위키백과의 장기간 편집 기록 분석을 통한 위키백과 문서의 분류와 집단지성의 형성과정

DOI: 10.3938/PhiT.25.006

윤진혁 · 이상훈 · 정하웅

Categorization of Wikipedia Articles and Formation of Collective Knowledge via the Massive History of an Open-editing Encyclopedia

Jinhyuk YUN, Sang Hoon LEE and Hawoong JEONG

Wikipedia, which is a public Internet encyclopedia with a vast amount of content, is written by voluntary editors with diverse backgrounds in a collective fashion; anyone can access and edit most articles. This open-editing nature may give us the impression that Wikipedia is an unstable and unreliable source, but many studies indicate that Wikipedia is even more accurate and consistent than conventional encyclopedias. Researchers have attempted to understand such unexpected credibility, but they have usually used the number of edits as the unit of time, without considering real time. In this work, we investigate the formation of such collective intelligence through a systematic analysis using the entire record of 34,534,110 English Wikipedia articles between 2001 and 2014. From this enormous amount of data, we reveal the patterns on both timewise and lengthwise editing scales, which suggests that it is crucial to consider the real-time dynamics. By considering real time, we classify the growth patterns into four different types that are not observed when utilizing only the number of edits as the unit of time. To explain these re-

sults, we introduce a mechanistic model that harnesses the dynamics of article editing dynamics based on both editor-editor and editor-article dynamics. The model successfully generates the crucial properties of real Wikipedia articles such as the distinct types of articles for the editing patterns represented by the interrelationship between the numbers of edits and editors and the article's length. In addition, our model predicts that articles which are infrequently referred to tend to grow faster than ones which are frequently referred to and that counterintuitively articles inducing a large motivation to edit actually shrink the number of participants. We suggest that this decline of participating editors eventually causes inequality among the editors, which will be exacerbated with time.

서론

인류는 오랜 시간에 걸쳐 지식을 축적해 왔으며, 이를 통해 문명을 발전시켜 왔다. 이러한 지식의 전승을 위해 고대의 동굴 벽화부터 현대의 인터넷에 이르기까지 다양한 기록 매체가 동원되었고, 이를 통해 인류는 고도의 문명을 구축할 수 있었다. 이러한 기록 매체의 대표적인 예로 백과사전을 들 수 있다. 백과사전은 수많은 분야의 전문 지식을 모아 놓은 것으로, 전통적으로 출판사의 촉탁에 의해 전문가들이 모여 각자의 전문분야를 편집해 오는 방식을 취해 왔다.

본 연구팀에서 분석^[1]한 위키백과(Wikipedia)^[2]는 이러한 중앙의 통제 없이 자발적으로 참여하는 수많은 사람들이 모든 글을 자유롭게 편집하고 열람할 수 있는 공개 온라인 백과사전이다. 자발적인 참여라는 특성은 필연적으로 비전문가의 편집 참여를 가능하게 하고, 이로 인해 많은 사람들이 위키백과의 신뢰도에 의문을 표해왔다. 그러나 연구를 통해 위키백과

저자약력

윤진혁 학생은 KAIST 물리학과 석·박사 통합과정에서 빅 데이터를 이용한 복잡계 분석을 수행하고 있다.

이상훈 박사는 2010년 KAIST 물리학과에서 연결망에서의 상전이와 임계 현상에 관한 연구로 박사학위를 받았고, 스웨덴 Umeå 대학, 영국 Oxford 대학에서 박사후 연구원, 성균관대학교 에너지과학과 BK21+ 연구교수를 거쳐 현재 고등과학원 연구원으로 재직 중이다.

정하웅 교수는 서울대학교 물리학과에서 박사학위를 받았고, University of Notre Dame에서 연구조교수를 거쳐 현재 복잡계 및 통계물리에 관해 연구를 진행하며 KAIST 물리학과 석좌교수로 재직 중이다.
(hjeong@kaist.ac.kr)

REFERENCES

- [1] J. Yun, S. H. Lee and H. Jeong, Phys. Rev. E **93**, 012307 (2016).
- [2] <http://www.wikipedia.org>.

