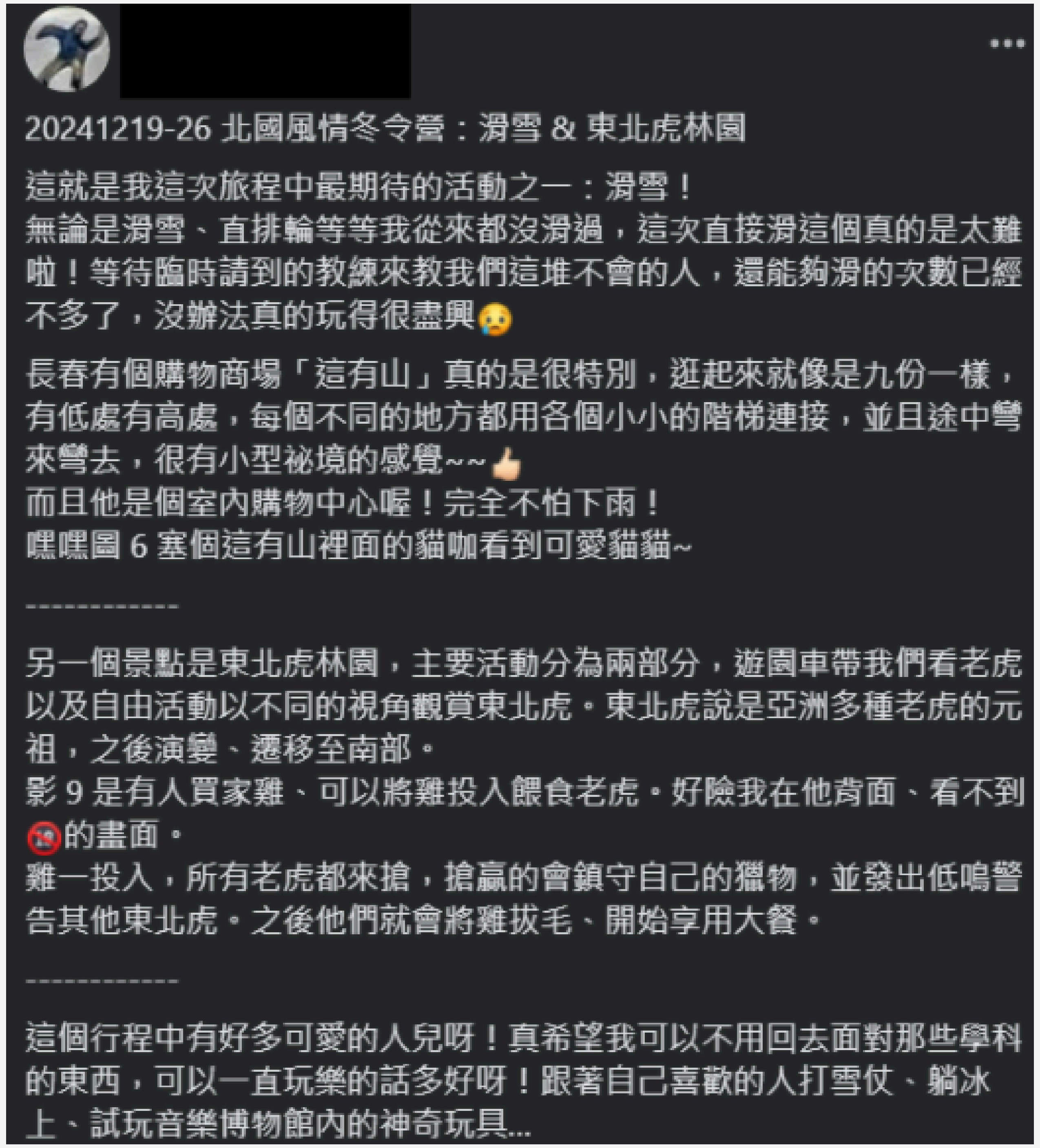
109 大學專題

由社群媒體發文
預測發文者輪廓

指導教授： 沈錳坤

專題學生： 周彥綸、馬茂元

專題發想背景


[redacted]

20241219-26 北國風情冬令營：滑雪 & 東北虎林園

這就是我這次旅程中最期待的活動之一：滑雪！
無論是滑雪、直排輪等等我從來都沒滑過，這次直接滑這個真的是太難啦！等待臨時請到的教練來教我們這堆不會的人，還能夠滑的次數已經不多了，沒辦法真的玩得很盡興😢

長春有個購物商場「這有山」真的是很特別，逛起來就像是九份一樣，有低處有高處，每個不同的地方都用各個小小的階梯連接，並且途中彎來彎去，很有小型祕境的感覺~~👍
而且他是個室內購物中心喔！完全不怕下雨！
嘿嘿圖 6 塞個這有山裡面的貓咖看到可愛貓貓~

另一個景點是東北虎林園，主要活動分為兩部分，遊園車帶我們看老虎以及自由活動以不同的視角觀賞東北虎。東北虎說是亞洲多種老虎的元祖，之後演變、遷移至南部。
影 9 是有人買家雞、可以將雞投入餵食老虎。好險我在他背面、看不到🔞的畫面。
雞一投入，所有老虎都來搶，搶贏的會鎮守自己的獵物，並發出低鳴警告其他東北虎。之後他們就會將雞拔毛、開始享用大餐。

這個行程中有好多可愛的人兒呀！真希望我可以不用回去面對那些學科的東西，可以一直玩樂的話多好呀！跟著自己喜歡的人打雪仗、躺冰上、試玩音樂博物館內的神奇玩具...

