Analyzing Future Nodes in a Knowledge Network

Sukhwan Jung*, Tuan Manh Lai†, Aviv Segev*†
*Graduate School of Knowledge Service Engineering
†School of Computing
Korea Advanced Institute of Science and Technology
Daejeon, South Korea
raphael@kaist.ac.kr, laituan245@kaist.ac.kr, aviv@kaist.edu

***Abstract*—The paper proposes new methods for knowledge prediction using network analytics and introduces pEgonet, sub-networks within knowledge networks consisting of to-be-neighbors of new knowledge. Preliminary results show that it is feasible to predict how future knowledge is added in the knowledge network by utilizing basic properties of pEgonet. The paper presents initial work which will be expanded to derive a method to predict labelled future knowledge, with its impact and structures.**

***Keywords*-Knowledge Network; Technology Forecasting; Network Analysis; Node Prediction;**

I. Introduction

Knowledge networks represent currently revealed knowledge in a given domain, with each node representing an individual knowledge concept and link representing their relationships. Analyzing how a new node is introduced to the knowledge network will give insight into the prediction of new knowledge in the future and how it will expand a given domain.

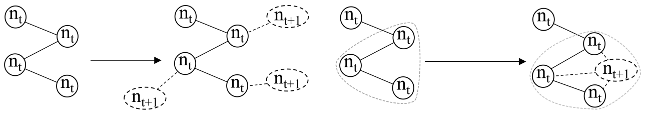
The identification of new knowledge is a task with high demand. Big data analysis focuses on efficiency when processing existing data to discover knowledge that is currently unrecognized because of data abundance. The problem of this approach is that it only works if the new knowledge in question is already present, hidden in data. The paper presents a new perspective of viewing the available data as a base for predicting the creation of new knowledge which is currently absent in the domain. The goal of this paper is to predict how new nodes will be connected to the given knowledge network in the future and to show its preliminary result.

The paper presents two methods to analyze how new knowledge appears over time in the knowledge network, which is defined as a graph representation of the domain knowledge at each timeslot, with knowledge concepts as nodes represented by keywords and common usage of concepts as links. One of the proposed methods is node-centric, which analyzes relationships between node properties and the number of its new neighboring nodes, or knowledge, and derives a function which can measure the likelihood of a given node being the connector to new nodes in the future. This method aims at measuring the impact given knowledge has in the future by measuring how many additional new knowledge concepts will stem from it. Another proposed method is a group-centric method, which analyzes structural characteristics of neighbors of a node in the past within a given knowledge network to derive a function capable of identifying sub-networks on which each new knowledge will be based. This method focuses on predicting new knowledge with comprehensive details provided by merging known properties about its neighbors such as keywords, degree, and histories. Figure 1 illustrates the differences between the two methods used in the paper. Figure 1a illustrates a node-centric approach where future nodes are predicted from each existing node, showing that three of the nodes will be neighbors of future nodes. Figure 1b shows a group-centric approach, identifying neighbors of a future node to show that those three nodes are actually common neighbors of a single future node. The impact of those identified future nodes will then be predicted by measuring how many additional newer neighbors they will have, using a combination of node-centric and group-centric methods.

The rest of the paper is organized as follows. Section II reviews the related work. Section III describes the data and methods used for analyzing future knowledge in knowledge networks. Section IV presents the preliminary experiment results, and section V provides directions of future work and concluding remarks.

II. Related Work

Many research fields aim to identify and predict new knowledge. Topic Detection and Tracking (TDT) [1] is a multi-site research project aiming to predict novel topics by effectively identifying the first article or report mentioning the new topic. The topic-conditioned First Story Detection (FSD) method [2] tried to identify the earliest report on a certain event in news articles. Manually assisted technology trend analysis was done to identify roots of new technologies with their projected impacts [3]. A similar approach was tried with multiple data sources to predict technology trends, while showing that different data sources exhibit different forecast speed [4]. All of these studies focus on analyzing unstructured documents using Natural Language Processing (NLP) technologies.



(a) Node-centric approach predicts future neighbors from every current node.

(b) Group-centric approach identifies set of nodes with common future neighbor.

Figure 1: Visual explanations of the two methods.

Researchers using structured networks of data instead of using unstructured documents in prediction fields aim to reduce the resource cost by utilizing already available structured knowledge networks such as DBpedia and Web of Science, which contain up-to-date knowledge of the world, or citation and coauthoring networks which contain activities leading up to this knowledge [5].

