 웹 문서의 특정 키워드에 대한 랭킹을 계산하는 ‘랭킹 시스템’으로 구성되었다.(Brin and Page, 1998)

‘백럽 크롤러’가 처음으로 작동하기 시작했을 때, 백럽에 의해 제기된 문제는 다음과 같은 것이었다. 수집된 웹 문서들로부터 획득된 데이터에 대해 페이지랭크 알고리즘은 문제없이 특정 키워드에 대한 모든 웹 문서의 페이지랭크를 계산해 낼 수 있을 것인가? 그렇게 계산된 ‘중요도’가 과연 디렉토리 밖의 웹 페이지를 찾는 독자들에게 ‘질의 키워드와의 관련성’의 의미로 받아들여질 수 있을 것인가? 과연 백럽 크롤러가 수집한 웹 페이지 안에서 추출한 ‘하이퍼링크’들은 ‘페이지랭크’ 알고리즘이 ‘중요도’를 계산하기 위해 필요한 독자 하이브리드가 만든 ‘하이퍼링크’로 순순히 동작해줄 것인가? 이 모든 것은 불확실했고, 어느 하나라도 작동하지 않는다면 ‘페이지랭크’는 다른 검색엔진들과 마찬가지로 빠르게 증식 중인 하이브리드들의 대표가 될 수 없었다.

백럽이 최초로 수집한 웹 문서들은 스탠포드 대학 내의 웹 문서들이었는데, 두 창업자는 백럽이 만들어낸 랭킹에 대한 피드백을 받기 위해, 스탠포드 대학 내의 지인들에게 백럽의 랭킹을 확인할 수 있는 웹 주소를 알리기 시작했다.(Auletta, 2010) 그러자 최초에 ‘페이지랭크’가 제기했던 문제들은 하나씩 작동되기 시작했다. 스탠포드 대학 내의 웹 문서들에 대한 랭킹 계산이 문제없이 수행되어, 아래 그림과 같이 스탠포드 대학 내의 웹 페이지들은 ‘페이지랭크’와 결합되면서 더 많은 방문자를 유인하려는 목표를 더 높은 백럽 랭킹을 얻는 것으로 변경하였고, 스

탤포드 대학 내의 웹 페이지의 저자들은 ‘더 좋은 하이퍼텍스트’의 저자가 되려는 목표를 멈추고 ‘페이지랭크’를 작동시키기 위해 ‘더 좋은 하이퍼링크를 만드는’ 독자 하이브리드가 되어 자기 자신의 문서가 아닌 다른 문서에 투표하는 것으로 목표를 변경했다. 또한 백럽의 사용자들은 백럽의 랭킹이 웹 페이지를 찾는 독자가 생각한 ‘관련성’과 매우 유사하다는 것을 입증시켜 주었다. 이들의 숫자가 늘어나는 것은 백럽이 웹 페이지에 대한 검색에 대해 제기했던 문제제기가 작동하여, 백럽이 그와 관련된 행위자들의 OPP가 될 수 있다는 것을 의미했다. 이들의 연합은 ‘페이지랭크’의 문제제기를 따라 각자가 원하는 것을 ‘페이지랭크’와의 관계 위에서 재정의함으로써 성공적으로 구성되었다. 백럽은 자신의 행위자네트워크 안으로 징병된(enlist) 행위자들을 동원(mobilization)하여, 더 많은 웹 검색 사용자와 더 많은 웹 페이지를 끌어 모았다. 그러나 두 창업자는 아직 웹 검색엔진을 발명하지 않았다. 두 대학원생에게 백럽 크롤러는 여전히 자신들의 아이디어인 ‘페이지랭크’ 기술의 행위자네트워크 안의 중요한 행위자였을 뿐, 그 자체로 검색엔진으로서의 성장 메커니즘을 가진 새로운 행위자는 아니었다.

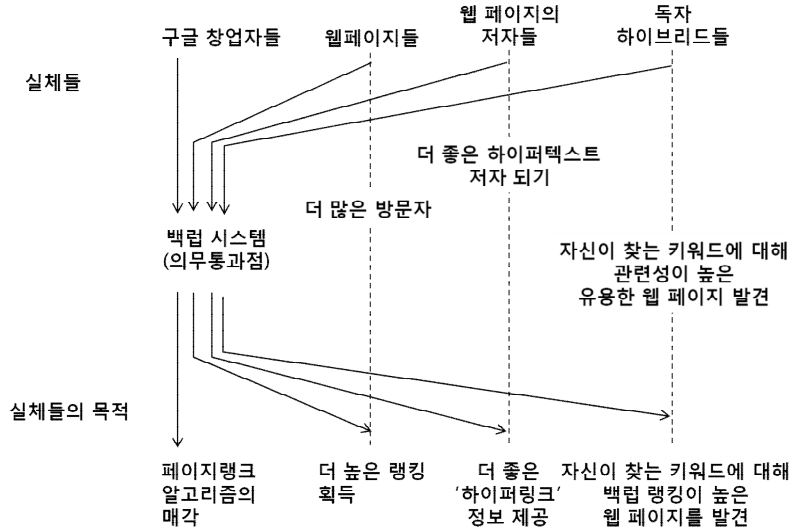


그림 10 OPP로서의 백업 시스템

두 창업자는 1997년 포털인 Exite.com에 ‘페이지랭크’를 160만 달러에 매각하려고 했으나 이미 임의의 웹 페이지를 배제한 야후!의 성공을 목격한 포털들은 새로운 검색 기술에 대한 흥미를 잃은 상황이었기 때문에 그 시도는 무산되었다. 아이러니하게도 두 대학원생이 ‘페이지랭크’의 행위자네트워크를 통해 달성하고자 했던 ‘의미 작용’, 즉 페이지랭크 ‘기술의 매각’을 포기한 실패의 순간²⁶이 바로 그들이 웹 검색엔진

²⁶ 두 창업자가 excite.com에 페이지랭크를 매각하려다 실패한 시점은 1997년이었던 것으로 보이는데, 테크런치(Techcrunch)나 포춘(Fortune) 등 일부 매체 기사(Siegler 2010)가 1999년으로 표기하고 있다. 하지만 1999년으로 거론되는 시점은 구글이 세쿼이아캐피탈 등 으로부터 거액의 투자를 받은 시점과 5개월 밖에 차이가 나지 않으며, 스티븐 레비(Steven Levy)의 *In the Plex*를 인용한 비즈니스인사이드(Business Insider) 등의 다른 기사에는 1997년 1월로 언급된 것으로 미루어 볼 때, 매각 실패의 시점은 1997년이다.

을 발명한 순간이었다. 매각에 실패한 백럽 크롤러는 이제 ‘페이지랭크’ 기술의 행위자네트워크를 대표하여 대학 내의 일부 사용자와 기술의 매수자와의 동맹이 아니라 스스로 검색엔진이 되는 새로운 동맹을 맺어야만 지속 가능한 행위자네트워크가 되었기 때문이다.

5.4 백럽 크롤러, ‘검색엔진’ 이 되다

구글이 등장하기 전부터 인터넷에는 ‘검색엔진’ 들이 있었다. 여기서 ‘검색’ 이라는 단어는 사용자가 입력한 질의와 대상 문서에서 추출된 인덱스를 비교하여 검색 질의에 연관된 문서들을 결과로 보여준다는 것을 의미했다. 그런 의미에서 보면 구글은 검색엔진이 아니었다. 구글은 검색 대상 문서로부터 검색 인덱스를 추출하지 않기 때문이다. 그런 이유에서 백럽 프로젝트의 참여자들과 그 주변의 사람들은 백럽을 검색엔진이라고 부르기보다는 크롤러라고 불렀다. 당시 백럽이 수행하던 기능을 기준으로 본다면 백럽은 랭킹엔진이라고 불리는 것이 적합했다.²⁷ 하지만 시장에서는 ‘페이지랭크’ 라는 랭킹엔진 기술이 아니라 성능이 개선된

²⁷ 래리 페이지는 1997년 1월에 미국 특허청에 제출한 특허출원서에서, ‘페이지랭크’ 가 ‘검색엔진’ 을 위해 사용될 수 있으며 그런 경우 단지 웹 페이지의 제목에 대한 색인 만으로도 높은 품질의 검색 결과를 제공할 수 있다고 설명하고 있다. 그리고 백럽 시스템을 ‘데모 시스템’ 으로 지칭하여, 상용 수준의 ‘검색 엔진’ 이 아님을 드러내고 있다. 또한 백럽이 아직은 ‘제목’ 만을 인덱싱하고 있다는 사실도 언급한다.(Page 1997)

상용 ‘검색엔진’을 필요로 했기 때문에, 구글은 ‘검색엔진’이 되어야만 했다.

우리의 주된 목표는 웹 검색엔진의 질적인 면을 개선하는 것이었다. 1994년 당시만 해도 완전한 검색엔진 인덱스를 갖추으로써 무엇이든 쉽게 찾을 수 있을 것이라 믿는 사람들이 있었다. “Best of the Web 1994 -- Navigators“에는, “최상의 네비게이션 서비스는 웹 상의 거의 모든 것을 쉽게 찾을 수 있도록 해줘야 한다 (일단 모든 데이터가 입력되고 나면)” 라고 실려있었던 것이다. 하지만 1997년의 웹은 그 전과 전혀 다른 양상이다. 최근 검색엔진을 사용해 본 사람이라면 누구나 인덱스 자체의 완전성이 검색 결과의 품질을 결정하는 유일한 요소가 아님을 얘기할 것이다. “쓰레기 검색 결과“(Junk results)가 원하는 결과를 압도해버리는 것은 드문 일이 아니다. ... (중략) ... 이것의 주된 원인은, 문서의 수는 기하급수적으로 증가해온 반면 **사용자가 문서를 찾는 능력**(강조-인용자)은 그렇지 못했기 때문이다. 사람들은 여전히 검색 결과 중 수십 개 정도만을 보려 한다. 그러므로 문서 모음의 크기가 증가함에 따라 대단히 높은 정확도(precision)를 가진 툴이 필요해진다.

