Regular paper

Journal of the KIECS. pp. 801-808, vol. 17, no. 5, Oct. 31. 2022, t. 115, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2022.17.5.801

# 의미적 유사성과 그래프 컨볼루션 네트워크 기법을 활용한 엔티티 매칭 방법

단홍조우[*] · 이용주[**]

# Entity Matching Method Using Semantic Similarity and Graph Convolutional Network Techniques

Hongzhou Duan[*] · Yongju Lee[**]

## 요 약

대규모 링크드 데이터에 어떻게 지식을 임베딩하고, 엔티티 매칭을 위해 어떻게 신경망 모델을 적용할 것인가에 대한 연구는 상대적으로 많이 부족한 상황이다. 이에 대한 가장 근본적인 문제는 서로 다른 레이블이 어휘 이질성을 초래한다는 것이다. 본 논문에서는 이러한 어휘 이질성 문제를 해결하기 위해 재정렬 구조를 결합한 확장된 GCN(Graph Convolutional Network) 모델을 제안한다. 제안된 모델은 기존 임베디드 기반 MTransE 및 BootEA 모델과 비교하여 각각 53% 및 40% 성능이 향상되었으며, GCN 기반 RDGCN 모델과 비교하여 성능이 5.1% 향상되었다.

## ABSTRACT

Research on how to embed knowledge in large-scale Linked Data and apply neural network models for entity matching is relatively scarce. The most fundamental problem with this is that different labels lead to lexical heterogeneity. In this paper, we propose an extended GCN (Graph Convolutional Network) model that combines re-align structure to solve this lexical heterogeneity problem. The proposed model improved the performance by 53% and 40%, respectively, compared to the existing embedded-based MTransE and BootEA models, and improved the performance by 5.1% compared to the GCN-based RDGCN model.

Keywords

Entity Matching, Lexical Heterogeneity, Graph Convolution Network, Hamming Distance, Entity Embedding
엔티티 매칭, 어휘 이질성, 그래프 컨볼루션 네트워크, 해밍 거리, 엔티티 임베딩

## Ⅰ Introduction

In recent years, with the rapid development of artificial intelligence, Linked Data research by utilizing machine learning and Big Data technology is gradually becoming a hotspot[1]. However, there

[*] 경북대학교 IT대학 컴퓨터학부
(caixiuming1984@163.com)
[**] 교신저자 : 경북대학교 IT대학 컴퓨터학부

· 접 수 일 : 2022. 07. 25
· 수정완료일 : 2022. 09. 02
· 게재확정일 : 2022. 10. 17

· Received : Jul. 25, 2022, Revised : Sep. 02, 2022, Accepted : Oct. 17, 2022

· Corresponding Author : Yong-Ju Lee
School of Computer Science and Engineering, Kyungpook National University,
Email : yongju@knu.ac.kr

are relatively few studies on how to embed knowledge into massively Linked Data and apply it to train neural network models for entity matching. The most fundamental problem is that different labels lead to lexical heterogeneity. Figure 1 illustrates the lexical heterogeneity problem with an example. In the two knowledge graphs KG1 and KG2, the ellipse is the entity identifier, and the rectangle is the attribute value, where “BMW 3 Series Compact Sport Sedam” and “BMW 320i m sport” do not match.