Many research fields related to network analysis, such as graph analysis and link predictions, can be used in knowledge prediction in structured networks. Analysis on a set of large real-world graphs across diverse domains showed that the evolving graph exhibits densifying behavior with decreasing node distances, which is in contrast to the projections made from analyzing static graphs [6]. A scalable graph anomaly detection algorithm utilizes characteristics of neighboring structures called egonet within the graph network [7]. Link prediction models the evolution of a network using its topological characteristics and primarily deals with the prediction of edges between existing nodes [8], [9]. Link prediction methods used to predict the formation of new social relations within social networks were shown to work with knowledge networks to predict emergence of new research fields as well [10].

As previous works utilize existing link structures, they are limited to predicting links within known sub-networks and unable to predict links connected to new nodes if they are introduced without any connecting links. This work presents a novel approach to predict how new nodes in expanding networks will be linked to existing nodes, which can aid in the prediction of emerging knowledge from the knowledge network.

III. Methods

The proposed methods utilize existing knowledge networks to predict where new nodes will be connected to the knowledge network in the future.

A. Data Used

The knowledge network is constructed from DBpedia for the following reasons. First, DBpedia is a large scale data repository which stores freely accessible and readily evaluated domain independent knowledge networks, ideal for undertaking preliminary experiments. Second, DBpedia, unlike other free knowledge networks available online, has versioned records as it is extracted regularly from Wikipedia.

Data set	Description
article-categories	Links between articles and categories
page-links	Internal links between Wikipedia articles
instance-types	DBpedia type information for articles
infobox	Infobox information for articles
redirect	Redirects between articles

Table I: DBpedia data sets used.

A constructed knowledge network consists of DBpedia categories as its nodes and DBpedia instances overlapping between categories as its links. DBpedia Data Set 3.9, 2014, and 2015 were used to create three timeslots. Table I shows the data sets used to fill the properties of given nodes and links. The data size of the latest data set, 2015, ranges from 6.8 million records for redirections, 20.2 million records for categories and types, 73.7 million records for infoboxes, to 162.2 million records for page-link.

B. pEgonet Identification

The paper introduces a new terminology called pEgonet by expanding from social network analytics research [7] a terminology *egonet*, which is defined as the collection of a given node n, its neighbors N_n, and links E_n among those nodes. pEgonet of new node n appeared at timeslot t is defined as a subset of its egonet at t, $\{n+N_{n,t}, E_{n,t}\}$, in the previous timeslot t-1 when n was nonexistent $\{N_{n,t-1}, E_{n,t-1}\}$. This represents a sub-network which will have a common new neighbor in the future timeslot.

DBpedia Data Set 3.9, 2014, and 2015 are denoted as t_1, t_2, and t_3 to generate three timeslots to be used in the preliminary research. There are two different groups of pEgonets that can be extracted from the data. The first group consists of all the pEgonets for new nodes first appearing in t_2, and the second group includes the pEgonets of new nodes first appearing in t_3. Figure 2 shows the pseudo-code for identifying all the pEgonets in t_2 generated from new nodes that appear in t_3.

C. Node-centric analysis

Node-centric analysis focuses on measuring the impact a given node will have in the future, which is defined as the number of future nodes that will stem from it. The definition of pEgonet dictates that a node which belongs to many pEgonets will be a neighbor of many new nodes in the next timeslot. Hence the paper aims to identify the characteristics of nodes that correlate to the number of pEgonets in which they are members.

Node properties shown in Table II are extracted as the first step of this analysis. There are four base predictor variables, and three average predictor variables which were derived from them. As the baseline function, linear functions were then analyzed to identify the function best fitted for the number of pEgonet membership, which is defined as *futureNode*. The correlation coefficients between dependent

```
Let $L$ = List of all of the new nodes at $t_3$
for all $n$ such that $n \in L$ do
    for all $a$ such that $a$ is a neighbor of $n$ at $t_3$ do
        for all $b$ such that $b$ is a neighbor of $n$ at $t_3$ do
            if $a$ exists at $t_2$ and $b$ exists at $t_2$ and edge ($a$, $b$) exists at $t_2$ then
                Add $a$, $b$, and edge ($a$, $b$) to the pEgonet of $n$
            end if
        end for
    end for
end for
```

Figure 2: Pseudo-code for extracting list of pEgonets in t_2 for new nodes in t_3.

variable *futureNode* and all the predictor variables were also calculated.