사실 우리는, “관계되는“이라는 단어의 의미를 아주 높은 수준으로 규정하려 했는데 그 이유는 ‘약간’ 관계되는 문서들만 해도 수만 개 이상이기 때문이다. 극도로 높은 수준의 정확도는 심지어 재현율(recall rate)을 희생해서라도 달성해야 할 만큼 중요하다.(Brin and Page, 1998: p.108)

이와 같은 ‘검색 품질’의 목표는 구글이 어떤 행위자들을 동맹으로 끌어들이길 원했는가를 보여준다. 만약 구글이 끌어들이려 한 행위자들이 자신의 글에서 사용한 어휘를 검색 키워드로 입력했을 때 자신의 문서가 반드시 검색 결과에 나타나는 재현율(recall rate)²⁸을 검색엔진의 가장 중요한 질적 지표로 생각하는 ‘텍스트의 저자’였다면, 구글의 도전과제는 정확도(precision)가 아니라 재현율이 되었을 것이다.(Langville and Meyer, 2011: p.8) 그러나 ‘페이지랭크’는 ‘텍스트의 저자’를 동맹에서 잃더라도 더 정확한 결과를 찾는 ‘독자’와의 동맹을 선택했다. 그런데 한가지 주목할 것은 구글 창업자의 논문에서 사용되는 ‘정확도’도 전통적 정보 검색엔진 기술의 평가 방법론에서 말하는 ‘정확도’가 아니라는

²⁸ ‘재현율’ (recall reate)은 검색 대상 안에 포함된 관련 문서의 전체 중, 검색 엔진이 검색 결과에 포함시킨 건수의 비율을 의미하며, ‘정확도’ (precision)과 함께 검색 엔진의 성능 평가에서 가장 중요하게 사용되는 성능 지표다.(참조 : Precision and recall, Wikipedia, https://en.wikipedia.org/wiki/Precision_and_recall, 2016.05.16 Retrieved)

점이다.²⁹ 이는 키워드 기반 검색엔진들이 최소한 검색결과에 나온 문서들의 본문 중에 해당 질의어를 포함하는 관련성을 보장하는 것과 가장 크게 다른 점이었다. 또한 구글의 검색엔진은 관련 문서를 최대한 많이 찾으려고 하는 순간 ‘정확도’가 크게 떨어지는 문제를 안고 있었다. ‘하이퍼링크’의 생성자가 오류를 범하여 앵커텍스트와 전혀 관련이 없는 문서를 링크할 경우에도 구글의 검색 결과는 이를 포함할 수 밖에 없었기 때문이다.

하지만 구글이 원한 동맹의 대상들은 많은 관련 문서를 찾기를 원하지 않았고 검색결과 첫 페이지에 나타난 가장 관련성이 높은 페이지 몇 개를 원하고 있었다.(Singhal and Kaszkiel, 2001) 이것은 ‘텍스트의 저자’를 흉내 내려 한 전통적인 정보검색 기술로서는 받아들일 수 없는 목표였지만, 웹 페이지의 독자들이 원하던 바로 그 ‘정확성’이었다. 그럼에도 구글은 전통적 검색엔진의 평가지표인 ‘재현율’과 ‘정확도’를 그대로 사용했는데, 다른 용어를 사용하여 전통적 검색엔진과는 다른 동맹 대상을 발견하기 보다는 기존 용어의 의미를 변형시킴으로써, 검색

²⁹ ‘정확도’ (precision)는 ‘관련있는 검색결과 개수/검색결과 개수’로 정의되는데, 일반적으로 분모가 되는 ‘검색결과 개수’를 10개, 20개, 100개, 1,000개 등으로 제한하여 사용한다. 따라서 상위 몇 개의 검색 결과를 기준으로 계산할 것인가에 따라 의미가 크게 변하는 지표다. 구글의 창업자들과 링크 기반 검색엔진의 옹호자들은 “사용자들의 이용 행태”를 기준으로 볼 때 전통적 평가방법에서 사용된 상위 1,000개의 결과에서의 정확도는 적절치 않다고 주장한다.(Singhal and Kaszkiel 2001)

엔진의 동맹 대상들의 목표를 변경시켜 끌어들이려 한 것이다. 구글은 스스로 전통적 검색엔진의 평가 기준을 충족시킴과 함께 자신이 제시한 평가 기준에 대한 변화된 정의를 받아들이는 행위자를 충분히 증식시킨 후에야 혁신적인 ‘검색엔진’이 될 수 있었다.

1997년 백럽 크롤러가 ‘검색엔진’이 되기 위해 제기했던 문제제기는 이것이었다. 대학 안에서 작동하는 백럽 크롤러가 아닌 상용화된 구글 검색엔진이 되기 위해 동맹을 맺어야 할 행위자들은 누구인가? 백럽 크롤러의 ‘페이지랭크’가 다른 검색엔진과 연결되어 있던 행위자들에게 ‘상용 검색엔진의 검색 결과’처럼 보이려면 보이기 위해서는 무엇이 필요한가? 만약 그렇게 받아들여지는 것이 거부된다면 어떻게 해야 할 것인가?

초기 백럽은 마치 초기의 웹 검색엔진들이 메타태그만을 인덱싱한 것과 비슷하게 ‘앵커 텍스트’와 ‘제목’만을 인덱싱하고 있었는데,³⁰ 이미 상용 검색엔진 사용자들은 전문 인덱싱(full text indexing)을 하는 알타비스타와 같은 검색엔진을 사용하는 경험에 익숙해져 있었다. 하지만 링크 분석만으로 검색 결과를 만드는 페이지랭크에게 전문 검색이 검색

³⁰ 구글이 전문 색인을 하기 시작한 것은 1997년인 것으로 보이지만, 1999년까지도 ‘본문의 내용’을 이용한 검색 품질 개선의 효과에 대해서는 회의적이었다. Page(1999)에서는 PageRank의 장점으로 ‘전문 인덱싱하는 것에 비해 비용이 매우 적게 든다’는 것을 제시하고 있다.

결과의 품질을 높여주는가? 두 창업자는 그렇지 않다는 것을 알고 있었지만, 구글이 본문 전체에 대한 색인을 만드는 것은 상용 검색엔진이 되기 위한 일종의 제식(ritual)이었다.³¹ 심지어 구글의 창업자들은 구글의 검색엔진 알고리즘인 ‘페이지랭크’와 ‘전문 검색’을 서로 다른 두 개의 검색엔진 접근법으로 생각했다. 하지만 검색엔진이 되겠다고 결정한 순간, 백럽의 랭킹 목록을 알타비스타의 ‘검색 결과처럼’ 보이기 위해 두 개의 접근법을 통합해야 했다.³² 이 문제는 구글에게 새로운 문제를 야기시켰다.

³¹ 구글은 검색엔진으로서의 두 가지 형태를 염두에 두었던 것으로 보인다. 하나는 제목만을 인덱싱하고 페이지랭크를 계산하여 비용이 매우 낮은 검색 엔진이고, 다른 하나는 다른 상용 검색엔진과 같이 전문(Full Text) 색인을 하는 대규모 시스템이면서 페이지랭크를 활용하여 품질을 개선한 검색 엔진이다. 하지만 래리 페이지의 ‘페이지랭크’ 관련 논문에서 표현된 내용은, 상용 검색엔진의 조건을 전문 검색으로 보는 시각을 드러낸다. (Page et al., 1999).

³² 구글은 스스로 검색엔진이 되기 위해 필요한 것들을 1998년에 발표한 논문에서 ‘Future works’로 표현하고 있다. 이 논문은 적극적으로 구글의 검색 결과를 알타비스타의 검색 결과와 비교하는 설명 방식을 취하고 있는 것으로 보아, 검색엔진으로서의 자기 정체성이 어느 정도 형성되어 있었던 것으로 추정된다. 다만 이 시기 이후에 발표된 래리 페이지의 논문(Page 1999)에서도 구글은 아직 전문(Full text) 검색을 제공하는 것에 대해서는 비용이 높다는 점에서 부정적이었다는 점에서 여기에 표현된 것들이 어떻게 구현될 것인가에 대해서는 명확한 대안을 갖고 있지는 않았던 것으로 보인다. (Brin and Page, 1998: p121)

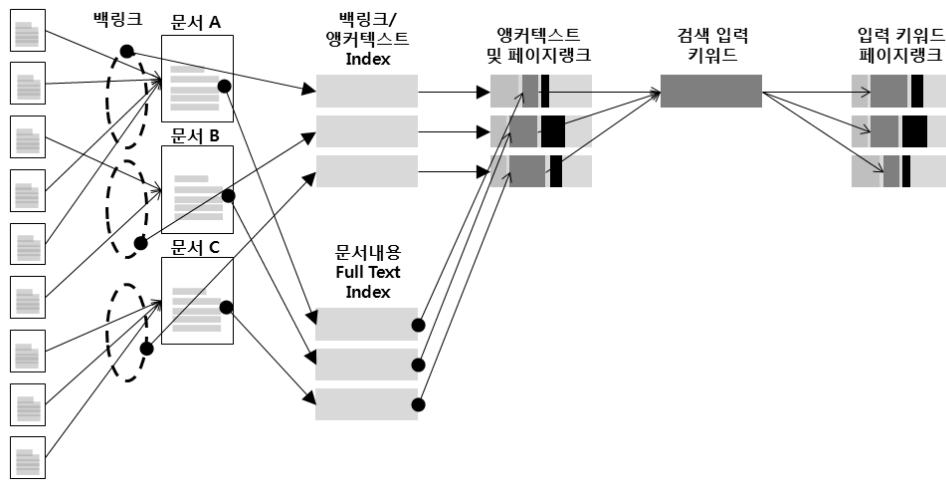


그림 11 전문 검색엔진(Full Text Search Engine)으로서 구글의 랭크병합

가장 중요한 것은 전문 검색을 할 경우, 검색 결과를 만드는데 있어 ‘페이지랭크’에 의해 계산된 웹 문서 중요도와 본문으로부터 만들어진 검색 키워드와의 유사도의 비중을 어떻게 보고 검색 결과를 만들 것인가의 문제였다. 단적으로 만약 검색 사용자들이 다른 상용 검색엔진에서와 같은 기대로 두 개의 단어를 입력한다면, 본문에 두 개의 단어를 모두 포함하고 있는 랭킹이 매우 낮은 페이지를 검색 결과의 상위에 올려야 할 것인지, 아니면 둘 중 하나만 포함하더라도 ‘페이지랭크’ 값이 큰 페이지로의 링크를 상위에 올려야 할 것인지를 결정해야 한다는 것이다. 이는 결국 ‘본문의 내용’과 ‘링크 분석’ 중에서 어떤 것을 얼마나 선택해야 하는가의 문제였다.(Dwork et al., 2001) 이 문제가 쉽지 않

은 문제라는 사실은 구글의 두 창업자도 분명하게 이해하고 있었다. 래리 페이지는 이 문제를 ‘랭크 병합’ (Rank Merging)이라고 불렀는데³³, 그런 이유로 구글을 ‘랭크 병합’을 사용하는 검색엔진이라고 지칭했다. 즉, ‘본문의 내용’에 대한 점수와 ‘페이지랭크’를 결합하여 검색 결과를 제공한다는 의미였다.(Page, 1999) 이는 구글 검색엔진이 ‘본문을 작성하여 의미를 담으려 한 저자’와 ‘다른 텍스트를 읽고 하이퍼링크를 남긴 독자’라는 서로 질적으로 완전히 다른 행위자를 동시에 대변하려는 이중적 연결을 시도하고 있음을 뜻했다. 구글의 랭크 병합이 ‘얼마나 올바른가’는 그것이 검색 결과의 품질을 떨어뜨리는가 아닌가가 아니라, 이를 통해 ‘얼마나 많은’ 행위자를 자신의 네트워크 안으로 끌어들이고 있는가에 의해 결정되었다. 구글은 ‘본문을 작성하여 의미를 담으려 한 저자’에 대해 ‘배제’라는 길을 택하지 않고 ‘검색 품질의 희생’이라는 길을 택한 것이다.