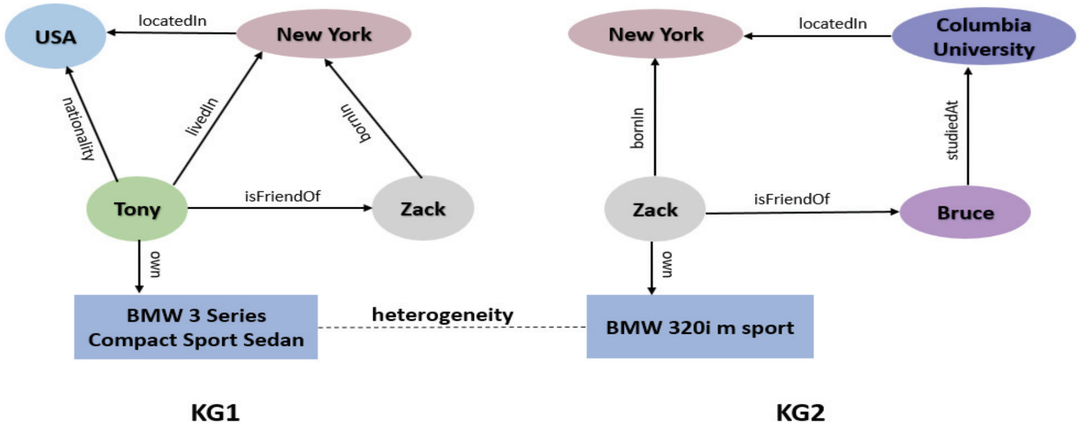

Fig. 1 An example of lexical heterogeneity problem

Despite recent considerable progress in applying embedding techniques to the field of link auto-completion, the accuracy of knowledge graph entity matching based on these techniques is still insufficient. In this paper, to address this lexical heterogeneity problem, we propose an extended GCN (Graph Convolutional Network) model combining re-align structure. It can better express complex edge structures and relationships. Compared with the traditional embedded-based MTransE[2] and BootEA[3] models, the improvements are 53% and 40%, respectively. Compared with the GCN-based RDGCN[4] model, the performance is improved by 5.1%.

The rest of the papers is organized as follows. Related works are introduced in Section 2 including word embedding, knowledge graph embedding, and GCN. In Section 3, we present an extended GCN model and a re-align structure. Experiment and analysis are described in Section 4. Finally, in Section 5, conclusions are drawn and future research is suggested.

## II Related works

### 2.1 Word embedding

Word embedding is the search for the vector value corresponding to each word so that the similar ones are located closer to each other. The typical research includes NNLM[5], Word2Vec[6], and Glove[7]. However, input data of word embedding use preprocessed text-based corpus and knowledge graph (KG) data are stored in triple structure of the RDF form[8], so graph-embedding technique should be used for KG rather than word embedding. RDF2Vec[9] is a method for embedding graphs, but vectors embedded from RDF2Vec are not optimized for entity matching that is related to specific relation; the matching method for entity matching is not optimized for entity matching.

### 2.2 Knowledge graph embedding

The TransE[10] model proposed by Google learns that sum of the head and relation vector is equal to tail. TransE performs better when it is a 1:1 relationship, but problems occur when it is 1:N, N:1, and N:N. The TransH[11] model solved TransE problem by projecting relations on hyperplane. However, entity (head, tail) and relation are clearly different objects, so TransR[12] proposed the idea of embedding entity and relation in separated space. Meanwhile ConvE[13] conducted 2D convolution against the embedding vector and applied nonlinearity to increase the expressive power of the vector. Until recently, although considerable progress has been made in research on the application of knowledge graph embedding techniques in the field of link auto-completion, there is still a lack of accuracy to utilize these techniques to carry out entity matching. The reason is that entity matching considers only the graph structure created from embeddings, so it is difficult to make detailed matching for terminal entities.

### 2.3 GCN

GCN presented a new direction as a technique that used convolutional operations of GNN (Graph Neural Networks)[14] which models correlations between nodes and nodes in a graph. This approach has limitations in properly modeling relational information. Because general GCN does not have direction and operates on unlabeled graphs, and useful relational information in the knowledge graph is not properly utilized. RGCN (Relational GCN)[15] can be used to model multiple relationship graphs, but this requires too excessive set of parameters, using separate weighting matrices for each relationship. DPGCNN (Dual-Primal Graph CNN)[16] carried out convergence operation in turn on the dual graph corresponding to the original graph. In order to improve the edge representation, the model learns vertex and the edge feature based on graph attention scores. Meanwhile, under the inspiration of DPGCNN, RDGCN (Relation aware Dual-graph Convolutional Network)[4], which could better express the relationship and characterize the relationship between different knowledge graphs, was proposed. However, DPGCNN and RDGCN can serve as a good starting point, but the accuracy of matching is still low when using deep learning, so it is necessary to increase the accuracy by considering both word and semantic similarity. Table 1 summarizes above related works.

