CS 247 Project Proposal

Hui wang, Noor Nakhaei, Eric Dang, Wenhao Zhang

TOTAL POINTS

1 / 1

QUESTION 1

1 Proposal 1 / 1

✓ **- 0 pts Correct**

There exists a line of research called Knowledge Graph completion. Searching for some papers in that domain might be useful. The proposal is great! The overall structure could be improved by merging the first two sections for a longer Introduction. Also, in terms of evaluation, I am very interested in the details of the Prec, Recall, and F1. Please detail them in the final report and presentation.

Project proposal: Emerging Relation Detection from News in Heterogeneous Information Networks(HIN)

Eric Dang
erickdang@g.ucla.edu

Hui Wang
logicvay2010@g.ucla.edu

Noor Nakhaei
noornk@ucla.edu

Wenhao Zhang
wenhaoz@cs.ucla.edu

1 INTRODUCTION

The project topic we propose is **User profile prediction in heterogeneous information networks[1–5, 7–9]**. This proposal is organized as follows. Firstly, we briefly formulate the problem and explain our goal. Then we elaborate on how we intend to implement our project step by step. Then we introduce the datasets and metrics that we will use. We would also talk about the milestones and task assignments for this assignment. At last, we discuss some roadblocks/challenges to be expected in this project.

2 PROBLEM DEFINITION

A heterogeneous information network (HIN) is a network whose nodes and links may belong to different types. HINs are ubiquitous in our daily life and many real world networks can be modeled as HINs, e.g., bibliographic network, social network and knowledge base etc. Mining HIN has become a hot research topic which attracts a lot of attentions from researchers due to its capability of capturing meta structures with various rich semantic meanings wide applications in real-world scenarios including recommender systems, clustering, and outlier detections. HINs are becoming more and more popular in real world, however, directly mining such complex relationships is neither efficient nor effective. Network embedding, given its capability of preserving network structure and node proximity in networks, has drawn much attention from researchers. The new spaces obtained by network embedding models can be then fed to many existing machine learning algorithms to help improve performances in various tasks such as node classification, clustering, and link prediction etc. Therefore, the widely studied network embedding techniques recently have also been extended to help analyze HINs. HIN embedding targets the problem of exploiting various types of relationships among nodes and the network structures which are carried by meta-path, a sequence of node types and/or edge types[11].

Internet generates a sheer amount of information (e.g. news, tweets, blog, etc.) on a daily basis. These pieces of information introduce new entities and relations into our current knowledge graphs (e.g. Freebase, DBpedia) in real time. However, with the rapid growth of real-world knowledge, it is difficult to keep our KGs up to date. Therefore, some relations between new entities are absent in our KGs. These unlabeled relations are defined as *emerging relations* [10]. For example, when a celebrity give birth to a new baby, the current KGs have no record of the relation between the parents and their baby. This relation is one example of the *emerging relations*. However, this *emerging relation* can be learned from the news and tweets that describe this topic (Fig 1).

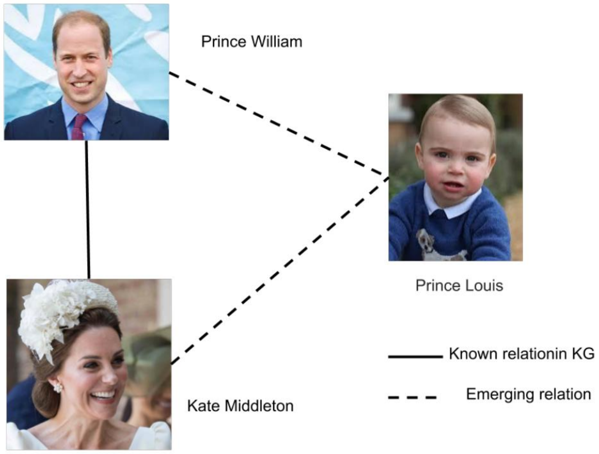


Figure 1: An example of emerging relation

In this project, we are interested in detecting these relations in real-time tweets and news. We first goal is to implement the *HEER* algorithm in [10]. Then we would like to test the effectiveness of our implementation on various datasets. **Please suggest if the workload is reasonable for ∼ 5 weeks**

Figure 2 explains the intuition of how HEER algorithm works in detecting emerging relations. We construct the heterogeneous textual graph from the news texts. In our example, the entities in **Heterogeneous Textual Gpraph (HTG)** are *name*, *baby*, *Charlotte*, *hopes*, and etc. You might realize that these entities are not referring to the same type of concepts. Some of them are names, and others are meta-data of an entity (e.g. writes, reveals). Hence the name of **Heterogeneous Textual Gpraph** [1–5, 7, 8, 10]. The relations between entities are encoded as edges in this HTG. The Knowledge graph serves as an "incomplete" dataset of facts of the world. Our goal is to update and complete our knowledge base by adding the "emerging" relations found in HTG. In Figure 2, the entity **Charlotte** was absent in our KG until we learned this entity and its relations with **Kate Middleton**, **George** and **William** in HTG. Then we added this new information to our KGs.