또 다른 난점은, 백랍이 짧은 문구의 앵커텍스트만을 처리하고 있었기 때문에 장문의 본문을 처리하기 위한 텍스트 처리 기술이 없었던 것에서 비롯된 것이었다. 당시 구글은 매우 단순한 텍스트 처리를 위해 N-gram을 사용하고 있었는데, 이것은 문서를 의미 단위로 분절하여 형태소를 분

³³ Dwork(2001) 등의 후속 연구에서는 ‘Rank Aggregation’으로 표현되었는데, 이는 복수의 랭킹을 통합하는 것 뿐 아니라, 스팸 필터 등을 적용하면서도 검색 품질의 훼손을 최소화하는 것을 목표로 하는 연구에 사용된 용어다.

석하는 다른 텍스트 검색엔진에 비해 연산 측면과 다국어 처리 측면에서 이점이 있기는 했지만, 다른 상용 검색엔진이 제공하고 있는 ‘형태소 분석’ (stemming)에 의한 동일어근 검색이나 ‘st**’ 같은 검색어를 입력하면 ‘star’, ‘stay’, ‘stun’ 같은 어휘들을 동시에 검색해주는 ‘철자 확장 검색’ (Wildcard search)과 같이 복잡한 텍스트 처리가 어려운 문제를 안고 있었다. N-gram 색인을 이용하여 상용 검색엔진 수준의 기능들이 구현된 사례가 없었기 때문이었다.³⁴ 또한 검색 키워드에 대한 ‘AND’, ‘OR’ 와 같은 ‘부울리안’ (Boolean) 논리연산 조건검색들을 제공하는 것도 구글에게는 골치 아픈 문제였다. 만약 검색 사용자가 < ‘구글’ and ‘검색엔진’ >을 검색 키워드로 입력한다면, 다른 상용 검색엔진들은 본문 안에 ‘구글’ 이라는 색인과 ‘검색엔진’ 이라는 색인을 함께 포함하고 있는 웹 페이지를 검색 결과로 표시할 수 있었지만, ‘페이지랭크’ 와 본문 내 키워드 매칭의 부울리안 연산 값을 어느 정도로 검색 랭킹에 반영해야 하는지 알 수 없었기 때문에 이런 간단한 논리 연산자의 적용도 단순한 문제가 아니었다.

구글이 ‘검색엔진’ 이 된다는 것은 더 이상 ‘페이지랭크’ 라는 기

³⁴ 그런 이유 때문에 구글이 형태소 분석에 의한 동일어근 검색을 제공할 수 있었던 것은 2003년이 되어서야 가능했다. “Stemming“. *Wikipedia*. 2015 [updated 04:16, 22 September 2015; cited 23 September 2015]. <https://en.wikipedia.org/wiki/Stemming>

술이 번역의 중심이 될 수 없으며, ‘구글 검색엔진’이라는 기업으로 블랙박스화된 행위자의 행위자네트워크가 검색 사용자뿐 아니라 사업적 목표를 가진 행위자들을 동맹 안으로 끌어들여 확장된다는 것을 의미했다. 구글은 ‘검색엔진’이 되기 위해, ‘검색엔진’의 행위자네트워크 안에서 생존 가능한 행위자가 되기 위해 목표를 변형하고 있었다. 구글은 스스로 하이브리드가 되고 있었다.

6. 새로운 하이브리드의 증식 : 구글의 위기와 새로운 번역의 전략

2001년이 되자 구글이 ‘페이지랭크’를 통해 대표하고 있었던 ‘하이퍼링크’를 만드는 독자는 빠르게 오염되기 시작했다. ‘하이퍼링크’를 만드는 독자의 자리에는 기계에 의해 하이퍼링크를 생성하는 기업들이 들어서기 시작했다. 하나는 닷컴기업들이었고, 다른 하나는 링크 분석 기술로 재무장한 ‘웹 스파머’들이었다. 닷컴 기업들은 사람을 대신해서 자신들이 유용할 것이라고 생각하는 많은 링크를 생성하고 있었고, 이른바 ‘링크 바밍(Link bombing)’이라 불리는 웹 스팸밍 알고리즘은 빠르

게 수많은 ‘하이퍼링크’로 연결된 웹 문서들을 쏟아내기 시작했다.(Metaxas and DeStefano, 2005) 이들은 모두 구글과 마찬가지로 ‘하이퍼링크를 만드는 사람들’을 흉내 내고 있었다. 구글은 이들에 대해서도 페이지랭크라는 질서를 부여하며 번역해냈는데, 이는 한편으로는 구글을 확장시켰지만 다른 한편으로는 구글을 야후!와 같은 위기로 끌고 가고 있었다. 닷컴기업들의 웹 페이지는 제후를 통해 대량으로 링크를 교환하여 ‘페이지랭크’가 높은 웹 페이지들을 만들어 냈고, ‘링크 바머’(Link Bomber)들은 ‘페이지랭크’ 알고리즘에서 높은 랭킹을 만들 수 있는 웹 링크 구조들을 생성하여 ‘페이지랭크’ 자체를 ‘역공학(reverse engineering)’으로 번역해 냄으로써 구글을 길들이려고 ‘힘겨루기’를 하고 있었다. 이들은 구글을 위협하는 하이브리드였지만, 동시에 구글의 질서를 웹 전반으로 확산시켰다. 구글은 야후!와 유사한 선택의 기로에 놓여 있었다. 동맹을 맺음으로써 이들의 증식을 성장의 매개로 활용할 것인가? 아니면 이들을 동맹에서 배제함으로써 번역되지 않은 행위자들을 새롭게 규합할 새로운 동맹의 탄생을 용납할 것인가? 만약 동맹을 성사시키려 한다면, 이들에게 어떤 OPP를 제시할 수 있을 것인가?

6.1 N-gram : 블랙박스 안의 행위자

구글의 두 창업자의 최초 아이디어가 ‘페이지랭크’ 이기는 했지만,

그것을 사람들이 사용할 수 있는 서비스로 만들기 위해서 반드시 필요한 것 중 하나는 ‘앵커텍스트’를 처리하는 알고리즘이었다. 하지만 ‘페이지랭크’는 본문을 검색하는 검색엔진이 아니었기 때문에 가장 간단한 텍스트 처리 알고리즘이 선택되었는데, 그것이 바로 N-gram 기술이었다. N-gram은 텍스트 검색 기술이 연구되던 초기에 활용되었지만, 곧 그 한계가 드러나 텍스트 검색에서는 대학의 연구 수준에서만 사용될 뿐 상용 기술로 발전될 가능성은 매우 낮게 평가된 기술이었다. 1997년 당시 이미 N-gram에 대한 연구가 이루어지는 분야는 그것은 광학 문자 인식(OCR: Optical Character Recognition)과 음성 인식 분야³⁵ 정도로 축소되어 있었다. 하지만 구글 초기 N-gram을 이용한 앵커텍스트의 색인화는 ‘페이지랭크’나 ‘백립 크롤러’에게 큰 문제가 되지 않았다. 왜냐하면 ‘페이지랭크’는 텍스트의 본문을 분석하는 기술이 아니었기 때문이다. 게다가 N-gram을 이용하여 만든 구글 검색엔진은 매우 훌륭하게 웹 문서들의 중요도를 계산해 냈고, 다른 어떤 텍스트 기반 검색엔진 보다 빠른 속도로 동맹을 늘여 나가고 있었다.

N-gram은 구글의 행위자네트워크를 확장시키는데 다른 장점도 가지고 있었다. N-gram은 ‘어휘(word)’와 같은 다른 인덱스와 비교했을 때

³⁵ “n-gram“, Wikipedia. 2015 [updated 10:13, 18 September 2015; cited 23 September 2015].<https://en.wikipedia.org/wiki/N-gram>

비교할 수 없이 빠른 탐색 속도를 제공하며, 규모가 큰 시스템을 만들 때도 매우 단순한 방식으로 시스템의 크기를 늘일 수 있기(scalability) 때문이다.³⁶ 이는 구글 창업자들이 웹 검색엔진이 갖춰야 하는 중요한 가치로 여겼던 ‘대용량(large scale)’ 시스템을 구현하는데 가장 중요한 해결책이었다.(Brin and Page, 1998) N-gram의 또 다른 장점은 ‘언어 독립성’이었다. 이는 문서가 어떤 언어로 작성되었는가와 무관하게 ‘형태소 분석’ 같은 복잡한 작업을 거치지 않고 문서를 색인할 수 있다는 뜻이다. 이 때문에 구글은 다른 검색엔진들이 대상 언어마다 각각에 대한 구문론적 지식을 검색엔진에 포함시켜 형태소 분석을 하는데 매우 많은 투자를 했던 것과 달리 투자에 대한 부담 없이 서비스 지역을 확장했다. 1999년에 상용서비스로 전환된 구글 검색은 2003년에는 이미 74개 언어로 된 문서에 대한 검색 서비스를 28개 언어로 된 사이트를 통해 제공하고 있었고,(예병일, 2003) 현재는 150개의 언어로 된 문서의 검색을 지원하고 있다.³⁷ 게다가 당시 많은 텍스트 기반 검색엔진들이 달성하려 했던 ‘언어 간 정보 검색’ (Cross language Information Retrieval)도,(Ballesteros and Croft, 1997) ‘페이지랭크’ 에게는 당연한 기능이었다. ‘페이지랭크’ 는 이로써 컴퓨팅 효율적 측면에서뿐 아니라 언어적 측면에서도 최

³⁶ Ibid.

³⁷ 구글 연혁, Google, 2015, <http://www.google.com/about/company/history/>

초로 ‘인터넷 규모(internet-wide scale)’로 확장 가능한 행위자네트워크가 되었다. 하지만 N-gram의 이러한 행위능력은 ‘페이지랭크’와의 결합을 통해서만 작동되는 것이었고, 반대로 낮은 컴퓨팅 비용으로 언어 장벽을 넘어 검색 랭킹을 제공하는 ‘페이지랭크’의 행위능력 또한 N-gram과의 결합을 통해 얻어진 하이브리드로서의 능력이었다. 구글은 임의의 언어로 웹 페이지를 만드는 저자 하이브리드들과 임의의 언어로 검색하는 독자 하이브리드들을 빠르게 동맹 안으로 끌어들여 증식시킴으로써 야후!와 같은 디렉토리를 갖지 못한 언어권에서 웹 자체를 성장시키고 있었다.

