Journal of KIIT. Vol. 19, No. 3, pp. 9-14, Mar. 31, 2021. pISSN 1598-8619, eISSN 2093-7571 9
http://dx.doi.org/10.14801/jkiit.2021.19.3.9

Word2Vec의 다차원 특징을 이용한 머신러닝 기반 추천 정확도 개선

박세준*, 변영철**

Improving Recommendation Accuracy based on Machine Learning using Multi-Dimensional Features of Word2Vec

Sejoon Park*, Yungcheol Byun**

본 연구는 중소벤처기업부와 한국산업기술진흥원의 "지역특화산업육성(R&D, S2855401)"사업의 지원을 받아 수행된 연구결과임

요 약

YouTube, Netflix등의 인터넷 서비스는 각자의 독자적인 추천 시스템을 구축하여 큰 성장을 했다. 본 논문은 추천 시스템에 대한 중요도가 증가하는 가운데 Word2Vec의 다차원 특징을 이용한 아이템 추천 방식을 제안한다. Word2Vec을 이용하여 아이템 간의 연관성을 찾고 학습 데이터 차원의 증가에 따라 추천 정확도 또한 증가하는 것을 확인했다. 연구 결과 Word2Vec의 벡터 차원별로 XGBoost의 분류모델을 이용하여 추천 정확도를 구했을 때 Word2Vec 1차원 벡터의 추천 정확도는 83.73%에서 4차원 벡터 추천 정확도 85.79%로 증가하는 것을 확인하며 차원이 일정수준까지 증가할수록 추천 정확도도 증가한다는 것을 확인했다.

Abstract

Internet services such as YouTube and Netflix have grown greatly by establishing their own recommendation systems. This paper proposes an item recommendation method using the multidimensional features of Word2Vec while the importance of the recommendation system is increasing. Word2Vec was used to find associations between items, and it was confirmed that recommendation accuracy also increases as the learning data dimension increases. As a result of the study, when the recommendation accuracy was calculated using the classification model of XGBoost for each vector dimension of Word2Vec, Word2Vec 1 dimensional vector recommendation accuracy increased from 83.73% to 4 dimensional vector recommendation accuracy 85.79%, and it was confirmed that the recommendation accuracy increased as the dimension increased to a certain level.

Keywords
machine learning, recommender system, collaborative filtering, pattern recognition, word2vec, multi-dimension, xgboost

* 제주대학교 컴퓨터공학과
- ORCID: https://orcid.org/0000-0003-2220-9318
** 제주대학교 컴퓨터공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1579-5323

Received: Dec. 02, 2020, Revised: Feb. 02, 2021, Accepted: Feb. 05, 2021
Corresponding Author: Yungcheol Byun
Dept. of Computer Engineering, Jeju National University, Jejudaehakro 102, Jeju, Jeju Special Self-Governing Province, Korea
Tel.: +82-64-754-3657, Email: ycb@jejunu.ac.kr

Ⅰ. 서 론

2020년 현재 COVID-19가 전 세계적으로 확산됨에 따라 각 정부에서는 서로 최소한의 접촉을 하는 언택트을 하도록 권하고 있다. 이로 인해 쇼핑몰 이용자들은 직접 오프라인 매장을 이용하기보다는 인터넷을 이용한 온라인 쇼핑을 선호하기 시작했으며 언택트가 쇼핑에서도 활발히 일어나고 있다. 한국 통계청 자료에 따르면 본격적으로 한국에 COVID-19가 확산된 2020년 2월부터 2020년 9월까지 온라인 쇼핑 총 거래액이 전년 동월대비 평균 15%이상 증가한 것으로 나타났다[1]. 언택트 활동이 활발하게 일어나 온라인 쇼핑 이용자의 증가로 추천 시스템의 중요성이 더욱 증가하고 있다.