Fig. 1. Distribution of time between two consecutive edits in Wikipedia. Each symbol corresponds to the range of ages to which articles belong. The time differences follow fat-tailed distribution, which is a sign of the burstiness, with a daily periodic pattern (1 day=86400 s).

의 정확도가 우리의 편견보다 훨씬 뛰어나다는 사실이 알려졌고,^[3,4] 학술지에 게재되는 논문들 중 일부는 실제로 위키백과를 인용하기 시작하였다.^[5] 이러한 선입관과 실제 현상 사이의 괴리는 많은 학자들에게 관심을 불러일으켜 위키백과와 같은 공유 자료(communal data)에 대한 다양한 연구가 시작되었다. 연구자들은 위키백과의 참여자와 그들의 행동양식, 특히 대립과 화해 과정 등을 통해 위키백과에 참여하는 참여자들에 대한 분석을 실시했다. 또한 위키백과 전체 문서 개수의 변화를 통해 위키백과의 성장을 분석하기도 하였다. 하지만 이러한 분석에서는 기술적인 문제로 인해 개개별 문서의 특징을 구분짓는 과정과 그 이유에 대한 이해가 부족하였다.

최근의 복잡계(complex systems)와 빅 데이터(big data)에 대한 관심은 우리에게 위키백과를 분석할 수 있는 또 다른 방법론을 제시해 준다. 본 연구팀에서는 위키백과 문서들의 성장양상 분석을 통해 어떠한 중앙의 통제를 받지 않음에도 불구하고 문서들이 단지 네 가지로 구분되는 매우 규칙적인 성장을 하고 있다는 사실을 밝혀내었다. 또한 모형을 통해 이러한 네 가지의 분류를 만들어낸 원인을 규명하고, 더 나아가 미래의 위키백과가 어떻게 변할지에 대한 전망도 제시할 수 있었다.

위키백과 편집 기록 데이터

본 연구에서 사용된 자료는 영어판 위키백과의 탄생 시점인 2000년 1월부터 2014년 12월 8일까지 행해진 모든 편집에 대한 기록 사본^[6]이다. 이 안에는 34,534,110개의 서로 다른 문서의 편집 기록이 존재한다. 기존의 연구들^[7,8]은 다양한 이유로 전체 데이터가 아닌 일부만을 사용하였으나, 우리는 위키

Fig. 2. Size difference between two consecutive edits in Wikipedia, in the unit of kilobyte (KB). Each symbol corresponds to the range of articles' size when the latter of two consecutive edits occurred.

백과의 전체 성장 양상을 통계적으로 관측하기 위해 모든 문서(article, category, talk 등)의 전 기간 데이터를 사용하였다. 이 기간 동안 총 587,606,564번의 편집이 행해졌으며, 개개별 문서는 최소 1번부터 최대 1,044,048번까지 편집이 이루어졌다.

문서 구분에 이어 우리는 서로 다른 편집자들을 구분하기 위하여 편집자가 로그인한 경우에는 편집자의 ID를, 그렇지 않은 경우에는 편집 시점의 IP 주소를 사용하여 각각의 ID와 IP 주소는 개별 편집자에 대응된다고 가정하였다. 우리는 이 과정에서 흥미로운 사실을 하나 알 수 있었는데, 대부분의(83.9%, 40,536,451명 중 34,006,990명)의 편집자는 로그인을 하지 않았으나 이들은 전체 편집의 단지 22.5%(132,475,517번)만을 수행하였다는 점이다. 즉, 자주 편집하는 사람들은 자신의 고유 ID를 사용하는 경향이 매우 강하였다.