하지만 구글 검색 서비스의 초기에 N-gram에 주목한 사람은 거의 없었다. 구글 검색에서 검색 결과의 순서를 결정하는데 본문의 내용이 미치는 영향이 제한적이어서,(Page, 1999) N-gram의 역할은 매우 미미했기 때문이다. 하지만 구글이 광고주들을 자신의 네트워크로 끌어들이는 과정에서 N-gram은 페이지랭크 보다 더 중요한 행위자로 떠올랐다. 광고주들은 사용자들이 특정한 키워드를 검색할 때 자신들의 광고가 노출되기를 원했고, 사용자가 입력한 질의와 검색 대상인 웹 페이지에 어떤 광고주가 구매한 키워드가 포함되어 있는지를 판단하는 것은 N-gram의 몫이었기 때문이다. 두 창업자는 “광고를 기반으로 한 검색엔진은 편향될 수 있으며, 그 편향은 너무 미묘해서 전문가들도 이를 찾아내기는 어렵다”(Brin

and Page, 1998: p114)고 생각하여 이에 부정적이었지만, 에릭 슈미트(Eric Schmidt)가 CEO에 임명되자 변화는 빠르게 진행되었다. N-gram과 페이지랭크의 새로운 하이브리드는 각자 독립적으로는 완성시킬 수 없었던 동맹의 확장을 완성시키고 있었다. 그럼에도 N-gram은 주목 받지 않았는데, 그 이유는 구글의 검색 알고리즘은 몇 번 “페이지랭크가 죽었다!”는 논란에 휘말렸음에도(Langville and Meyer, 2011) 여전히 ‘페이지랭크’라는 블랙박스를 열어야 할 정도의 위기를 맞지 않았기 때문이었다.

아미트 싱할은 N-gram을 이용하여 키워드 기반 검색엔진들이 구현한 거의 모든 텍스트 인덱싱 기술들을 차례로 구현해 나갔다. ‘형태소 분석’이나 ‘부울리안 검색’, ‘철자 확장 검색’(wildcard search), ‘의미 추출’(text mining)과 같이 전통적 검색엔진에서 사용되던 기능들이 그의 팀에 의해 N-gram 색인을 통해 구현되어 나갔다.³⁸ 그와 그의 팀은 이 과정을 통해 ‘N-gram의 흑마술사’라는 호칭을 얻었는데,(Levy, 2010) 이것이 ‘흑마술’이라고 불린 이유는 명백했다. N-gram 알고리즘

³⁸ 구글의 N-gram 분야의 독보적 기술력은 Research at Google 사이트에서 확인되는데, 구글이 정확도가 높은 음성 인식이나 도서관 장서 스캐닝 같은 기술을 개발할 수 있었던 이유는 바로 N-gram 기술 때문이었다. 구글이 2015년 9월까지 발표한 N-gram 관련 논문은 158건이고 그 중 음성인식 관련 논문은 135건이다. 이는 구글 학술검색에서 검색되는 N-gram 관련 논문과 특허의 대부분이 2000년 이전의 것들인 것과 확연하게 비교되는 현상이다.

들은 이중적 의미로 ‘인간의 눈에는 보이지 않는’ 것을 보는 행위자였기 때문이다. 하나는 N-gram이 인간을 위해 만들어진 정보 단위가 아니라는 의미였고, 다른 하나는 그것이 ‘페이지랭크’라는 블랙박스 안에서 보이지 않게 작동한다는 의미였다. 싱할의 팀이 N-gram을 이용하여 텍스트를 처리하는 최초의 컴퓨터 알고리즘의 발명자는 아니었지만, ‘페이지랭크’라는 행위자네트워크 안으로 이식된 N-gram은 다른 연구자들이 다루던 N-gram과는 전혀 다른 의미와 행위 능력을 갖게 되었다.

사실 싱할의 팀이 N-gram을 이용하여 의미를 재구성하는 방식은 원리적으로 새로운 것은 아니었다. 이미 N-gram은 광학문자인식이나 음성 인식과 같은 자연어 처리 기술에서 구글이 사용하게 되는 확률적 모델과 같은 일군의 요소 기술들을 갖춘 접근법으로 어느 정도의 성과를 보여주고 있었다.(Suen, 1979) 그러나 텍스트 처리 분야에서 연구자들이 그 방식을 택하지 않았던 이유는 그것이 ‘비효율적’이기 때문이었다. 하지만 싱할의 팀에게는 그것이 ‘효율적인’ 방법이 되었는데, 그 이유는 싱할의 팀이 만들려는 행위자네트워크가 다른 텍스트 처리기술 연구자들이 만들려고 한 행위자네트워크와는 달랐기 때문이었다. 다른 연구자들의 텍스트 연구는 인간을 대신하여 높은 정확도로 의미를 읽을 수 있는 범용적 기술을 확보하는 목표를 가진 것이었지만 싱할의 팀은 이미 ‘페이지랭크’에 의해 구축된 행위자네트워크를 조금만 다른 방향으로 확장시키

면 충족되는 것이었다.(Severyn, et al., 2014) N-gram 기반 텍스트 기술은 그런 이유로 구글에게는 충분히 ‘효율적인’ 기술이었다. 하지만 이러한 장점들이 설계자들의 ‘의도’ 안에 어느 정도 예견되었다 하더라도, 이들의 결합이 가져온 또 다른 결과들은 설계자들이 전혀 예상치 못한 것들이었다.(홍성욱, 2010) N-gram으로 중심을 이동시킨 ‘페이지랭크’는 고전적 ‘페이지랭크’ 그 자체보다도 더 빠르고 광범위하게 동맹을 확산시키는 OPP가 되고 있었다.

6.2 구글을 둘러싼 새로운 하이브리드의 중식과 위기

구글의 이러한 중심 이동은 구글에게 또 다른 방향에서의 활로를 제공했다. 2001년이 되자 구글은 ‘텍스트의 저자’를 훔쳐 내는 검색엔진들과의 경쟁에서 앞서 나가기 시작했다. 구글의 네트워크로 더 많은 사용자들이 연결되기 시작했는데, 이는 ‘하이퍼링크를 만드는 독자’를 훔쳐 내는 ‘페이지랭크’ 알고리즘이 웹 페이지를 찾는 사용자들의 OPP가 되는데 성공한 이유도 있지만 상당 부분은 ‘텍스트의 저자’를 훔쳐 내는 검색엔진의 실패에서 기인한 것이었다. 검색엔진의 실패는 검색엔진들이 디렉토리에 등록되지 않은 웹 페이지의 저자와 임의의 웹 페이지를 찾는 독자들을 배제하고 정제된 콘텐츠를 전달하는 ‘포털’ (Portal)이 되기로 목표를 변경하면서 더 급속히 진행되었다.(Hu, 2001) 이러한 검색엔

진의 방향전환을 이끈 웹 스팸머들은 ‘텍스트의 저자’를 흉내 내는 검색엔진들이 무엇을 검색 대상 텍스트의 ‘의미’라고 연산해 낼 것인지를 알고 검색엔진들을 오독하게 만들 수 있는 방법, 즉 텍스트 기반 검색엔진과의 ‘대화 방법’을 찾아냄으로써 검색엔진을 공격하고 있었다. 특정한 키워드를 매우 많이 포함한 웹 페이지를 자동으로 만들어 내거나(keyword stuffing), 메타 태그에 해당 키워드를 삽입하거나(meta-tag stuffing) 본문 내에 보이지 않게 키워드를 끼워 넣는(Hidden or invisible text) 것과 같은 다양한 방법으로 검색엔진들을 속였다.(Metaxas and DeStefano, 2005) 이들의 목표는 더 많은 검색엔진 사용자들을 자신의 고객 사이트로 끌어들이는 것이었다. 검색엔진이 텍스트의 독자를 포괄적으로 흉내 내야 하는 것에 반해 웹 스팸머들은 그러한 검색엔진들을 속이는 파편적 패턴을 찾기만 하면 되었기 때문에, 이를 막으려는 검색엔진은 웹 스팸머에 비해 너무 많은 자원이 들어가는 싸움이었다. 하지만 2001년 알타비스타가 검색엔진 주위에 몰려든 행위자들에 대한 배제를 위해 야후!가 선택한 길로 방향을 바꾸었을 때, 웹 스팸머들 역시 더 이상 검색엔진을 매개로 증식할 수 없는 상태로 전락했다.

이렇게 되자 웹 스팸머들은 링크 기반 검색엔진인 구글에 대한 ‘역공학’(reverse engineering)을 시도했다. 이들은 자신들의 스팸밍 시스템이 구글의 페이지랭크 알고리즘이 특정 페이지의 중요도를 오독하도록

유도할 수 있다면, 텍스트 검색엔진을 공격하여 거두었던 수익을 구글을 통해서도 달성할 수 있을 것이라 기대했다. 이러한 스팸밍 시스템은 ‘링크 팜(link farm)’이라고 불리었는데, 그것이 지향한 것은 앞서 설명한 ‘키워드 스테핑(keyword stuffing)’과 유사해 보이면서도 중요한 차이를 갖는다. 그 차이란 단순히 하나는 키워드를 이용하고 다른 하나는 링크를 이용한다는 것이 아니었다. ‘키워드 스테핑’ 같은 웹 스팸밍이 검색엔진과 마찬가지로 현실의 텍스트 독자를 훔쳐 내는 반면, ‘링크 팜’과 같이 링크를 읽는 인간 독자는 현실에 존재하지 않는다는 점이다. 전통적 웹 스팸머는 직관적으로도 텍스트 시스템의 행위자네트워크에서 비롯된 인간 행위자의 계보 안에서 태어난 하이브리드이지만, 링크 스팸머는 직관적으로 더 이상 인간 행위자와의 관련성을 찾기 힘든 오직 웹 안에만 존재하는 하이브리드였다.³⁹

6.3 구글의 새로운 번역 전략 : N-gram 필터

싱할의 팀은 이러한 링크 스팸밍에 대응하기 위해 ‘페이지랭크’를 속이기 위해 만들어진 웹 페이지들을 골라내는 알고리즘을 개발해야만

³⁹ 웹 스팸에 대한 일반화된 대응법에 대한 연구(Gyöngyi et al., 2004)도 시도되었는데, 구글 창업자들의 지도교수인 가르시아 몰리나(Garcia Molina) 등도 이 주제에 대한 논문을 발표했다.