Table 1. Summary of related works

| Category | Characteristic | Weakness |
|---|---|---|
| Word embedding | Use preprocessed text-based corpus | KG must use graph embedding methods |
| Knowledge graph embedding | Sum of the head and relation is equal to tail | Lack of accuracy to carry out entity matching |
| GCN | Model multiple relationship graphs | Accuracy of matching is still low |

## III Entity matching method using semantic similarity and GCN techniques

### 3.1 Relation aware GCN model

In this paper, graphs with dual relations are constructed from original graphs and then dual relations of these graphs are computed, which are learned through mutual interactions with dual graphs and original graphs. The role of dual graphs facilitates better integration into existing graph representations. Figure 2 shows overview of the learning model for dual relational graphs. DAL

(Dual Attention Layer) interactively affects the representation of graphs in PL (Primal Layer) for more accurate relational integration.

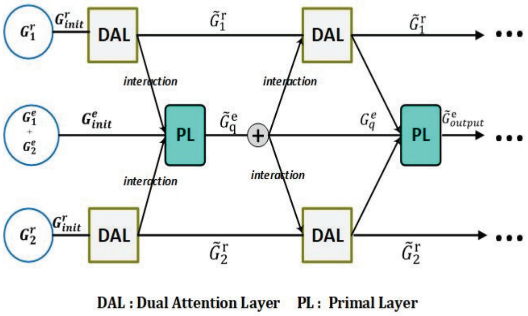

Fig. 2 Learning model of relation-aware dual graph

Eq. (1) and Eq. (2) represent the processing of DAL. $\widetilde{G_i^r}$ in Eq. (1) represents the d-dimensional output representation of a vertex $v_i^r$, where $N_i^r$ represents the set of adjacent indices. $\alpha_{ij}^r$ is a dual attention score, and ELU is the activation function. In Eq. (2), $a^r$ is a fully connected layer[17] that maps the 2-dimensional input to a scalar, $c_i$ is the relation representation generated from the previous PL. $w_{ij}^r$ represents the weight matrix in DAL.

$$\widetilde{G_i^r} = ELU\left(\sum_{j\in N_i^r} \alpha_{ij}^r G_j^r\right) \qquad (1)$$

$$\alpha_{ij}^r = \frac{\exp\left(ELU\left(w_{ij}^r a^r\left[c_i \parallel c_j\right]\right)\right)}{\sum_{k\in N_i^r}\exp\left(ELU\left(w_{ik}^r a^r\left[c_i \parallel c_k\right]\right)\right)} \qquad (2)$$

Eq. (3) and Eq. (4) represent the processing of PL. In Primal Layer, the vertex embedding is affected by using the relational expression created in the Dual Attention Layer. Eq. (3) and Eq. (4) show the scores of the primal attention in the Primal Layer. $\widetilde{X_{qt}^r}$ represents the output expression obtained from the dual graph, and Leaky ReLU is the activation function of Primal Layer. $\alpha_{qt}^e$ is a primal attention score from Primal Layer. $a^e$ is a fully connected layer that maps the 2-dimensional input to a scalar.

$$\widetilde{G_q^e} = Leaky\ \ ReLU\left(\sum_{t\in N_q^e}\alpha_{qt}^e G_t^e\right) \qquad (3)$$

$$\alpha_{qt}^e = \frac{\exp\left(ELU\left(a^e\left(\widetilde{X_{qt}^r}\right)\right)\right)}{\sum_{K\in N_q^e}\exp\left(ELU\left(a^e\left(\widetilde{X_{qk}^r}\right)\right)\right)} \qquad (4)$$

After several rounds of interaction between the dual relationship graph and the primal graph, a relationship-aware entity representation can be obtained. Finally, GCN with highway gate is applied to the integrated adjacent information structure. The output expression $H^{(l+1)}$ generated by the GCN layer is as follows.

$$H^{(l+1)} = ReLU\left(G^{(l)}W^{(l)}\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}\right) \qquad (5)$$