나이에 따른 눈금 바꿈(rescaling)

이전에 위키백과의 성장에 대한 연구가 없었던 것은 아니나,

REFERENCES

- [3] T. Chesney, First Monday **11**, 11 (2006).
- [4] J. Giles, Nature **438**, 900 (2005).
- [5] M. D. Bould, E. S. Hladkovicz, A.-A. E. Pigford, L.-A. Uffholz, T. Postonogova, E. Shin and S. Boet, BMJ **348**, g1585 (2014).
- [6] <https://dumps.wikimedia.org/backup-index.html>.
- [7] A. Kittur, B. Suh and E. H. Chi, Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08) (ACM, New York, 2008), p. 477.
- [8] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye and V. Raman, Proceedings of the 4th International Symposium on Wikis (WikiSym '08) (ACM, New York, 2008), Article No. 26.

Fig. 3. Interplay between time-rescaled measures for Wikipedia in December 2014. Each point corresponds to an individual Wikipedia article. (a) The number of edits N_e^* versus the number of editors N_p^* and (inset) magnified view on $x \approx 10^{-1}$ and $y \approx 10^{-1}$. (b) The number of editors N_p^* and the article size S^* . The bimodal structure in (a) is highlighted in (c) by manually dividing the upper (purple) and lower (blue) modes, but this division of modes does not divide the modes of (d). Likewise, the bimodal structure in (b) is highlighted in (f) by manually dividing the upper (purple) and lower (blue) modes, but this division of modes does not divide the modes of (e).

이러한 연구들은 한 번의 편집이 미치는 영향이나 개별 편집 사이의 시간 간격에 대한 고려가 부족하였다. 그림 1은 두 개의 연속된 편집간의 시간 간격을 두 번째 편집이 일어난 시간에 따라서 그린 분포이다. 그림에서 확인할 수 있듯 편집간의 시간 간격은 꼬리가 두꺼운(fat-tailed) 분포를 따르는 것을 볼 수 있다. 다시 말해 위키백과 편집의 대부분은 짧은 시간 간격을 가지고 일어나지만, 아주 긴 시간 간격을 가지고 있는 편집들도 무시할 수 없을 정도로 많이 존재한다. 또한 그림 2에서 볼 수 있듯 그 시점의 문서 크기와는 무관하게 문서 크기 변화량은 거의 일정하다는 결과를 확인할 수 있었다. 문서가 커질수록 한 번의 편집이 주는 영향이 줄어드는 것이다.

이를 보정하기 위해서 본 연구에서는 아래와 같이 각 문서의 특징을 나타내는 변수를 나이, 다시 말해 그 문서의 최초 편집부터 마지막 편집까지의 시간 $[T(i)]$ 으로 나눠주는 방법을 택했다.

$$N_e^*(i) = N_e(i)/T(i) \quad (1)$$

$$N_p^*(i) = N_p(i)/T(i) \quad (2)$$

$$S^*(i) = S(i)/T(i) \quad (3)$$

여기서 $N_e(i)$ 는 i 라는 문서의 최종 편집 횟수, $N_p(i)$ 는 문서 i 의 최종 참여 편집자 수, 그리고 $S(i)$ 는 i 의 현재 크기를 의미한다. 재미있게도 이렇게 단순한 눈금 바꿈(rescaling)을 통해서 기존에 발견하지 못했던 문서들의 분류 체계를 확인할 수 있었다. 일견 복잡하게 보이는 위키백과의 수많은 문서지만, 그 속을 들여다보면 단지 4가지의 분류만이 존재하는 것이다. 그림 3의 (a)와 (b)에 나타난 네 가지의 분류가 그것이다.

Table 1. Classification of Wikipedia articles in Fig. 3.