스마트폰의 보급과 모바일 인터넷 기기의 발전이 활발해짐에 따라 SNS나 동영상 스트리밍, 온라인 쇼핑몰과 같은 인터넷 서비스는 기존의 웹에서 시작하여 점차 모바일 인터넷 서비스로도 간편하게 이용할 수 있게 됐다[2]. 인터넷을 이용하는 이용자들의 증가로 인터넷이 다루는 데이터의 규모 또한 커졌다. 그로 인해 인터넷 이용자들은 많은 인터넷 정보 속에서 자신이 원하는 정보를 얻는데 많은 시간을 보낸다. 인터넷 이용자들이 원하는 정보를 쉽게 찾을 수 있도록 돕기 위해서는 추천 시스템이 중요하다. 또한 YouTube, Netflix 등의 세계적인 동영상 스트리밍 서비스는 자신들만의 독자적인 추천 시스템을 구축하여 서비스 이용자들의 니즈를 충족시켰다[3].

추천 시스템은 인터넷 서비스가 사용자에게 제공하는 가장 일반적인 시스템이다. 인터넷 서비스는 사용자에게 최대한의 편의를 제공하려고 노력하며 발전함에 따라 인터넷 서비스 사용자들은 계속 발전하는 서비스를 이용하면서도 끊임없이 편의를 요구한다. 이 모든 것이 활발한 소비활동 있기에 톱니바퀴처럼 돌아가고 있는 것이다[4].

본 논문은 2장에서 관련 연구와 기술에 대해서 설명한다. 추천 시스템과 Word2Vec 기법, 추천 시스템의 성능 향상을 위해 Word2Vec을 이용한 연구에 대해 언급한다. 3장에서는 본 연구에서 Word2Vec과 머신러닝의 역할과 사용한 데이터 셋에 대해 설명한다. 4장은 연구 환경에 대한 소개와 연구 결과를 언급한다.

Ⅱ. 관련 연구 및 기술

2.1 추천 시스템

대표적인 추천 시스템은 바로 협업 필터링이다. 협업 필터링은 사용자 기반의 협업 필터링과 아이템 기반의 협업 필터링으로 나뉜다.

우선 사용자 기반 협업 필터링은 사용자를 중점으로 사용자 간의 유사한 점을 찾은 후 그 결과 값을 기반으로 추천해주는 형태이다. 다음으로 아이템 기반 협업 필터링은 사용자 기반 협업 필터링과는 반대로 아이템 간의 유사성을 찾고 해당 아이템과 비슷한 유사도의 아이템을 추천해주는 방식이다. 정리하자면 사용자 기반 협업 필터링은 작은 관점에서 사용자 개개인에 따라 아이템을 추천해주는 방식이고 아이템 기반 협업 필터링은 큰 관점에서 볼 때 전체 사용자가 함께 선호했던 아이템을 추천해주는 방식이다[5][6].

2.2 Word2Vec

Word2Vec은 단어 간의 연관성을 찾고, 단어 데이터들을 벡터상의 수치로 표현한다. Word2Vec에서 단어 간의 연관성을 찾는 방법에는 두 가지가 있다.

우선 CBOW(Continuous Bag of Words)는 문장 내에 있는 모든 단어들을 각각의 단어로써 나눈다. 처음 단어부터 마지막 단어까지 순차적으로 슬라이딩을 하여 해당 단어와 주변의 단어의 연관성을 찾는다. 단어 데이터들은 벡터상의 수치로 표현되며, 다른 데이터에 적용시킬 때 주변 단어들을 토대로 중심 단어를 예측하는 것이 CBOW이다.

다음으로 Skip-Gram은 단어 간의 연관성을 찾는 방법은 CBOW와 같으나 중심 단어를 토대로 주변 단어를 예측해내는 방법으로 단어를 추출해내는 방법은 CBOW와 반대이다[7].

본 연구는 Word2Vec의 Skip-Gram 방식을 사용하여 구매할 아이템을 예측하고 각각으로 나눈 단어를 하나의 고유 아이템으로 인식하고 학습을 진행했다.

2.3 추천 시스템의 성능을 높이기 위해 Word2Vec을 이용한 연구

해당 연구는 사용자 기반 협업 필터링의 예측 정확도를 높이기 위해서 Word2Vec을 활용했다. 각 사용자의 쇼핑 클릭 내역을 시간 순서대로 나열한 데이터에 Word2Vec을 이용하여 연관성을 찾았다. 중심 아이템을 중심으로 이전과 이후에 클릭한 아이템의 연관성을 찾을 때 서로 여러 번 중복된 아이템일수록 벡터상의 수치가 가깝게 표현된다. 이러한 Word2Vec의 특성을 이용하여 추천 시스템에 적용했을 때 추천 정확도는 Word2Vec을 이용하여 연관성을 찾기 전 추천 정확도보다 2.1%의 추천 정확도 증가를 보였다[8].