Future work will include more analysis with additional data, including link information. Some of the current predictor variables will be removed from the analysis to make sure that linear regression does not suffer from the multicollinearity problem. In addition, history data of Wikipedia articles will be used to see if modification patterns are related to the generation of new knowledge, as well as external link counts and types. Relationships between pEgonet membership likelihood and node properties will then be used to formulate a prediction method which can identify where and how many future nodes will be connected in the future.

D. Group-centric analysis

The identification of future neighbors from each node suffers from an overlapping problem as illustrated in Figure 1, as it can only predict if a given node will have new neighbors in the future. New analysis is needed to identify how new nodes will be connected to the current knowledge network. Group-centric analysis aims to identify the neighbors of a specific new node rather than the new neighbors of a specific current node in the future. This is the same as predicting the membership of pEgonet, given that pEgonet itself is the neighbors of a future node in a past timeslot. In this analysis, evaluations of pEgonet nodes were done to show that pEgonets have distinguishable characteristics.

Characteristics of pEgonets are compared against randomly selected sub-networks as there are no widely accepted baseline methods in link structure prediction for new nodes. Node size of each pEgonet in two groups mentioned in Section III-B were identified, and random sub-networks with the same size within the same timeslot were generated for every node size found. Links/nodes ratios of pEgonets and random sub-networks were then analyzed to identify the differentiating characteristics of pEgonet.

Future work includes comparing network communities against pEgonet node groups by the use of community detection algorithms, identifying disparities between their characteristics on which pEgonet identification method will be based. The result from linear regression for node properties will be incorporated as well. Human experiments will be done to evaluate the practicality of the proposed method, by comparing found results with NLP based prediction methods and actual knowledge expansion records in the given domain.

Variable	**Definition**
size	Number of Wikipedia instances a node has
type	Number of DBpedia types used within a node
redirect	Number of redirects used
infobox	Number of DBpedia infoboxes used
avgType	Average number of types per instance
avgRedirect	Average number of redirects per instance
avgInfobox	Average number of infoboxes per instance

Table II: Predictor variables in linear regression models.

IV. Preliminary Experiments

A. Node-centric analysis

The best matching functions for *futureNode* were found for each timeslot, with coefficients calculated from using actual pEgonet data.

For all nodes existing in t_1, the results of linear regression analysis suggest that: $futureNode = 2.57 + 0.01size - 0.07type + 0.07redirect + 0.00infobox + 2.51avgType - 0.48avgRedirect - 0.01avgInfobox$ with mean squared error of 591.64 and R-squared value of 0.61.

For all nodes existing in t_2, the results of linear regression analysis suggest that: $futureNode = 0.38 + 0.01size - 0.02type + 0.03redirect + 0.00infobox + 2.52avgType - 0.22avgRedirect - 0.01avgInfobox$ with mean squared error of 180.83 and R-squared of 0.73. Both analysis show that the average number of types per instances is the most important factor in predicting the number of future nodes.

Comparison between two results shows that *size*, *infobox*, and *avgType* have comparably consistent coefficients, which indicate that they are more suited for representing *futureNode* in multiple timeslots. Both *redirect* and *avgRedirect* fluctuate considerably per timeslot, showing redirection data is unstable in predicting *futureNode* in multiple timeslots.

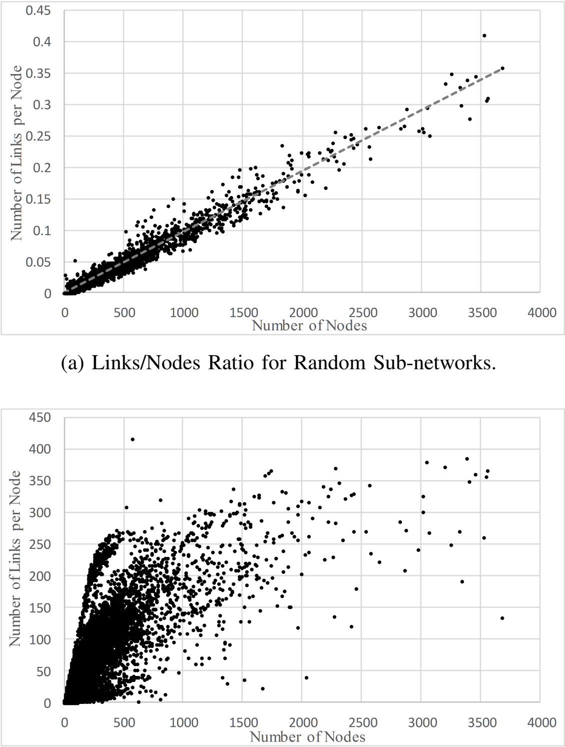
Correlation analysis showed that *futureNode* is highly correlated with the all base predictor variables with $size = 0.76$, $type = 0.75$, $redirect = 0.78$, and $infobox = 0.76$, while it has nearly zero correlations with averaged predictor variables with $avgType = 0.03$, $avgRedirect = 0.0006$, and $avgInfobox = 0.0008$. This suggests that *size* is the main predictor for *futureNode*, and larger nodes are more likely to be a neighbor of many future nodes. Also, the analysis revealed that there is a high correlation between the four base variables *size*, *type*, *redirect*, and *infobox*, with average correlation of 0.98 on average.