Productivity	much	little
Number of edits		
many	category 1	category 2
few	category 3	category 4

그림 3(a)는 단위 시간당 편집에 참여한 사람 수와 단위 시간당 편집된 횟수를 비교한 그래프이다. 위 그래프의 점 하나는 위키백과의 문서 하나를 나타낸다. 그래프에서 볼 수 있듯 한 그룹은 한 번의 편집을 위해 0.5명의 참여를 필요로 하고, 다른 그룹은 0.9명의 참여를 필요로 하는 것을 알 수 있다. 이와 유사하게 그림 3(b)는 한 명의 편집자가 평균적으로 얼마나 많은 데이터를 만들어 냈는지를 보여주는데, 위와 비슷하게 한 그룹은 평균적으로 약 30바이트, 한 그룹은 900바이트를 만들어내는 차이를 보였다.

일반적으로 이렇게 두 가지의 다른 변수 공간에서 유사한 형태의 분류(classification)가 나타날 경우, 양쪽의 상관관계(correlation)가 있다고 추정할 수 있다. 그림 3의 (c)-(f)는 그러한 상관관계의 유무를 확인하는 그림인데, 신기하게도 두 가지 그래프의 분류는 서로 아무런 상관관계가 없는 것으로 나타났다. 정리하자면 위키백과의 문서들은 편집자 1인당 자료 생성량과 수정 횟수라는 두 가지 독립된 기준에 따라 표 1에 나타난 네 가지로 구분할 수 있다는 것이다.

위키백과 편집 모형

앞부분에서 우리는 나이에 따른 눈금 바꿈 분석을 통해 위키백과의 문서들이 단지 4개의 분류로 구분 가능하다는 사실을 밝혀내었다. 본 연구에서는 더 나아가 이러한 분류의 근원을 규명하기 위하여 행위자 기반 모형(agent-based modeling)^[9]을 통해 현재 위키백과의 형성 과정을 분석해보았다. 이 소개 글에서는 간략히 모형의 특징만을 기술하였으므로 자세한 내용은 연구논문^[1]을 확인하기 바란다.

실제 위키백과는 수천만 개의 문서를 가지고 있으나, 우리의 모형은 단 한 개의 문서만을 고려하였다. 이 문서에는 여러 명의 가상 편집자 후보들이 연결되어 있고, 이들 사이에도 의견 교환이 일어난다고 가정하였다. 이러한 의견 교환은 미리 주어진 연결망에서 이루어졌는데, 우리는 에르되스-레니(Erdős-

REFERENCES

- [9] C. Castellano, S. Fortunato and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).

Fig. 4. Snapshots of (a) $\langle N_p^*(i) \rangle / \langle N_e^*(i) \rangle$ corresponding to the proportionality constant in Fig. 3(a) and (b) $\langle S^*(i) \rangle / \langle N_p^*(i) \rangle$ corresponding to Fig. 3(b), where $\langle \dots \rangle$ is the averaged quantity over all of the agents ($= \sum_i \dots / N$). $N=10^5$ agents are connected by the Erdős-Rényi random network with connecting probability 0.1. Each point is the average of 1000 independent runs of simulation with $\varepsilon_w = \varepsilon_m = \varepsilon_j = 0.1$.

Rényi) 연결망을 사용하였다.^[10] 위키백과와 편집자들 모두 0과 1 사이의 숫자로 표현되는 개별 의견을 가지고 있는데, 0과 1은 각각 보수와 진보처럼 양 극단의 의견을 나타낸다. 초기에 편집자들은 0과 1 사이에서 연속균등분포를 따르는 임의의 의견 값을 가지며, 위키백과는 한쪽 극단에서 시작한다는 가정 하에 1의 값을 가진다.

모형은 크게 2가지의 과정으로 이루어져 있다. 그중 첫 번째는 편집자끼리 의견을 조율해 합의점을 찾는 과정이며, 두 번째는 편집자가 위키문서를 편집하는 과정이다.