III. Word2Vec 다차원 특징 기반 추천 기법

본 연구에서 제안하는 방법은 Word2Vec 사용 유무에 따른 추천 정확도의 변화를 확인한 연구에서 Word2Vec의 다차원 특징을 이용한 연구로 발전하여 진행하였다. Word2Vec은 벡터의 차원을 조절하는 하이퍼 파라미터 Size에 값을 설정할 수 있고 사용자는 벡터의 차원을 데이터에 따라 자유자재로 설정할 수 있다. 본 연구는 해당 하이퍼 파라미터인 Size, 즉 벡터의 차원이 증가함에 따라 추천 정확도가 증가하는 것을 확인했다.

Word2Vec은 데이터에 윈도우 사이즈를 설정할 수 있으며 예시 그림 1에서의 윈도우 사이즈는 2이다.

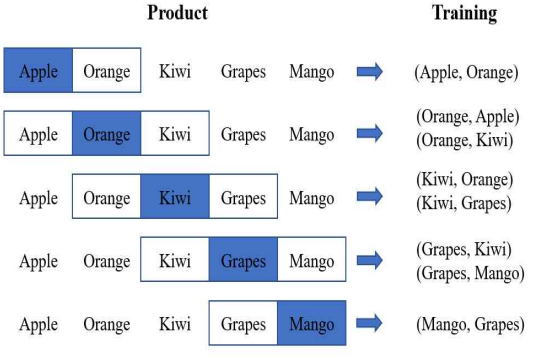

그림 1. Word2Vec Skip-Gram 예시
Fig. 1. Example of Word2Vec Skip-Gram

윈도우 사이즈는 설정한 값에 따라 중심이 되는 단어와 중심 단어 앞과 뒤에 단어 각각 2개씩을 연관시킨다. 서로 자주 중복된 단어들은 벡터상의 수치에서 가까운 위치에 나타낸다.

표 1. 데이터 셋
Table 1. Dataset for recommendation

Total number of records	10000
Maximum number of clicks	14
Minimum number of clicks	4
Training data	80%
Test data	20%

그림 1과 같이 Word2Vec의 특성을 온라인 쇼핑에 적용하는 것이 본 논문의 추천 방법이다. 표 1은 '이제주몰'을 이용한 이용자들의 클릭 내역, 즉 쇼핑 내역으로 본 논문에 사용한 데이터 수에 대한 표이다. 전체 데이터의 개수는 10000개이며, 50일간 10000명의 해당 온라인 쇼핑몰 이용자들의 쇼핑 내역을 나열해 놓은 데이터이다. 데이터 셋의 형태는 표 2와 같다.

표 2. 실험 데이터 셋 형태
Table 2. Experimental data set's shape

User / Click	1	2	…	14
User A	Item	Item	…	Item
User B	Item	Item	…	Item
User C	Item	Item	…	Item
…	Item	Item	…	Item

데이터 수집 단계에서는 각각의 사용자들이 클릭한 아이템의 고유 번호를 클릭한 순서대로 1부터 14까지 나열한다. 해당 데이터에서는 가장 많이 클릭한 사용자의 클릭 횟수는 14번이고, 14번의 클릭보다 많아지면 최근 클릭한 14개의 아이템의 목록을 나열하고 그전 클릭한 아이템은 삭제한다. 또한 가장 마지막에 클릭한 아이템인 14번째 아이템은 해당 사용자의 최종 구매 아이템이다.

본 논문은 Word2Vec을 이용하여 아이템들 간의 연관성을 찾고 수치로 나타낸 데이터를 머신러닝 모델에 학습시킨다. 1부터 13까지의 아이템을 인풋 데이터로, 마지막 구매한 14번째 아이템을 아웃풋데

이터로 학습하여 인풋데이터의 패턴을 학습하여 결국에 구매한 아이템을 추천하는 방식이다.

본 논문은 Word2Vec을 이용하여 온라인 쇼핑 아이템을 추천하였을 때 Word2Vec을 적용시킴으로써 추천 정확도의 증가를 확인했고 더 나아가 Word2Vec의 다차원 특징을 이용하여 벡터 차원의 증가에 따른 추천 정확도 증가를 확인했다.