했다. 만약 이것에 실패한다면 구글 검색엔진은 앞서 알타비스타가 그러했던 것처럼 웹 문서를 찾는 행위자들의 OPP의 지위를 유지할 수 없게 될 것이 분명했다. 구글은 링크 스팸머들이 주로 사용하는 페이지랭크를 높이는 링크 연결기법을 수집하여 그들의 패턴을 찾아냄으로써 그러한 패턴에 대한 필터링을 시도했다.⁴⁰ 그 패턴들은 인간 독자에게는 읽혀지지 않았지만, 컴퓨터 저자의 ‘하이퍼링크’ 문체(style)로 웹 링크들의 집합체 내에 기록되어 있었다. 싱할의 팀은 이를 위해 새로운 종류의 ‘랭크 병합’을 시도했는데, 그것은 N-gram을 응용한 스팸필터를 이용한 랭킹과 링크 분석을 통한 ‘페이지랭크’의 병합이었다. N-gram은 텍스트를 넘어서 더 범용적인 행위자가 되기 시작했다.(Taylor et al., 2007) 그리고 2002년 추수감사절을 앞둔 쇼핑 시즌이 개막되기 직전에 스팸머들에 대한 공격을 감행했다. 이른바 ‘구글 댄스(Google Dance)’라고 불리는 이날의 ‘공격’은 약 1달에 걸쳐 진행되었는데, 링크 스팸머들은 자신들의 고객인 온라인 상점들이 구글의 공격에 의해 심각한 매출 손실을 입고 링크 스팸머들과의 계약을 해지함으로써 자신들의 네트워크를 더 이상 유지할 수 없게 되었다.(Auletta, 2010) 그 후로도 구글은 링크 스팸머들의 새로운 공격을 무산시키는 알고리즘 업데이트를 단행했

⁴⁰ Google, ‘구글 검색 속으로 - 스팸차단’, www.google.com, <https://www.google.com/insidesearch/howsearchworks/fighting-spam.html>, 2016.03.23 Retrieved

다.(Dean 2015) 링크 스팸머들은 구글의 가이드라인에 따라 고객의 웹사이트를 정상적으로 높은 페이지랭크를 얻을 수 있도록 도와주는 검색엔진 최적화(Search Engine Optimization : SEO) 컨설턴트로 목표를 전환해야 했다. 구글은 ‘배제’라는 번역의 방식을 성공적으로 작동시킴으로써, 스팸머들을 길들였고 동맹으로 끌어들이며 동원 가능한 행위자로 만들어냈다.

링크 스팸머에 대한 N-gram을 OPP로 한 번역 전략은 성공적으로 작동했다. 그러나 그 과정에서 N-gram은 외부로 드러나지 않았다. 여전히 구글의 핵심 알고리즘은 ‘페이지랭크’로 불렸다. 그러나 ‘페이지랭크’라 불린 블랙박스 안에서 N-gram 기반의 수많은 필터들은 ‘페이지랭크’에게 새로운 행위능력을 제공하고 있었다. 구글 검색엔진은 인간의 눈에는 보이지 않는 비인간 하이브리드들과 동맹을 확대하면서 인간의 의미론 보다 더 세분화된 의미론을 구성하기 시작했다. 하지만 동시에 구글이 구축한 ‘정화의 축’으로 환원될 수 없는 성분을 가진 하이브리드들은 계속 증식되며 구글의 행위자네트워크 안으로 들어와 구글을 위기로 끌고 가고 있었다. 악성코드, 포르노 배포자, 저작권자, 검열(censorship)을 원하는 정부들, ‘잊혀질 권리’(right to be forgotten)와 같은 새로운 하이브리드들은 무시할 수 없는 속도로 증식되고 있었다.

6.4 새로운 하이브리드를 위한 대표 선출

구글은 2001년부터 진행한 대대적인 알고리즘 업데이트인 ‘구글 댄스’ (Google Dance)를 통해 웹 스팸머들을 제압하는데 성공하자(장중혁, 2014: p371), 다른 영역에까지 ‘심판’의 영역을 확대했다. 하나는 ‘포르노’였고, 다른 하나는 ‘악성 코드’였다. 이러한 심판을 수행한 것은 ‘필터(filter)’들이었는데, 이 필터들은 인간 독자를 위한 텍스트가 아닌 이미지나 영상, 프로그램 코드를 위한 것이었지만, ‘의미 없는 기호’들을 조합하여 연산에 의해 의미를 추출할 수 있도록 한다는 점에서 N-gram과 같았다. 포르노나 악성코드의 패턴은 N-gram과 같은 이진 데이터의 패턴으로 정의되었다. 웹 페이지 내의 텍스트와 이미지, 영상들을 ‘의미론적으로’ 판정하는 필터들에 의해 포르노를 포함한 페이지들은 검색 결과에 나타나지 않도록 번역되었다. 또한 필터들은 악성코드를 배포하는데 활용되는 웹 페이지를 찾아내기 위해서 인간 텍스트 독자가 읽을 수 없는 자바 스크립트(Java script)를 해석하여 그 코드가 컴퓨터 시스템에 제공하는 의미를 추출해 냈다.

2006년이 되자 구글은 포르노 관련 이미지와 동영상을 걸러내기 위해 포르노 콘텐츠들로부터 의미론적 벡터를 추출하여 학습시켜 임의의 이미지와 동영상의 포르노 여부를 판정하는 기술을 발표했다.(Rowley et al., 2006) 이는 텍스트가 아닌 이미지나 동영상, 사운드와 같은 내용을 읽어

들이는 의미론을 멀티미디어 콘텐츠를 주파수 도메인(Frequency Domain)에서 재해석하여 나타나는 특징점 벡터들로 구성한 것인데, 구글은 이 기술을 기반으로 2007년에 ‘세이프서치’라는 성인 콘텐츠 필터링 서비스를 출시했다. 하지만 이 기술 개발 과정의 연구 결과를 담은 논문에서, 구글 연구진은 성인 콘텐츠를 담고 있는 이미지를 얼마나 제대로 걸러내는가를 표시하는 성능 지표인 재현율(recall rate)을 50%로 할 경우 대략 10%의 비성인 콘텐츠가 잘못 필터링되고, 재현율을 90%로 할 경우 필터링된 이미지의 35%가 비성인 콘텐츠라고 밝혔다. 그럼에도 구글은 이 기술을 통해 자신이 지배하는 사유화 영역 안에서 권력작용을 구축하였는데, 이것이 비인간행위자를 매개로 한 ‘자동화된’ 권력작용을 충분히 정당화시킬 수 있을 정도로 성숙된 기술인가의 의문은 끝내 제기되지 않았다. 포르노와 악성코드를 배포하는 웹 페이지가 증가할수록 구글의 권력은 성장했다. 구글은 더 명시적으로 웹 페이지들을 차단했다.(허완, 2015) 구글에게 중요한 것은 기술의 완성도가 아니라, 네트워크 안의 행위자들이 구글의 권력행사에도 여전히 구글이 자신들을 대표하도록 투표할 것인가였다.

하지만 구글은 이러한 권력이 사적으로 작동되고 있다는 느낌을 주지는 않았다. ‘페이지랭크’라는 질서는 구글이 만든 질서였고 전적으로 사유화된 질서였지만, 구글은 자신의 권력이 사용자들로부터 공적 가치를

지킬 권한을 위임받은 것처럼 움직였다. 구글의 입장에서 스팸어나 악성 코드, 포르노 배포자들 혹은 감시자들이란 구글이 만든 세계 안으로 넘어 들어온 침입자였고, 이들이 '검색에 악영향'을 주며 '사이트 소유자에게도 부정적인 영향'을 준다고 판단한 구글이 선량한 행위자들을 보호하는 것이기도 했다.⁴¹ 사유화된 영역 안에서 무엇이 나쁜 행위인가를 결정하는 것은 전적으로 소유자의 권한이라는 것은, 구글에게는 의심할 여지가 없는 정의(justice)였다. 그러나 구글은 야후!나 알타비스타와 같이 나쁜 행위자들을 동맹에서 제외시키지 않고 구글의 OPP를 통과하도록 - 그 단죄가 성공하건 실패하건 - 길들임으로써, 선한 행위자들 뿐 아니라 나쁜 행위자들도 대표하기를 원했다. 그럼으로써 이들이 증식되고 있는 모든 곳으로 구글의 사적 영역을 확장해 나갔다.(이해성, 2007)

6.5 보일의 실험실과 진공펌프로서의 인터넷

인터넷은 공적 권력의 진공상태를 통해 새로운 공적인 문제와 이에 대한 사적 권력이 전세계적 규모로 발생할 수 있는 조건을 제공했다. 하지만 구글의 사적 권력은 이러한 문제들을 해결함으로써 권력을 확장한 것은 아니었다. 구글의 권력은 자신의 문제 해결 능력에 의해서가 아니라 새로운 문제를 일으키는 하이브리드들의 증식에 의해 확장되고 있었다.

⁴¹ ibid

구글은 사업의 초기부터 인터넷 상에서 발생하는 저작권과 사생활침해, 상표권 침해 등과 관련하여 각국의 법정에서 소송의 당사자가 되었다.⁴² 일부 소송은 이겼고 일부는 합의했고 일부는 패배했다. 그러나 그 결과가 무엇이건 구글은 법정에서 새로운 권력을 위임 받았는데, 구글은 각국의 법정이 내린 판결에 따라 자신의 사적 영역 안에서 활동하는 행위자들을 통제할 수 있는 권한을 부여 받았다.(허완, 2015)⁴³ 그런데 문제는 그 사적 영역이 새로운 하이브리드의 증식으로 계속 팽창하고 있다는 사실이다. 대개의 경우 구글의 기술 선택은 기술 기업이 할 수 있는 최선의 선택이며 어떤 권력도 인터넷 안에서 증식되는 하이브리드들에 대한 법률의 권력작용으로 구글보다 나은 기술적 대안을 제공할 가능성은 거의 없다. 구글이 인터넷을 국가권력에 대한 이상적인 진공 상태로 만들 수 있다면 구글은 국가들로부터 권한을 위임 받아 전세계를 아우르는 거의 유일한 권력이 될 수도 있다. 하지만 흡스가 보일의 진공펌프에 대해 반론했던 것과 마찬가지로 국가권력의 진공은 잘 만들어지지 않으며 인터넷 안으로 국가 권력이 새어 들어왔다. 반독점 규제 뿐 아니라 미디어 규제, 저작권, 과세 등의 법정을 통해 국가들은 인터넷 안에서 국가 권력에 지배되지 않는 사실들이 만들어지는 것이 위험한 일임을 보이려고 했

⁴² Perfect 10, Inc. v. Amazon.com, Inc.(Google Inc.), 연방제9항소법원 508 F.3d 1146(9th Cir. 2007)

⁴³ 구글, 애드워즈, 상표권 정책, <https://support.google.com/adwordspolicy/answer/6118?hl=ko>, 2016.05.16 Retrieved