3 PROPOSED WORKFLOW AND MILESTONES

From the example of figure 2, we understand that the workflow of our project can be mainly divided into 3 steps:

(1) Construct a Heterogeneous Textual Graph from News
(2) Joint Embedding of the News and the KG

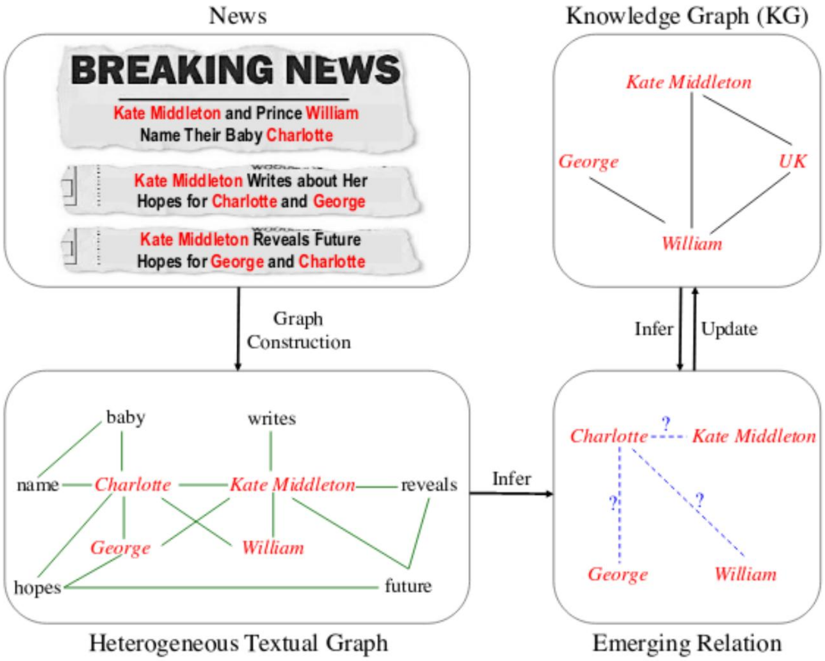


Figure 2: Detecting emerging relations by inferring from the heterogeneous textual graph and the KGs. The entities are in red. The co-occurence links in the heterogeneous textual graph are in green and the relations in KG are in black. Reprinted from [10]

(3) Implement HEER algo and detect Emerging Relations

This workflow is also summarized in Fig 3

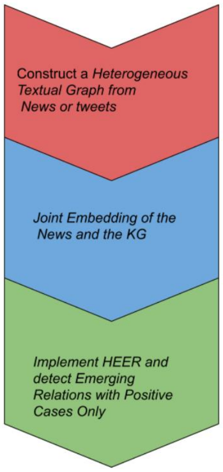

Construct a Heterogeneous Textual Graph from News or tweets
Joint Embedding of the News and the KG
Implement HEER and detect Emerging Relations with Positive Cases Only

Figure 3: Workflow and milestones

4 DATASETS DESCRIPTION

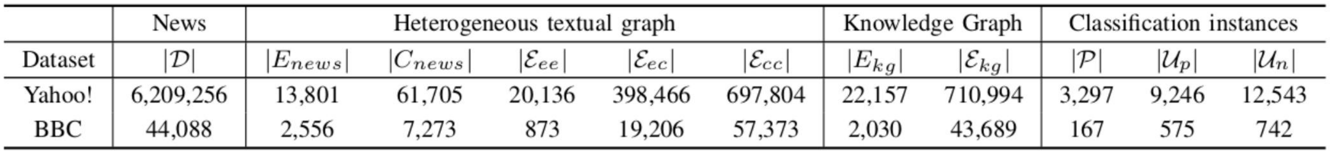
We will use four real-world datasets in this project.

- **Yahoo! News**: A collection of online English news from Yahoo! News in October 2015. Only the headline information is considered for Yahoo! News datasets. The link for this dataset is https://www.dropbox.com/s/yad2tfaj9ve3vuf/yahoo news titles.tar.gz? dl=0
- **BBC News** [6]: Documents in five topical areas are collected from the BBC news website from 2004 to 2005. We consider each sentence in the document as a piece of news. The link for this dataset is http://mlg.ucd.ie/files/datasets/bbc-fulltext.zip
- **Trump tweets**: A collection of all 30,385 Trump tweets from 5/4/2009 ∼ 1/27/2017. The link to the dataset is https://data.world/lovesdata/trump-tweets-5-4-09-12-5-16
- **DBpedia Ontology**: we plan to use DBpedia or a subset of it as our Knowledge Graphs. It contains the entities and the relations. The current release is 3.5. Link to dataset is https://wiki.dbpedia.org/data-set-35.