Ⅳ. 연구 환경 및 성능 평가

4.1 연구 환경

본 연구는 Word2Vec의 벡터 차원 증가에 따른 학습시간을 비교한다. 벡터 차원이 증가할수록 데이터의 양은 그만큼 증가하기 때문에 추천 정확도와 학습시간을 고려하여 좋은 결과의 차원을 추천한다. 연구 환경에 따라 학습에 걸리는 시간은 차이가 있기에 본 연구를 진행한 연구 환경에 대한 정보는 표 3과 같다[9].

표 3. 연구 환경
Table 3. Research environment

Programming language	Python 3.7.6
Operating system	Window 10 pro 64bit
Browser	Google chrome
Library and framework	Jupyter notebook
CPU	Intel(R) Core(TM) i5-9600k@3.70GHz
Memory	16GB

4.2 성능 평가

본 연구는 Word2Vec 사용 유무에 따른 추천 정확도의 변화를 확인한 연구에서 Word2Vec의 다차원 특징을 이용한 연구로 발전하여 진행하였다. Word2Vec은 벡터의 차원을 조절하는 하이퍼 파라미터 Size가 있어 프로그래머가 벡터의 차원을 설정할 수 있다. Word2Vec의 벡터 차원을 1차원부터 5차원까지 증가시키면서 추천 정확도를 측정하였다. 머신러닝 모델은 XGBoost의 분류 모델인 XGBoost Classifier를 이용하여 학습을 진행하였고 학습에 대한 결과는 다음과 같다.

Word2Vec 차원별 벡터를 학습했을 때 결과는 표 4와 같다. 각 차원에서 총 10번의 학습을 진행하였고 1차원 벡터에서 평균 83.73%의 추천 정확도를 보였다. 다음으로 2차원 벡터의 학습 결과는 평균 85.63%로 Word2Vec 1차원 벡터의 결과보다 1.9% 증가하는 것을 확인했다. Word2Vec 3차원 벡터를 학습했을 때 평균 85.68%로 Word2Vec 2차원 벡터의 결과보다 0.05% 증가함을 확인했다. Word2Vec 4차원 벡터의 학습 결과, 평균 85.79%의 추천 정확도가 나타났다. 3차원 벡터의 결과보다 4차원 벡터에서 결과가 0.11% 증가하는 것을 확인했다[10].

표 4. Word2Vec 차원별 벡터의 추천 정확도
Table 4. Recommendation accuracy according to Word2Vec dimensional vector

Word2Vec dimension	Recommendation accuracy (%)
1 Dimension	83.73
2 Dimension	85.63
3 Dimension	85.68
4 Dimension	85.79
5 Dimension	85.72

Word2Vec 5차원 벡터에서 추천 정확도 결과에 대한 표이다. 10번 학습하여 평균 85.72%의 추천 정확도를 보였다. 4차원 벡터에서 5차원 벡터로 학습을 진행했을 때 추천 정확도는 0.07% 감소하는 것을 확인했다.

앞서 결과에서 확인할 수 있듯이 1차원 벡터에서 2차원 벡터로 진행했을 때 추천 정확도는 가장 크게 증가했지만 2차원 벡터에서 4차원 벡터까지 학습을 진행했을 때 미미하게 증가하다 5차원 벡터에서 추천 정확도가 감소하는 것을 확인했다. 이는 Word2Vec의 벡터 차원이 증가할수록 그에 대한 예측 정확도도 증가하다가 일정 수준의 벡터 차원에 도달하게 되면 예측 정확도는 감소하는 것을 의미한다.

그림 2는 Word2Vec 벡터 차원에 따른 XGBoost Classifier의 추천 정확도에 대한 그림이다. Word2Vec 1차원 벡터에서 2차원 벡터로 증가할 때 눈에 띌 정도의 추천 정확도 증가를 보이다 이후 미미하게 추천 정확도가 증가하다가 5차원 벡터에서 추천 정확도가 감소하는 것을 보였다.