보통 사람들은 서로의 의견을 조율할 때 근거를 제시하여 상대방을 설득하려 한다. 우리는 위키백과와 일반 매체라는 두 가지로 근거를 분류할 수 있다고 가정했으며, 편집자가 위키백과를 근거로 사용할 확률을 p 라는 값으로 두었다. 모형에서는 매 시간 연결망에서 이웃한 두 명의 편집자를 선택해 서로의 의견을 조율한다. 이 편집자들은 p 라는 확률에 따라 선택된 근거 (위키백과 혹은 일반매체)에 가까운 편집자가 근거에 먼 편집자를 설득해 의견을 바꾼다.

의견 조율 과정 이후에 한 명의 편집자를 선택해 편집자가 위키백과를 자신의 의견과 유사하게 편집하는 과정을 거친다. 이때 편집은 미리 주어진 q 라는 기본 편집 빈도에 편집자와 위키백과의 의견 차이를 곱한 확률로 일어난다. 또한 우리는 이러한 편집 이후에 편집자의 그 문서에 대한 애착(attachment)이 더 커진다고 생각하였다. 애착이 더 큰 편집자는 다른 편집자들에 비해 더 높은 확률로 그 문서를 수정하며, 이를 통해 편집자들 개개인의 편집 횟수에 나타난 꼬리가 두꺼운 분포(fat-tailed distribution)를 자연스럽게 재현할 수 있었다.

그림 4는 위의 모형을 통해 위키백과가 시간에 따라 어떤 형식으로 형성되어 왔는지를 재구성해본 그림이다. 그림 4(a)

Fig. 5 The Gini coefficient values for Wikipedia articles classified by various measures. Each sample in the data points corresponds to an individual Wikipedia article. The average coefficient is plotted as functions of (a) N_p^*/N_e^* , (b) the estimator of q : ratio of the number of edits in 2014 to the page view statistics, and (c) the article size at the end of 2014. Panels (d) and (e) display the observation from our model, as functions of (d) q and (e) the time step. The model results in (d) and (e) are also obtained using the same condition as in Fig. 4 but with the fixed $p=0.5$.

의 경우 한 번의 편집을 위해 얼마나 많은 편집자가 필요한지, 즉 그림 3의 (a)에 대응되는 그래프이다. 이 그림에서 우리는 재미있게도 위키백과를 참조할 확률인 p 와는 무관하게 위키백과를 수정할 확률인 q 에만 의존하여 변화한다는 사실을 확인할 수 있었다. 유사하게 그림 4(b)의 경우 그림 3(b)에 대응되는 그래프인데, 정반대로 q 와는 거의 무관하게 p 에만 의존하는 것을 볼 수 있었다. 즉 그림 3에서의 실제 위키백과 분석 결과와 유사하게 최종적인 문서의 형태가 p 와 q 라는 2가지 변수에 따라 4가지로 구분된다는 것을 그림 4에서 확인할 수 있으며, 이런 구分的 기준은 서로 독립적이라는 것도 파악할 수 있었다. 위의 분석을 통해 우리가 제시한 모형이 위키백과 성장을 잘 설명한다는 것을 알 수 있었으며, 이러한 모형을 통해 위키백과를 대상으로 한 추가적인 분석 진행의 가능성을 열게 되었다.

위키백과의 현재와 미래

마지막으로 본 연구에서는 편집의 수에 비해 참여자가 적어지는 것이 소수의 편집자(super editor)^[11]에 의한 독점 현상에

REFERENCES

- [10] P. Erdős and A. Rényi, Publ. Math. Inst. Hungar. Acad. Sci. 5, 17 (1960).
- [11] Y. Gandica, J. Carvalho and F. Sampaio dos Aidos, Phys. Rev. E 91, 012824 (2015).

기인하는지 분석하였다. 이를 위해 일반적으로 사용하는 불평등 지표인 지니계수를 사용하였는데, 1에 가까울수록 불평등한 상태라는 것을 의미한다.^[12] 그림 5(a)를 보면 단위 편집당 편집자의 수가 감소할수록 지니계수가 커지는 것을 볼 수 있는데, 이는 소수 편집자의 독점현상이 커질수록 참여하는 편집자의 수가 줄어든다는 사실을 의미한다. 문서의 크기가 이러한 불평등성에 큰 영향을 주지 않는다는 점을 고려하면 [그림 5(c)] 단순히 미리 만들어진 정보량이 커서 새로운 편집자가 진입하기 힘든 것이 아니라^[13] 구조적인 문제에 의해 새로운 편집자가 진입하기 힘든 것이라는 사실을 추정할 수 있다.