Table 4 summarizes the statistics of **Yahoo! News** and **BBC News** datasets.

One example in the *DBpedia Ontology* is as follows,

```
<owl:Class rdf:about="http://dbpedia.org/ontology...">
  <rdfs:label xml:lang="en">basketball player<...>
  <rdfs:subClassOf .../ontology/Athlete">...
```



	News	Heterogeneous textual graph					Knowledge Graph		Classification instances		
Dataset	$\lvert\mathcal{D}\rvert$	$\lvert E_{news}\rvert$	$\lvert C_{news}\rvert$	$\lvert\mathcal{E}_{ee}\rvert$	$\lvert\mathcal{E}_{ec}\rvert$	$\lvert\mathcal{E}_{cc}\rvert$	$\lvert E_{kg}\rvert$	$\lvert\mathcal{E}_{kg}\rvert$	$\lvert\mathcal{P}\rvert$	$\lvert\mathcal{U}_p\rvert$	$\lvert\mathcal{U}_n\rvert$
Yahoo!	6,209,256	13,801	61,705	20,136	398,466	697,804	22,157	710,994	3,297	9,246	12,543
BBC	44,088	2,556	7,273	873	19,206	57,373	2,030	43,689	167	575	742

Figure 4: Statistics of Yahoo! News and BBC News datasets

```
</owl:Class>
```

We can see that this entry has two entities, *basketball player* and *Athlete*. The relation (i.e. *Basketball player* is a **subclass** of *Athlete*) is also encoded here.

5 METRICS

We'll be using **AUC**, **Accuracy**, **F1 measure** to evaluate the performance [10]. We'll study the details of testing HEER algorithm later. The equations of caluclating these metrics are,

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1\ measure = \frac{precision \times recall}{precision + recall} \tag{3}$$

$$OverallAccuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

Where **TP** is number of true positives; **FP** is the number of false positives; **FN** is the number of false negatives; **TN** is the number of true negatives.

6 TASK ASSIGNMENT

- Data cleaning; Constructing a heterogeneous textual graph from news. (**Eric Dang**)
- Joint Embedding of the news and KG (**Hui Wang**)
- Implement HEER and detect Emerging Relations (**Noor Nakhaei; Wenhao Zhang**)

7 CONCERNS AND ROADBLOCKS TO BE EXPECTED

- Our major concern is whether the amount of workload we propose here can be finished by the end of the quarter. Our final presentation might be slightly different from the goal we propose here. We might simplify the project if we are running of time to complete our initial goal.
- We're aware that this project might be computationally intensive. We might rent computation if this becomes too challenging for our machines. Current plan is to use *Google Colab* if we need GPU acceleration.

REFERENCES

[1] Ye Liu, Luming Zhang, Liqiang Nie, Yan Yan, and David S Rosenblum. 2016. Fortune teller: predicting your career path. In *Thirtieth AAAI conference on artificial intelligence.*

[2] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and S Yu Philip. 2017. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering* 29, 1 (2017), 17–37.

[3] Yu Shi, Qi Zhu, Fang Guo, Chao Zhang, and Jiawei Han. 2018. Easing embedding learning by comprehensive transcription of heterogeneous information networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2190–2199.

[4] Yizhou Sun and Jiawei Han. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery* 3, 2 (2012), 1–159.

[5] Yizhou Sun, Xiang Ren, and Hongzhi Yin. [n. d.]. CONTEXT-RICH RECOMMENDATION: INTEGRATING LINKS, TEXT, AND SPATIO- TEMPORAL DIMENSIONS. ([n. d.]), 237.

[6] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[7] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. 2015. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* 58, 1 (2015), 1–38.

[8] Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu, Bradley Sturt, Urvashi Khandelwal, Brandon Norick, and Jiawei Han. 2014. Personalized entity recommendation: A heterogeneous information network approach. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 283–292.

[9] Jingyuan Zhang, Chun-Ta Lu, Mianwei Zhou, Sihong Xie, Yi Chang, and S Yu Philip. 2016. Heer: Heterogeneous graph embedding for emerging relation detection from news. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 803–812.

[10] Jingyuan Zhang, Chun-Ta Lu, Mianwei Zhou, Sihong Xie, Yi Chang, and S Yu Philip. 2016. Heer: Heterogeneous graph embedding for emerging relation detection from news. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 803–812.

[11] Sheng Zhou, Jiajun Bu, Xin Wang, Jiawei Chen, Binbin Hu, Defang Chen, and Can Wang. 2019. HAHE: Hierarchical Attentive Heterogeneous Information Network Embedding. 12.