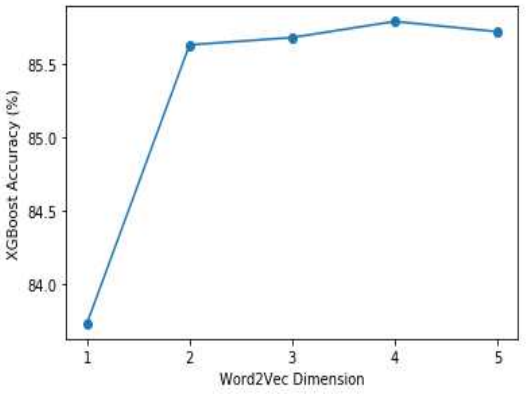

그림 2. Word2Vec 벡터 차원에 따른 XGBoost Classifier 추천 정확도

Fig. 2. Recommendation accuracy according to Word2Vec dimensional vector using XGBoost classifier

그림 3은 Word2Vec 벡터 차원에 따른 XGBoost Classifier의 학습시간에 대한 그림이다. Word2Vec의 벡터 차원이 증가할수록 학습할 데이터의 수가 증가함에 따라 학습에 필요한 시간 또한 시간과 비례하여 증가했다.

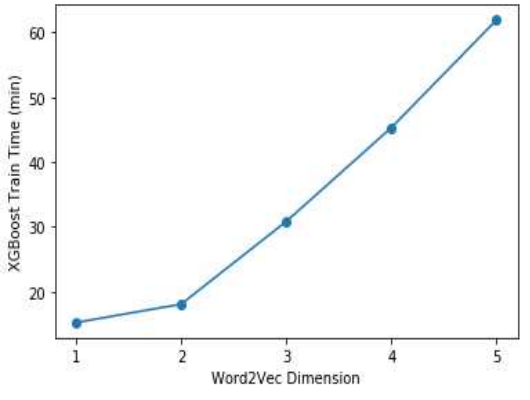

그림 3. Word2Vec 벡터 차원에 따른 XGBoost Classifier 학습시간

Fig. 3. Train time according to Word2Vec dimensional vector using XGBoost classifier

본 연구에서는 추천 정확도와 학습 속도를 비교해보았을 때 1차원 벡터에서 2차원 벡터로 증가했을 때의 추천 정확도의 폭과 학습시간을 고려하여 추천 정확도와 학습 속도 대비 Word2Vec 2차원 벡터에서의 학습이 가장 좋은 결과를 보였다.

4.3 교차 검증

기존 머신러닝은 Train 데이터와 그것을 기반으로 예측한 결과인 Test 데이터로 성능을 평가했다. 그러나 교차 검증은 Train 데이터와 Test 데이터 사이에 Validation 데이터를 넣어 과적합이 일어나는지를 확인할 수 있으며 예방할 수 있다. 본 연구에서도 Validation 데이터를 추가하여 60%의 Train 데이터와 20%의 Validation 데이터, 20%의 Test 데이터를 통해 추천 정확도를 나타낸 결과는 표 5와 같다. 추천 정확도는 Validation 데이터와 Test 데이터의 평균을 나타냈다.

표 5. Word2Vec 차원별 벡터의 교차 검증을 이용한 추천 정확도

Table 5. Recommendation accuracy using cross-validation of vector according to Word2Vec dimensional vector

Word2Vec Dimension	Recommendation accuracy (%)
1 Dimension	81.10
2 Dimension	83.05
3 Dimension	83.75
4 Dimension	84.30
5 Dimension	84.55

V. 결 론

본 논문은 Word2Vec의 다차원 특징을 이용하여 학습을 진행하였을 때 차원의 증가에 따라 정확도의 증가가 이루어지는 것을 확인했다. 연구 결과에서 알 수 있듯이 차원이 증가함에 따라 추천 정확도도 증가하는 것을 확인할 수 있었다. 대표적으로 XGBoost의 경우 1차원에서부터 4차원까지 추천 정확도가 이루어지다가 4차원에서의 85.79%의 추천 정확도에서 5차원의 85.72%로 떨어지는 것을 확인했다. 추천 정확도는 일정수준의 벡터 차원에 이르기 전까지 증가하는 것을 확인했다. 일정수준의 벡터 차원은 학습에 사용하는 데이터와 모델에 따라 다르지만 그전까지의 차원의 증가는 추천 정확도의 증가에 영향을 미치는 것을 확인할 수 있다.