역설적이게도 모형에서 q 값을 높여 수정을 쉽게 만들면 이러한 불평등이 더 증가하는 모습을 볼 수 있었는데 [그림 5(d)], 문서 형성 초기에 소수의 편집자(super editor)들이 문서를 쉽게 독점하기 때문으로 볼 수 있다. 더 우려되는 점은 이러한 불평등이 시간에 따라 점점 더 증가할 것이라는 사실이다 [그림 5(e)]. 누구나 편집할 수 있다는 위키백과의 가장 큰 장점이 독점에 의해 사라지고 있는 것이다.

맺음말

정보화 시대는 인류에게 지식의 형성을 위한 폭넓은 협력을 가능하게 해주었다. 위키백과는 이러한 지식 공유 자산의 대표적인 예로, 지난 10여 년간 전 세계 사람들의 지식을 공유하는 장으로 활용되어 왔다. 이번 연구는 현재까지 위키백과가 어떤 식으로 발전하고 변화해 왔으며, 어떤 미래를 가지게 될지 예측하고 분석하는데 그 의의를 둘 수 있다.

점점 높아지는 정보 처리 능력을 통해 우리는 기존에 분석할 수 없던 대용량 데이터(big data)라는 새로운 분야를 개척할 수 있었다. 본 연구는 한걸음 더 나아가 이러한 대용량 데이터와 물리학적 모형의 결합을 통해 데이터만으로 얻을 수

없던 구조에 대한 정보를 얻어내었다. 우리는 현재 위키백과 문서들에는 명시적인 소유권이 없음에도 불구하고, 소수의 편집자들이 독점하는 상태이며 독점이 점점 심화됨을 확인할 수 있었다. 현재 위키백과의 높은 신뢰도에 이런 소수 편집자들의 공이 있다는 사실을 부정할 수는 없다.^[14,15] 하지만 이러한 소수의 노력만으로 위키백과와 같은 지적자산을 유지하는 것은 거의 불가능하다.^[16] 이미 우리의 우려와 같이 위키백과의 성장이 정체되고 있다는 것이 보고되었으며,^[17] 이를 해결하기 위해서는 다양한 대책이 강구되어야만 할 것이다.

빅 데이터(big data)와 통계 물리적 모형을 결합한 이러한 방식의 연구는 기존에 알기 힘들었던 것들에 대한 새로운 통찰력을 얻을 수 있는 방식이라 생각된다. 우리는 이러한 기초가 지속되어 다양한 대상에 대하여 후속연구가 이어지기를 바란다. 말로 이 글을 끝맺고자 한다.

그림 저작권 관련 정보

본 글에 사용된 모든 그림은 <http://journals.aps.org/pre/abstract/0.1103/hysRevE.93.012307> 논문에 사용된 그림입니다.

REFERENCES

- [12] C. Gini, Variabilità e mutabilità (Variability and mutability) (C. Cuppini, Bologna, 1912).
- [13] S. L. Bryant, A. Forte and A. Bruckman, *Proceedings of the 2005 International ACM SIG-GROUP Conference on Supporting Group Work* (GROUP '05) (ACM, New York, 2005).
- [14] T. Chesney, *First Monday* **11**, 11 (2006).
- [15] J. Giles, *Nature* **438**, 900 (2005).
- [16] J. Yun, P.-J. Kim and H. Jeong, *PLoS One* **10**, e0117388 (2015).
- [17] B. Suh, *Proceedings of the Fifth International Symposium on Wikis (WikiSym '09)* (ACM, New York, 2009), Article No. 8.