다. 그러나 구글은 보일이 했던 것처럼 끝없이 새로운 필터를 추가하며 ‘페이지랭크’ 머신을 고쳐나갔고, 각 국가의 법정과 행정 규제는 보일의 실험실에 가두어진 신사들(gentlemen)이 보일이 질문한 내용 안에서만 답해야 했던 것처럼 인터넷과 구글이 제시한 프로토콜을 통해서만 선택해야 한다.(Latour, 1993: p.18)

인터넷은 홉스가 보일의 실험실과 진공펌프에 대해 우려했던 것과 같이 국가가 통제력을 행사할 수 없는 공간이 될 가능성이 있었다. 인터넷 안에서는 지상의 유일한 권력으로서의 국가의 충만론(plenism)은 거부되었고, 보일의 실험에서 진공펌프라는 비인간행위자를 길들일 수 있는 인간행위자만이 그 안의 진공에 대해 권력을 가질 수 있었던 것과 같이 오직 인터넷의 비인간행위자들과 협상할 수 있는 인간행위자만 권력을 가질 수 있다. 아마도 홉스였다면 국가들에게 인터넷을 비난하며 권력의 유일성이 해체되고 있다고 경고할 것이다. 인터넷은 새로운 전쟁을 예고하는 ‘권력의 공백’을 의미하기 때문이다. 또한 진공이 인정되면 그것이 자연철학으로 침투하여 권위를 분할할 것이라는 홉스의 걱정과 라투어의 해석처럼, 인터넷 안에서 만들어진 진공(vacuum)은 곧 인터넷 밖의 국가로 침투함으로써 국가의 권위를 분할하고 있었다.(Shapin and Schaffer, 1985; Latour, 1993: p.20)

하이브리드는 인터넷 밖에서도 증식되었다. 2011년 말 미국 하원이 발

의한 온라인 저작권 침해 금지 법안(SOPA: Stop Online Piracy Act)⁴⁴과 상원이 발의한 지식재산권 보호 법안(PIPA : Protect Intellectual Property Act)⁴⁵에 반대하는 캠페인을 통해 이를 무력화시켰다. 구글과 그 협력자들은 이 두 법안이 너무 포괄적이어서 결과적으로 ‘인터넷에 대한 감시’가 적법한 절차 없이 이루어질 가능성이 있다고 비판했다. 구글은 이 과정에서 자신들이 인터넷 밖에 있는 현실의 유권자들을 대표하는 것처럼 행동했다. 사람의 권리로서 ‘자유’라는 공적 가치는 의회가 아니라 구글에 의해 대변되고 있었다. 2012년 초 결국 이들은 이 두 법안의 상정을 철회시켰고, 미 의회에 강력한 로비스트 그룹을 보유하고 이 법안의 상정을 주도했던 미디어기업과 특허관련 기업들과의 정치적 경쟁에서 완승했다. 구글은 인터넷 안에서 미국 의회의 ‘나쁜 행위’를 판단하고 통제하는 권력을 인터넷 밖에서 획득한 것이다.(장중혁, 2014)

⁴⁴ 온라인 저작권 침해 금지 법안(영어: Stop Online Piracy Act, SOPA)은 2011년 10월 26일 미국 하원에서 제출된 법안으로, 인터넷 상에서 저작권이 침해된 경우에 대한 경우의 법적 조치를 확대하는 내용을 가지고 있다. 이 법안에서 가장 논란이 되는 부분은 법적 조치로 웹사이트의 접속 차단할 수 있다는 점이며, 이에 따라 인터넷 검열이라는 비판이 일어났다.(참조: “온라인 저작권 침해 금지 법안“, 위키피디아, 2016.05.17 Retrieved)

⁴⁵ 지식재산권 보호 법안(정식 명칭: 경제적 창의성에 대한 실질적 온라인 위협 및 지식재산권 침해 방지 법안, Preventing Real Online Threats to Economic Creativity and Theft of Intellectual Property Act of 2011, PROTECT IP Act, PIPA)은 미국 정부와 저작권 소유자가 저작권을 침해하거나 위조품을 만드는 웹사이트(특히 미국 이외의 지역에 등록된 웹사이트)로의 접속을 제한하게 하는 수단을 명시한 법안이다. (참조: “지적재산권 보호 법안“, 위키피디아, 2016.05.17 Retrieved)

구글의 사적 영역은 국가 간의 질서로도 확장되었다. 2012년 구글은 전세계 구글 사용자들에게 ‘인터넷의 자유’를 주제로 하는 캠페인을 벌이고 있었다. 하지만 그 캠페인의 실제 목표는 그 해 말에 두바이에서 열리는 WCIT⁴⁶-12에서 아프리카와 중동, 중국, 러시아 등이 상정하여 추진한 인터넷 거버넌스를 UN 산하기구인 국제전기통신연합⁴⁷ 아래로 옮겨야 한다는 내용을 담은 국제전기통신규칙(ITRs : International Telecommunication Regulations)에 반대하는 것이었다. 구글은 결과적으로 인터넷에 대한 사법적 시스템이 ‘국가와 국가들 사이의 질서’로 만들어지는 것을 반대한 것인데, 구글은 인터넷이 계속 ‘헌법 없는’ 공간으로 유지되기를 바라고 있었다.⁴⁸

구글은 법률이 아니라 국가와의 협상을 통해 권한을 위임 받기 원했다. 이는 각국 정부의 저작권 침해(Peterson, 2014)와 프라이버시 침해에 대한 외교적 요청과 정치적 요구 등을 구글이 수용하면서 더 빠르게 진

⁴⁶ WCIT(World Conference on International Telecommunications)는 유엔산하 전문기구인 ITU의 회의체 중 하나로서, 국제 통신 효율을 포함하여 국가 간 통신 규약에 대해 논의하고 조약의 효력을 갖는 협정서를 체결하는 회의체이다.

⁴⁷ 국제전기통신연합(國際電氣通信聯合, ITU, International Telecommunication Union)은 유엔의 산하 기구로 전기 통신의 개선과 효율적인 사용을 위해 국제 협력을 증진하고 전기 통신 업무의 능력 향상, 이용 증대 및 보급의 확대를 위해 기술적 수단의 발달과 효율적 운용을 목적으로 하는 정부간 국제 기구이다.

⁴⁸ 구글은 이러한 자신들의 정치적 활동에 따로 이름을 붙이지 않았지만, 이것은 20세기 후반과 21세기 초 거대기업들이 정치적 목소리를 내는 창구로 활용한 ‘기업의 사회적 책임’(Corporate Social Responsibility)이라는 트렌드 안에서 해석될 수도 있다.(Crouch 2011)

행 되었는데, 이는 일반적으로 ‘내용검열(censorship)’이라는 용어로 불려왔던 것이었다. ‘내용검열’은 의미작용이 어떻게 권력작용이 되는가를 보여주는 좋은 대상인데, ‘내용검열’은 단순히 내용 그 자체가 아니라 접속자의 위치와 내용, 그리고 ‘하이퍼링크’가 향하는 대상과 같이 이중적 행위자들의 결합으로 이루어진다. 전통적인 ‘페이지랭크’에는 검색자의 국적이 없지만 ‘내용검열’이 포함된 ‘페이지랭크’는 검색자의 국적에도 연루되었다.⁴⁹ 이는 구글이 전세계 대부분의 지역에서 해당 지역의 고유한 규제 기준을 인터넷 안에서 가장 잘 대표할 수 있는 권력 작용을 공급할 수 있는 국가의 대리인이 되고 있다는 의미였다. 구글은 경찰이 아니었지만 자신이 만들어낸 거대한 사적 영역 안에서 국가의 법률적 위임을 받은 사법적 대리인으로 행동했다. 인간-구글 하이브리드들이 증가하고 이들이 구글의 서비스에 의존하게 될수록 혹은 이들이 더 많은 문제를 일으킬 수록 구글은 국가로부터 더 많은 권한을 위임 받았다.

이제 구글의 검색엔진은 기술로서의 ‘페이지랭크’가 아니라 구글이 다양한 랭킹과 필터링 기준을 수용하여 ‘연산 가능한’ 논리를 기초로 검색 순위를 만드는 ‘길잡이’ (Directory) 사이트에 가깝게 변화했

⁴⁹ “Censorship by Google“, *Wikipedia*. 2015 [updated 21:03, 4 September 2015; cited 23 September 2015].https://en.wikipedia.org/wiki/Censorship_by_Google

다.(Dean, 2015) 나아가 ‘페이지랭크’는 인간행위자들에게 뿐 아니라 스팸머의 시스템이나 악성코드, 각국의 인터넷 규제 정책들에게도 OPP가 되었다. ‘페이지랭크’는 인터넷 안에 있는 인간과 비인간을 포함하는 모든 행위자들에게 적용되는 ‘헌법’과도 비슷하게 작동되고 있었다. 하지만 구글의 헌법은 국가의 헌법과 달리 원리적 충만론을 통한 ‘정화’가 아니라 새로운 기술적 하이브리드들을 에이전트로 결합시키면서 국지적으로 작동했다. ‘페이지랭크’, N-gram, 웹 페이지, 하이퍼링크, 국가의 검열 장치들, 저작권의 공정이용에 대한 판례, 이를 이해하고 있는 사용자들의 클릭과 같은 이중적 요소들을 ‘페이지랭크’라는 하나의 블랙박스화된 의미작용 혹은 권력작용으로 통합했다. 구글 ‘페이지랭크’라는 블랙박스는 기술적으로 매우 이질적인 요소를 ‘병합’한 것이었지만, 인간 하이브리드들에게 인터넷 안에서 가장 강력하게 ‘보편성’을 대체하는 네트워크가 되었다. 구글은 근대 사회가 갈라 놓았던 공적 영역과 사적 영역의 구분도 모호하게 만들었다. 구글의 권력은 ‘사람의 권리’ 담론 뒤에서 공적 가치인 ‘자유’를 활용하여 세계화된 사적 권력으로 성장하고 있었다. 자본주의가 ‘사적 소유’라는 ‘사람의 권리’ 담론 내의 공적 가치를 활용하여 사적 권력을 확장하고 있었던 것 처럼, 구글은 ‘자유’를 활용하여 결과적으로 같은 효과를 냈다. 이는 구글이 갈라져 있던 두 영역을 연결한 것이 아니라 또 다른 연결 통로를 만들었다는

뜻이었다.

7. 결론 : 구글 동맹은 어떻게 공적 영역을 사유화 하였는가

구글의 혁신적 검색엔진은 발명된 것이 아니라 구성되었다. 구글의 알고리즘은 탈역사적인 기술이 아니라 역사적으로 ‘축적된 행위자’들이 웹이라는 환경에 연결되면서 그 안에서 각자의 목표를 실현하려는 선택의 과정이 하나의 행위자네트워크로 연결되어 이루어진 역사적 결과다. 구글 탄생 전에 나타나 구글의 강력한 동맹이 된 ‘축적된 행위자’들은 부시의 메멕스, 넬슨의 하이퍼텍스트, 1980년대를 전후로 급속히 성장한 게임과 PC의 하이퍼미디어로부터 온 인간-소프트웨어 하이브리드들이었다.