$\widetilde{A} = A + I$ is the adjacency matrix of the primal graph $G^e$, $W$ is weights from convolutional layer, $D$ is a degree matrix. $I$ is the identity matrix, and ReLU is the activation function. Eq. (6) is a matrix of weights that can be learned for each layer. The final entity representation generated from the output of the GCN layer is sorted through the distance between the two entities.

$$\widetilde{D_{jj}} = \sum_k \widetilde{A_{jk}},\ \ H^{(l)} \in R^{d^{(l)} * d^{(l+1)}} \qquad (6)$$

### 3.2 Re-align structure based on hamming distance

To increase accuracy of matching, we propose a re-align structure based on the hamming distance (HD) shown in figure 3. First, entities are extracted from the GCN model. Then, we adopt the hamming distance to evaluate entity similarity shown in Eq. (7) Finally, the minimum hamming distance is added to the final alignment result.

$$HD = \left(\sum_{i=1}^{k} \mid G_e^1 - G_e^2 \mid \right) \qquad (7)$$

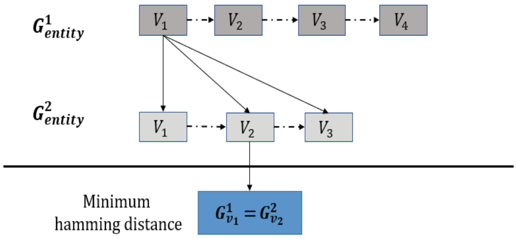

Fig. 3 Re-align structure based on hamming distance

Figure 4 shows the overall overview of our extended GCN model combining the re-align structure based on the hamming distance. In this model, we extract two knowledge graph entities from the dual graph convolutional network. At this time, the entity with the closest hamming distance becomes the matching value by the re-align structure.

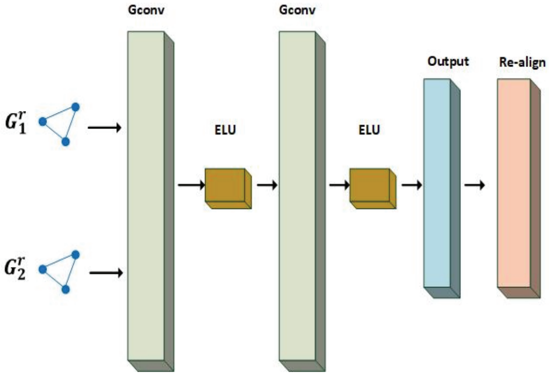

Fig. 4 Overview of the extended GCN model

## Ⅳ Experiment and analysis

### 4.1 Experiment data

Table 2. Experimental dataset

| Data set | Language | No. of relations | No. of attributes | No. of rel. triples |
|---|---|---|---|---|
| SRPRS Fr-En | French | 177 | 53,045 | 33,532 |
| | English | 221 | 60,800 | 36,508 |
| SRPRS De-En | German | 222 | 55,580 | 38,363 |
| | English | 120 | 73,753 | 37,377 |
| SRPRS DBP-WD | DBpedia | 253 | 64,021 | 38,421 |
| | WikiData | 144 | 133,371 | 40,159 |
| SRPRS DBP-YG | DBpedia | 323 | 58,853 | 33,748 |
| | YAGO 3 | 30 | 18,241 | 36,569 |

To evaluate the performance of the proposed extended GCN model, in this paper, a benchmark dataset SRPRS[18] is utilized shown in Table 2. The SRPRS dataset consists of cross-lingual (Fr-En and De-En) and mono-lingual (DBP-WD and DBP-YG) KG pairs.

### 4.2 Performance evaluation

The experiment consists of the following. First, using the SRPRS dataset, entity matching of the extended GCN model is performed to evaluate whether entities in two KGs match exactly. For experimental results, we use the Hit@K rate, which is commonly used in entity matching studies. Hit@K indicates whether the correct answer was found in the K-th among matching candidates.