Word2Vec의 벡터 차원을 1차원부터 5차원까지 학습하여 나온 추천 정확도와 학습시간을 비교하여 가장 결과가 좋은 Word2Vec 차원을 도출했다. 본 연구에서는 1차원부터 차원을 하나씩 증가시킬 때

마다 추천 정확도와 학습시간의 차이를 정리했다. 총 5차원까지 각 차원별 증가폭을 고려했을 때 Word2Vec 2차원에서 학습을 진행했을 때가 추천 정확도의 증가폭이 높고 학습시간은 적어 가장 좋은 결과를 보였다.

References

[1] Korea National Statistical Office, https://www.kostat.go.kr, [accessed : Oct. 16, 2020]

[2] Yunju Lee, Haram Won, aeseung Shim and Hyunchul Ahn, "A hybrid Collaborative Filtering-based Product Recommender System using Search Keywords", Journal of Intelligent Information Systems Society, Vol. 26, No. 1, pp. 151-166, May 2020. https://doi.org/10.13088/jiis.2020.26.1.151.

[3] Chang-Hwab Son, Jin-Ouk Kim and Gui-Ryong Ha, "A Study on Development of Hybrid Collaborative Filtering Algorithm", Journal of Business Research, Vol. 25, No. 4, pp. 47-66, Apr. 2010. https://doi.org/10.22903/jbr.2010.25.4.47.

[4] Soojung Lee, "Using Genre Rating Information for Similarity Estimation in Collaborative Filtering", Journal of the Korea society of computer and information, Vol. 24, No. 12, pp. 93-100, Dec. 2019. https://doi.org/10.9708/jksci.2019.24.12.093.

[5] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems" The adaptive web. Springer. Berlin. Heidelberg, pp. 291-324, 2007.

[6] Chul-Jin Kim, Ji-Hyun Jeong, Cheon-Woo Jo and Je-Kwang Yoo, "A Performance Evaluation Analysis of Product Recommendation Techniques", Journal of Knowledge Information Technology and Systems, Vol. 14, No. 5, pp. 515-525, May 2019. https://doi.org/10.34163/jkits.2019.14.5.008.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality", Advances in neural information processing systems 26, pp. 3111-3119, Oct. 2013.

[8] Se-Joon Park and Yung-Choel Byun, "A hybrid Collaborative Filtering based on Online Shopping Patterns using XGBoost and Word2Vec", The Journal of KIIT, Vol. 18, No. 9, pp. 1-8, Sep. 2020. https://doi.org/10.14801/jkiit.2020.18.9.1

[9] Zeinab Shahbazi and Yung-Cheol Byun, "Product Recommendation Based on Content-based Filtering Using XGBoost Classifier", International Journal of Advanced Science and Technology, Vol. 29, No. 4, pp. 6979-6988, Apr. 2020.

[10] Prince Waqas Khan and Yung-Cheol Byun, "Genetic Algorithm Based Optimized Feature Engineering and Hybrid Machine Learning for Effective Energy Consumption Prediction", IEEE Access 8, pp. 196274-196286, Oct. 2020. https://doi.org/10.1109/ACCESS.2020.3034101

저자소개

박 세 준 (Sejoon Park)

2020년 2월 : 제주대학교 컴퓨터공학과(공학사)
2020년 3월 ~ 현재 : 제주대학교 컴퓨터공학과 석사과정
관심분야 : 인공지능(머신러닝, 딥러닝), 자연어처리, 인지과학

변 영 철 (Yungcheol Byun)

1993년 2월 : 제주대학교 컴퓨터공학과(공학사)
1995년 8월 : 연세대학교 컴퓨터과학과 (공학석사)
2001년 8월 : 연세대학교 컴퓨터과학과 (공학박사)
1998년 ~ 2001년 : 삼성전자, SDS 전문강사
2001년 ~ 2003년 : 한국전자통신 연구원 선임연구원
2003년 ~ 현재 : 제주대학교 컴퓨터공학과 교수
2012년 ~ 2014년 : University of Florida 방문 교수
관심분야 : 딥러닝, 패턴인식, 시계열 데이터 처리, 추천 시스템, 지식발견, 딥러닝 기반 신재생에너지 시스템, 블록체인