The correlation analysis result is inconsistent with the linear regression result. Negative coefficients of the variable *type* suggest negative correlation with *futureNode*, while the correlation analysis suggested otherwise. One possible explanation for this is that high correlation between four base variables caused the multicollinearity problem where coefficients change drastically based on the changes of other coefficients.

B. *Group-centric analysis*

The links/nodes ratio of pEgonet and randomly selected sub-networks were analyzed to identify differentiating characteristics of a sub-network which will have a new node as its common neighbor. Figure 3 shows the scatter plot of links/nodes ratio for random sub-networks and pEgonets, with 1% of outliers removed from graphs.

Figure 3a shows the links/nodes ratio of random sub-networks have a linear trendline with slope coefficient of $1/10265$. This suggests that sub-networks without future common neighbors can be identified by matching their links/nodes ratio against the trendline. Figure 3b shows that pEgonets show a scattered graph with no clear trend function, while having 155.5 times larger average links/nodes ratio. This suggests that given sub-networks can be assumed to have future common neighbors if their links/nodes ratio is comparably larger than that of random sub-networks, but the function with which sub-networks with common neighbors can be identified is less clear.



(a) Links/Nodes Ratio for Random Sub-networks.

(b) Links/Nodes Ratio for pEgonets

Figure 3: Scatter plot showing characteristics of pEgonet.

V. Conclusion

Preliminary results show that it is feasible to predict the location and structure of future nodes in the given knowledge network. Proposed methods can also be used to predict the impact of predicted nodes on the network, as well as providing comprehensive background information on which the future nodes will be labelled.

As this research is still a work in progress, the two proposed methods will be expanded as described earlier, and a combined method will be proposed to predict future knowledge concepts in the form of future nodes and their link structures in the knowledge network. This will include experiments with more timeslots, with domain specific networks such as PubChem and co-authoring networks.

The proposed method will need to address the labeling issue of the predicted nodes, or knowledge concepts, as they currently have no labels, or keywords, to represent themselves. The new method will utilize textual data from its neighbors to generate keywords candidates to deal with this problem. Also a weighting function will be added, measuring how *useful* the new knowledge will be in the knowledge network in the form of its effect on further new knowledge in the domain.

References

[1] J. Allan, *Topic detection and tracking: event-based information organization*. Springer Science & Business Media, 2012, vol. 12.

[2] Y. Yang, J. Zhang, J. Carbonell, and C. Jin, “Topic-conditioned novelty detection,” in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 688–693.

[3] A. Segev, C. Jung, and S. Jung, “Analysis of technology trends based on big data,” in *2013 IEEE International Congress on Big Data*, 2013, pp. 419–420.

[4] A. Segev, S. Jung, and S. Choi, “Analysis of technology trends basedon diverse data sources,” *Services Computing, IEEE Transactions on*, vol. 8, no. 6, pp. 903–915, 2015.

[5] S. K. Arora, A. L. Porter, J. Youtie, and P. Shapira, “Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs,” *Scientometrics*, vol. 95, no. 1, pp. 351–370, 2013.

[6] J. Leskovec, J. Kleinberg, and C. Faloutsos, “Graph evolution: Densification and shrinking diameters,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, pp. 3–43, 2007.

[7] L. Akoglu, M. McGlohon, and C. Faloutsos, “Oddball: Spotting anomalies in weighted graphs,” in *Advances in Knowledge Discovery and Data Mining*, 2010, pp. 410–421.

[8] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[9] M. E. Newman, “Clustering and preferential attachment in growing networks,” *Physical Review E*, vol. 64, no. 2, pp. 25–102, 2001.

[10] S. Jung and A. Segev, “Analyzing future communities in growing citation networks,” *Knowledge-Based Systems*, vol. 69, pp. 34–44, 2014.