웹의 초기 야후!는 하이퍼미디어의 이용경험을 가진 독자들과의 동맹을 확대하면서 인터넷의 성장을 주도했다. 그러나 야후!는 닷컴기업들과 웹 스파머의 증가로 인해 사용자들에게 나쁜 경험을 줄 가능성이 있는 검증되지 않은 웹 페이지가 증가함에 따라, 이들과의 연결을 배제함으로써 오염된 웹 페이지로 인해 어려움을 겪던 검색엔진과의 경쟁에서 살아

남았다. 그러나 야후!의 ‘번역으로서의 배제’는 실패하고 있었다. 만약 이것이 성공적으로 작동했다면 야후!가 배제한 하이브리드들은 줄어들고 이들이 스스로 목표를 변경하여 야후! 디렉토리 등록을 받아들이거나 디렉토리 안에서만 정보를 찾는 행위자로 바뀌어야 했지만, 이들은 여전히 빠르게 증식되면서 야후!를 위기로 몰고 가고 있었다.

웹 스파머의 공격을 견뎌내지 못함으로써 야후!와의 경쟁에서 밀려난 다른 검색엔진들과 달리 구글 검색엔진은 전통적 의미의 정보 검색엔진이 기술적으로 진화한 결과가 아니었다. 구글 검색의 핵심 알고리즘인 ‘페이지랭크’는 발명자들 주위의 인간행위자들을 동원하여 백럽 크롤러라는 시스템이 되었는데, 백럽 크롤러는 당시 빠르게 증가하던 야후!가 배제한 웹 페이지의 저자와 디렉토리에 등록되지 않은 웹 페이지를 찾는 독자들과 동맹에 성공함으로써 ‘페이지랭크’ 기술을 매각하려는 발명자들의 목표를 실현시켜주는 듯 했다. 그러나 매각은 실패했고 ‘페이지랭크’의 행위자네트워크는 상용 검색엔진이 되어야 살아남을 수 있게 되었다.

그러나 구글이 상용 검색엔진이 되기 위해서는 ‘페이지랭크’가 구축한 행위자네트워크를 다른 영역으로 확장해야 했다. 기술적으로는 당시 상용 검색엔진이 되기 위해 필수적이었던 전문검색(全文檢索)을 수용해야 했다. 이는 ‘페이지랭크’가 가진 비용효율성과 검색품질을 일부 희생하

는 것이었지만, 이 과정에서 구글은 ‘페이지랭크’의 앵커텍스트 색인에 사용되던 N-gram이라는 텍스트 처리기술의 적용 영역을 본문으로 확장함으로써, 향후 구글이 맞이한 위기를 해소할 수 있는 중요한 기술을 획득했을 뿐 아니라 향후 구글의 사적 권력이 검색을 넘어 다른 영역으로 확장되는데 결정적 기여를 한 기술로의 확장을 시작했다.

구글의 위기는 다른 검색엔진과 마찬가지로 스팸머들의 공격에서 비롯되었다. 그러나 구글은 이들을 동맹에서 배제하고 포털이 되기로 번역의 전략을 바꾼 다른 검색엔진들과 달리, 자신의 질서를 위협하는 하이브리드들과의 동맹을 맺기 위해 N-gram 필터와 같은 새로운 에이전트를 투입하여 블랙박스화된 ‘페이지랭크’의 목표를 변경함으로써 이들을 길들이는데 성공했다. 구글은 이러한 성공에 힘입어 N-gram 필터 기술을 활용하여 저작권이나 포르노에 관련된 통제, 악성코드의 차단, 정치적 검열과 같은 웹 내의 다른 권력작용에 개입하기 시작했다.

때로는 소송에 짐으로써 권력을 획득했고, 때로는 국가와의 협상을 통해 권력을 획득했다. N-gram 필터들에 의해 웹과 구글을 위협하는 하이브리드들은 성공적으로 번역되고 있었다. 이렇게 길들여진 웹은 구글의 사적 권력이 지배하는 공간이 되었고, 하이브리드들이 빠르게 증식되는 만큼 구글의 사적 영역도 확대되었다. 그러자 곧 구글의 권력은 인터넷 밖을 향하고 있었는데, 이를 위해 구글은 국가들과의 결합이 느슨해진

‘자유’라는 공적 가치를 자신의 동맹 안으로 끌어들였다. 구글은 사적 권력이 공적 영역으로 팽창할 수 있는 새로운 경로를 찾고 있었다.

구글은 ‘페이지랭크’라는 단일한 행위자를 통해 인터넷 안에서 세계 각국의 법체계와 생산제관계의 권력작용을 통합해 뉘으로써, 이들이 원래 서로 뿔 수 없이 완전히 병합된 하나라는 것을 보여주었다. 인터넷 안의 사유화된 영역이 확장되고 그 내부에서 벌어지는 비가시적 위험이 통제되지 않는 문제를 안고 있다는 시각은, 국가의 충만함을 원리로 기획된 근대적 질서를 보일의 진공펌프가 위험하다는 홉스의 경고와 같은 선상에 있다. 근대 사회는 인간과 비인간의 이분법에 의한 정화작용과 함께 비인간에게 인간과 동등한 권리를 보장하게 하는 번역을 작동시킴으로써 성공과 동시에 위기를 잉태한다. 이는 마치 ‘속도 무제한’의 도로에서 인간 운전자와 자율주행 자동차가 동등한 권리를 갖고 주행하는 것이 위기를 증폭시키는 이유와 같다. 이것은 최근의 고도로 발전된 기술사회에만 적용되는 설명이 아니며, 기업과 국가와 같은 근대적 비인간행위자들의 행위능력과 책임이 인간행위자들의 그것과 비교하여 비대칭적이었음이 고도로 발전된 기술 행위자의 등장으로 드러난 것일 뿐이다.

근대인들은 라투어의 지적대로 사회와 자연을 분리하고 주체와 객체를 분리함으로써 – 실제로 그랬던 적은 없지만 – 세계를 정화(purify)하였는데, 그것을 통해 이루고자 한 것은 무엇일까? 근대인들은 ‘충만했던

신’을 대신하여 ‘국가로 충만한’ 세계를 꿈꾸었다. 어느 곳에서도 국가를 작동시키기 위해 공간 전체를 ‘국가’로 뒤덮었고, ‘헌법’의 보편성을 선언함으로써 누구에게나 작동한다는 믿음을 만들고자 했다. 그러나 실제의 공간은 국가로 충만하지도 헌법의 ‘보편성’이 작동하지도 않는다. 인터넷은 20세기 후반에 나타난 ‘국가 없는 공간’이다. 근대인들이 국가 없는 공간이 ‘무질서’로 가득한 공간이라고 착각하는 것과는 달리 그 안에서는 권력이 작동하고 있으며 다만 ‘국가’가 아닐 뿐이다. 국가들은 인터넷 안의 권력과 협상하며 그들에게 국가를 대리할 작위를 수여한다. 그러나 인터넷 안에 있는 국가의 비인간 대리인들은 언제나 반란을 일으킬 수 있다. 이는 대양 시대 해적의 사략선(privateer)이 국가를 대리했던 것과 비슷하다. 근대성은 국가와 해적을 동시에 허용함으로써 결코 패배하지 않는다.

구글은 그런 점에서 16~19세기의 해적과 같이 국가가 작동하지 않는 곳에서 권력을 획득하는 새로운 권력주체다. 해적들은 때로 국가와 대립하지만 국가와의 협력을 배제하지 않는다. 해적들은 ‘보편성’을 선언하지 않고 ‘실효적 지배의 충만함’을 택한다. 법체계의 무결성을 추구하지 않고 모든 행위자들이 자신이 설치한 OPP를 통과할 수 밖에 없도록 많은 비인간행위자들을 동원하여 권력을 구축한다. 물론 해적이 현재 인류 역사의 해결책이라는 뜻은 아니다. 해적과 그 후예인 주식회사는 스스

로 자신의 사적 권력을 확장하여 공적 가치를 파괴한다는 점에서 국가보다 더 나쁘다. 그러나 구글은 국가들과의 동맹이 느슨해진 ‘자유’와 같은 근대의 공적 가치와 연결된 행위자들을 동맹 안으로 끌어들이므로써 ‘사람의 권리’ 뒤에서 자신을 정당화하며 사적 권력을 전세계로 확장한다는 점에서 더 위험하다. 그리고 가장 큰 문제는 19세기에는 국가들의 합의에 의해 사략선을 폐지할 수 있었지만, 21세기에도 가능할 것인지는 알 수 없다는 사실이다.

참고문헌

- 미셸 푸코 (2003), *감시와 처벌*, 오생근 역, 나남.
- 미셸 칼롱 (2010), “번역의 사회학의 몇 가지 요소들: 가리비와 생브리의 만의 어부들 길들이기“, *인간•사물•동맹*. 홍성욱 역, 이음 : pp.57-94.
- 브루노 라투어 (2010), “행위자네트워크 이론에 관하여: 약간의 해명, 그리고 문제를 더 복잡하게 만들기 ” , *인간•사물•동맹*. 홍성욱 역, 이음: 97-124.
- 예병일 (2003), “[인터넷 사업 성공사례] 구글, 사업 속의 장인정신“, *www.linux.co.kr* ,
http://www.linux.co.kr/main/view_post.asp?post_seq_no=6843, 2016.05.16 Retrieved.
- 윤평중 (1997), “미셸 푸코“, *현대철학의 흐름*, 동녘 : pp.292-317
- 이광석 (2015), “인터넷 역사 방법론의 논점-기술계보학적 태도와 기술 주체의 복원” . *커뮤니케이션 이론* 11(2): pp.142-187.
- 이해성 (2007). “구글 검색 페이지 10곳중 1곳 악성코드 감염...액티브X 설치 요구하면 일단 의심해야“, *www.hankyung.com*,
<http://www.hankyung.com/news/app/newsview.php?aid=2007051434051> ,

2016.05.16 Retrieved.

이홍균 (2006), “사적 영역의 팽창에 의한 공적 영역의 파괴.” *현상과인식* 30.3: pp.80-101.

장중혁 (2014), “키워드의 제국, 구글의 초국가적 진화.” *서양사연구* 50: pp.363-394.

조지 랜도우 (2001), *하이퍼텍스트 2.0*, 문화과학사, 여국현 역.

존 로(2010), “ANT에 대한 노트: 질서 짓기, 전략, 이질성에 대하여” , *인간·사물·동맹*. 홍성욱 역, 이음: pp.39-56.