Table 3 lists comparison results of the embedding-based approach (i.e., MTransE[19] and BootEA[3]), RDGCN model, and our approach. Overall, accuracy based on the GCN model (i.e., RDGCN and our approach) is much higher than that based on the embedding-based approach. Comparing the performance of the RDGCN model and our approach, Hit@1, which found entity matching at once, showed a performance improvement of about 5.1% from 69.99% to 75.1%. Hit@10 showed a slight improvement of 77.68% from the previous 76.5%. However, Hit@50 and Hit@100 did not show significant improvement. This is related to the entity distribution of SRPRS dataset. The entity sparseness in SRPRS dataset is relatively large, so the K value of Hit and accuracy do not increase in the same proportion during the matching process. Figure 5 is a visualization of performance analysis results in Table 3.

Table 3. Experimental results

| Model | Hit@1 | Hit@10 | Hit@50 | Hit@100 |
|---|---|---|---|---|
| MTransE | 22.7 | 42.68 | 58.45 | 61.3 |
| BootEA | 35.63 | 62.41 | 71.32 | 79.56 |
| RDGCN | 69.99 | 76.5 | 88.3 | 92.52 |
| Our approach | 75.1 | 77.68 | 89.12 | 93.4 |

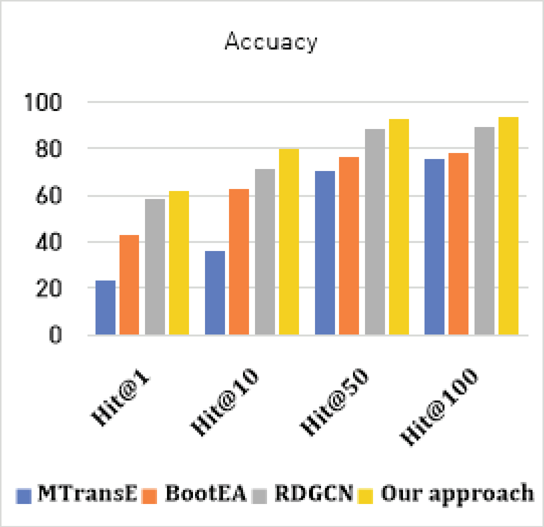

Fig. 5 Visualization of performance analysis results

## V Conclusion

Knowledge integration to utilize KGs in large-scale semantic Big Data faces the problem of entity heterogeneity. Despite of many studies using word embedding and graph embedding methods, accuracy is still not high enough to be applied to real life. Since it is learned based on the graph structure, there is a problem that fine matching of entities is not perfect.

In this paper, we proposed an extended GCN model and re-align structure to improve alignment accuracy. Considering the problem of triangular data structure due to characteristics of RDF data, we solve it by using the dual graph combined with the GCN model. In addition, since it takes a lot of time to check 1:1 entity matching of the similarity technique, the time required is reduced by using the re-align structure.

Compared with the existing RDGCN model, our approach improves accuracy by about 5.1%. The contribution of our paper is that it is possible to use dual graph convolutional networks that express complex edge structures and relationships better than existing GCN models and embedding-based approaches. However, the time required for the model proposed in this paper increases due to the expansion of similarity methods. To complement this issue, future research needs to compare the performance of other deep learning based entity matching techniques that consider the interaction between AI models and similarity techniques.

감사의 글

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1A1B02008553).

## References

[1] O. Kingsley, *Linked Open Data: State-of-the-Art Mechanisms and Conceptual Framework*. London, United Kingdom, IntechOpen, Oct. 2020.

[2] M. Chen, Y. Tian, M. Yang, and C. Zaniolo, "Multilingual Knowledge Graph Embeddings

for Cross-lingual Knowledge Alignment," In *Proc. International Joint Conference on Artificial Intelligence (IJCAI-17),* Melbourne Australia, Aug. 2017, pp. 1511-1517.

[3] Z. Sun, W. Hu, Q. Zhang, and Y. Qu, "Boot-strapping Entity Alignment with Knowledge Graph Embedding," In *Proc. International Joint Conference on Artificial Intelligence,* Stockholm, Sweden, 2018, pp. 4396-4402.