허완 (2015), “구글, 한국 이용자에게 ‘세이프서치’ 강제적용?“, *www.huffingtonpost.kr*,

http://www.huffingtonpost.kr/2015/12/14/story_n_8802318.html, 2016.05.16 Retrieved.

홍성욱 (2010), “7 가지 테제로 이해하는 ANT.” *인간·사물·동맹*. 이음 : pp.17-35.

Abbate, J. (2000). *Inventing the internet*. MIT press.

Arendt, H. (1958), *The human condition* , University of Chicago.

Ken, A. (2010) *Googled: The end of the world as we know it*. Random House.

Baker, E. (2006), *Media concentration and democracy: Why ownership matters*. Cambridge University Press.

Baker, N. (2002) *Double fold: Libraries and the assault on paper*. Random House.

Ballesteros, L. and Bruce Croft W. (1997), "Phrasal translation and query expansion techniques for cross-language information retrieval." *ACM SIGIR Forum*. Vol. 31. No. SI. ACM.

Beel, J., Gipp, G. and Wilde, E. (2009), "Academic Search Engine Optimization (aseo) Optimizing Scholarly Literature for Google Scholar & Co." *Journal of scholarly publishing* 41.2: pp.176-190.

Berners-Lee, T. (1989), "Information management: A proposal", <http://www.w3.org/History/1989/proposal.html>

Bolter, J. (2001), *Writing Space: Computers, Hypertext, and the Remediation of Print*. Taylor & Francis.

_____ (1994), "Authors and readers in an age of electronic texts." *Literary texts in an electronic age: Scholarly implications and library services* .

Brin, S., Davis, J. & Garcia-Molina, H. (1995), "Copy detection

mechanisms for digital documents” . In *Proceedings of the ACM SIGMOD Annual Conference*.

Brin, S., Page, L. (1998), 'The anatomy of a large-scale hypertextual Web search engine', *Computer Networks and ISDN Systems* 30 : pp 107-117

Broder, A., Najork, M. & Wiener, J. (2003), “Efficient URL caching for world wide web crawling.” *Proceedings of the 12th international conference on World Wide Web*. ACM.

Brown, S.D. (2002), ‘Michel Serres: Science, translation and the logic of the parasite ’ *Theory, Culture & Society* ISSN 0263-2764 Vol 19, 3. pp1-27.

Brügger, N. (2010) *Web History*. Peter Lang.

Bush, V. (1945), “ As We May Think ” , The atlantic monthly 176.1: pp.101-108. <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>

Bush, V. (1991) “Memex revisited.” *From Memex to hypertext*. Academic Press Professional, Inc.

Callon, M. (1986) “Some elements of a sociology of translation:

domestication of the scallops and the fishermen of St Brieuc Bay“
action and belief: a new sociology of knowledge?, London, Routledge.

_____ (1987) “Society in the making: the study of technology as a tool for sociological analysis.” *The social construction of technological systems: New directions in the sociology and history of technology* : pp.83-103.

_____ (2007) “What does it mean to say that economics is performative.” *Do economists make markets* (2007): pp.311-357.

Cho, J., Garcia-Molina, H. (2000), “Synchronizing a database to improve freshness.” *ACM Sigmod Record*. Vol. 29. No. 2. ACM.

Crouch, C. (2011), *The strange non-death of neo-liberalism*. Polity.

Cutts, M. (2009) “PageRank sculpting“, *mattcutts.com*,
<https://www.mattcutts.com/blog/pagerank-sculpting/>.

Dean, B. (2015), Google ’ s 200 Ranking Factors: The Complete List, *backlinko.com*, <http://backlinko.com/google-ranking-factors>.

Denning, J. (1995), “Plagiarism in the Web“, *COMMUNICATIONS OF THE ACM*, December 1995/Vol. 38, No. 12 : on p.29.

DiMaggio, P., et al. (2001), “Social implications of the Internet.” *Annual*

review of sociology : pp.307-336.

Duchamp, D. (1999) "Prefetching Hyperlinks." *USENIX Symposium on Internet Technologies and Systems*.

Dwork, C., et al. (2001), "Rank aggregation methods for the web." *Proceedings of the 10th international conference on World Wide Web*. ACM.

Ferro, S. (2015), The Internet Has Exposed Us To The Risk Of Doom, *www.businessinsider.com*, <http://www.businessinsider.com/thoughts-on-the-technology-devolution-2015-1>.

Fletcher, O. (2009), Google Porn Filter Gained China's Thumbs-up, *pcworld.com*.

Frana, P. (2004), "Before the web there was Gopher." *IEEE Annals of the History of Computing* 1 : pp.20-41.

Gyöngyi, Z., Garcia-Molina, H. & Pedersen, J. (2004), "Combating web spam with trustrank." *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment.

Habermas, J. (1991), *The structural transformation of the public sphere*:

An inquiry into a category of bourgeois society, MIT press.

Hawking, D., et al. (1999), "Overview of the trec-8 web track." *TREC*.

Hawking, D. (2000), "Overview of the TREC-9 Web Track." *TREC*.

Hu, J. (2001), 알타비스타 「 결국 검색 서비스로 복귀 」 , zdnet.co.kr,

http://www.zdnet.co.kr/news/news_view.asp?article_id=00000010040268,

2016.05.16 Retrieved.

Hughes, T. (1987), "The evolution of large technological systems." *The social construction of technological systems: New directions in the sociology and history of technology* :pp. 51-82.

Kleinberg, J. (1999) "Authoritative sources in a hyperlinked environment." *Journal of the ACM (JACM)* 46.5 : pp.604-632.

Landow, G. (1991) *HyperText: the convergence of contemporary critical theory and technology* (parallax: re-visions of culture and society series), Johns Hopkins University Press.

Langville, A., Meyer, C. (2011), *Google's PageRank and beyond: The science of search engine rankings*, Princeton University Press.

Latour, B. (1993), *We have never been modern*. Harvard University Press.

_____ (1994) “On technical mediation.” *Common knowledge* 3.2 : pp.29-64.

_____ (2004) “On using ANT for studying information systems: a (somewhat) Socratic dialogue ” . *The Social Study of Information and Communication Technology: Innovation, Actors, and Contexts* : pp.62-76

Law, J. (1987), “Technology and heterogeneous engineering: the case of Portuguese expansion.” *The social construction of technological systems: New directions in the sociology and history of technology* 1 : pp.111-134.

_____ (1992), “Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity.” *Systems practice* 5.4 : pp.379-393.

_____ (2009), “Actor network theory and material semiotics.” *The new Blackwell companion to social theory* : pp.141-158.

Levy, S. (2010), Exclusive: How Google ’ s Algorithm Rules the Web, *wired.com*, http://www.wired.com/2010/02/ff_google_algorithm/, 2016.05.16 Retrieved.

Li, Y., Rafsky, L. (1997), “Beyond relevance ranking: Hyperlink vector voting.” *RIAO*. Vol. 97.

McChesney, R. (2013), *Digital disconnect: How capitalism is turning the Internet against democracy*, The New Press.

McKnight, L., Bailey, J. (1998), *Internet economics*, MIT Press.

Metaxas, P., DeStefano, J. (2005), "Web Spam, Propaganda and Trust." *AIRWeb*.

Nelson, T. (1965), "Complex information processing: a file structure for the complex, the changing and the indeterminate." *Proceedings of the 1965 20th national conference. ACM*.

_____ (1974), *Computer lib: Dream machines*. self-published.

_____ (1988) "Managing immense storage." *Byte* 13.1 (1988): pp.225-238.

Paepcke, A., et al. (1996), "Towards Interoperability in Digital Libraries." *IEEE Computer* 29.5.

Page, L. (1997), "Improved Text Searching in Hypertext System" , US Patent.

Page, L., et al. (1999), "The PageRank citation ranking: bringing order to the Web." (1999).

Patry, W. (2009), *Moral panics and the copyright wars*. Oxford University Press.

Peterson, A. (2014), How Google Search is changing to fight online

piracy, *The Washington Post*.

Röscheisen, M. (1997), *A network-centric design for relationship-based rights management*, Diss. Stanford University.

Rowley, H., Jing, Y. & Baluja, S. (2006), “Large scale image-based adult-content filtering.” *VISAPP* (1).

Severyn, A., et al. (2014), “Opinion Mining on YouTube.” *ACL* (1).

Shapin, S., Schaffer, S. (1985), *Leviathan and the air-pump: Hobbes, Boyle, and the Experimental Life*. Princeton: Princeton University Press.

Shivakumar, N., Garcia-Molina, H. (1995), “SCAM: A copy detection mechanism for digital documents“, *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*, Austin, Texas.

Siegler (2010), When Google want to sell to Excite for under \$1 million – And they passed, *techcrunch.com*.

Singhal, A., Kaszkiel, M. (2001), “A case study in Web search using TREC algorithms” In *Proceedings of the 10th international World Wide Web conference*, Hong Kong, 1–5 May 2001 : pp. 708–716), on p.709.

Stoffe, B. (2013), Google Stock: Buy It and Hold It For Life,

www.fool.com, 2016.07.18 Retrieved.

Suen, C. (1979), "N-gram statistics for natural language understanding and text processing." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 2 : pp.164-172.

Taylor, B., Fingal, D. & Aberdeen, D. (2007), "The war against spam: A report from the front line." *NIPS Workshop on Machine Learning in Adversarial Environments*.

Vaidhyanathan, S. (2012), *The Googlization of everything:(and why we should worry)*, Univ of California Press.

Winner, L. (1993), "Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology." *Science, Technology, and Human Values* : pp.362-378.

Abstract

ANT Analysis on Google Search Engine

Jung-hyouk Zhang

Program in History and Philosophy of Science

The Graduate School

Seoul National University

This study set out to trace the process of changes in Actor-Network of the Internet over a period of ten years after the birth of Google, from 1998 to 2007. What made Google succeed against a lot of failure of search engines? And How Google defeat Yahoo!, the leader of internet portal? How much is it true that the myth of Google's success from their innovative technologies? This study tackles the questions with theoretical tools of ANT. Micro analysis of three major actors – ‘reader of hyper-media’, ‘author of hypertext’ and ‘reader of hypertext’ – is used for solving the questions, focused on the genealogical trace of the three actors.

Hybridization has made them the major actors of WWW. Yahoo! can succeed by the translation on ‘reader of hyper-media’, and fall into

failure because of missing the alliance with ‘author of hypertext’ and ‘reader of hypertext’. Google can get the influence on the Internet by the translation on them. The technologies of Google are not invented by inventors, constituted by changes of Actor-Network of the Internet. Google can get the technical success and the political influence occasionally by some strategic choice, not technical choice. And then Power of Google expands beyond the internet to real world and beyond modern order of nations to trans-national order.

keywords : Google, Search Engine, Actor-Network, Hybrid, Internet

Student Number : 97321-503