[4] Y. Wu, X. Liu, Y. Feng, Z. Wang, R. Yan, and D. Zhao, "Relation-Aware Entity Alignment for Heterogeneous Knowledge Graphs," In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, Macao, China, Aug. 2019, pp. 5278-5284.

[5] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A Neural Probabilistic Language Model," *Neural Information Processing Systems (NIPS)*, vol. 3, Jan. 2000, pp. 932-938.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," In *Proc. ICLR,* Arizona, USA, Sept. 2013.

[7] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1532-1543.

[8] Y. Sun and Y. Lee, "Word Embedding-based Space Index Structure for Knowledge Graph Search," *J. of the Korea Institute of Electronic Communication Science*, vol. 15, no. 2, Nov. 2021, pp. 192-194.

[9] P. Ristoski and H. Paulheim, "RDF2Vec: RDF Graph Embeddings for Data Mining," In *Proc. SEMWEB 17 October 2016 Computer Science,* Kobe, Japan, 2016, pp. 498-514.

[10] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating Embeddings for Modeling Multi-relational Data," *Neural Information Processing Systems (NIPS)*, Nevada, USA, Dec. 2013, pp. 2787-2795.

[11] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge Graph Embedding by Translating on Hyperplanes," In *Proc. AAAI Conference on Artificial Intelligence*, Québec City, Québec, Canada, July 2014, pp. 1112-1119.

[12] T. Trouillon, C. R. Dance, E. Gaussier, J. Welbl, S. Riedel, and G. Bouchard, "Knowledge Graph Completion via Complex Tensor Factorization," *Journal of Machine Learning Research*, vol. 18, issue 1, Jan. 2017, pp. 4735-4772.

[13] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D Knowledge Graph Embeddings," In *Proc. ACM International Conference on Bioinformatics, Computational Biology and Health Informatics,* New Orleans, USA, Sept. 2019, pp. 113-123.

[14] F. Scarselli, M. Gori, A. Tsoi, M. Hagenbuchner, and G. Monfardini, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, vol. 20, issue 1, Jan. 2009, pp. 61-80.

[15] M. Schlichtkrull, T. Kipf, P. Bloem, R. Berg, I. Titov, and M. Welling, "Modeling Relational Data with Graph Convolutional Networks," In *Proc. European Semantic Web Conference (ESWC)*, Heraklion, Crete, Greece, June 2018, pp. 593-607.

[16] F. Monti, O. Shchur, A. Bojchevski, O. Litany, S. Günnemann, and M. M. Bronstein, "Dual-Primal Graph Convolutional Networks," In *Proc. Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, Dublin, Ireland, Sept. 2018, pp. 1-14.

[17] G. Choi and Y. Jeong, "Efficient Iris Recognition using Deep-Learning Convolution Neural Network," *J. of the Korea Institute of Electronic Communication Science,* vol. 15, issue 3, June 2020, pp. 521-526.

[18] L. Guo, Z. Sun, and W. Hu, "Learning to Exploit Long-term Relational Dependencies in Knowledge Graphs," In *Proc. International Conference on Machine Learning*, California, USA, 2019, pp. 2505-2514.

[19] M. Chen, Y. Tian, M. Yang, and C. Zaniolo, "Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment," In *Proc. International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2017, pp. 1511-1517.

## 저자 소개

### 단홍조우(Hongzhou Duan)

2009년 하얼빈 이공대학교 소프트웨어 공학과 졸업(공학사)
2019년 경북대학교 대학원 컴퓨터학과 졸업(공학석사)
2021년 ~현재 경북대학교 대학원 컴퓨터학과 (박사과정)
※ 관심분야 : 시맨틱 웹, 지식 그래프 임베딩, 딥러닝, 빅 데이터

### 이용주(Yong-Ju Lee)

1985년 한국과학기술원 정보검색 전공(공학석사)
1997년 한국과학기술원 컴퓨터공학전공(공학박사)
1998년 ~현재 경북대학교 IT대학 컴퓨터학부 교수
※ 관심분야 : 링크드 데이터, 시맨틱 웹, 빅데이터, 지식 그래프