1. 現代人習慣使用社群軟體，並在上面發表可以表達自我的貼文

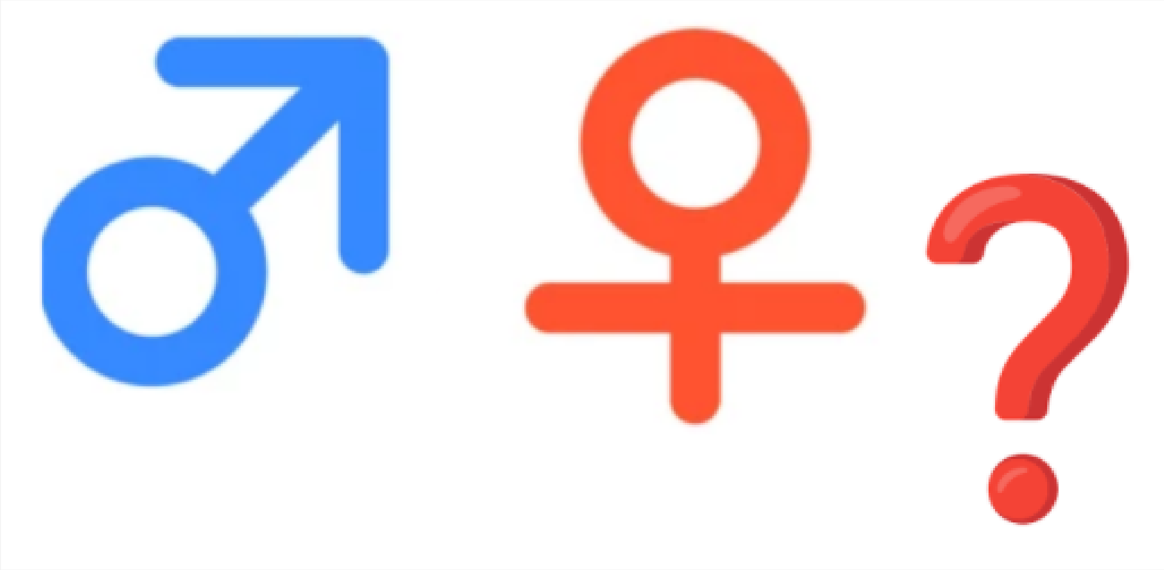
2. 現今各行各業為了更能迎合主要客群以及最大化廣告投放的效益，因此相當盛行分析用戶組成

專題動機

1. 某些社群軟體的使用者未必會完全公開其用戶資訊，如：性別、地區... ，我們希望透過深度學習來推敲這些使用者輪廓

2. 目前相關的 Data Mining 論文中，較少使用貼文關鍵字與 Knowledge Graph 的方法進行使用者輪廓(Demographic Data)的分析

專題目的



由使用者的貼文，預測該名使用者的
性別、所屬地區、年齡區間等等
使用者輪廓

實際測試 Knowledge Graph 預測
使用者輪廓的效果，並透過各種訓練技
巧，提高預測的準確率

相關研究

Explainable Reasoning over Knowledge Graphs for Recommendation by Xiang Wang et al

在本篇論文中，作者使用 Knowledge Graph 分別為使用者推薦歌曲與電影。 (接下來以歌曲說明)

Explainable Reasoning over Knowledge Graphs for Recommendation

Xiang Wang[1*], Dingxian Wang[2†], Canran Xu[2], Xiangnan He[1,3], Yixin Cao[1], Tat-Seng Chua[1]
[1]School of Computing, National University of Singapore, [2]eBay
[3]School of Information Science and Technology, University of Science and Technology of China
xiangwang1223@gmail.com, {diwang, canxu}@ebay.com, {xiangnanhe, caoyixin2011}@gmail.com, dcscts@nus.edu.sg

Abstract

Incorporating knowledge graph into recommender systems has attracted increasing attention in recent years. By exploring the interlinks within a knowledge graph, the connectivity between users and items can be discovered as paths, which provide rich and complementary information to user-item interactions. Such connectivity not only reveals the semantics of entities and relations, but also helps to comprehend a user's interest. However, existing efforts have not fully explored this connectivity to infer user preferences, especially in terms of modeling the sequential dependencies within and holistic semantics of a path.

In this paper, we contribute a new model named *Knowledge-aware Path Recurrent Network* (KPRN) to exploit knowledge graph for recommendation. KPRN can generate path representations by composing the semantics of both entities and relations. By leveraging the sequential dependencies within a path, we allow effective reasoning on paths to infer the underlying rationale of a user-item interaction. Furthermore, we design a new weighted pooling operation to discriminate the strengths of different paths in connecting a user with an item, endowing our model with a certain level of explainability. We conduct extensive experiments on two datasets about movie and music, demonstrating significant improvements over state-of-the-art solutions *Collaborative Knowledge Base Embedding* and *Neural Factorization Machine*.

Introduction

Prior efforts have shown the importance of incorporating auxiliary data into recommender systems, such as user profiles (Wang et al. 2018c) and item attributes (Bayer et al. 2017). Recently, knowledge graphs (KGs) have attracted increasing attention (Zhang et al. 2016; Shu et al. 2018; Wang et al. 2018a), due to its comprehensive auxiliary data: background knowledge of items and their relations amongst them. It usually organizes the facts of items in the form of triplets like (*Ed Sheeran*, *IsSingerOf*, *Shape of You*), which can be seamlessly integrated with user-item interactions (Chaudhari, Azaria, and Mitchell 2016; Cao et al. 2017). More important, by exploring the interlinks within a KG, the connectivity between users and items reflects their underlying relationships, which are complementary to user-item interaction data.

*The first three authors have equal contribution.

†Dingxian Wang is the corresponding author.

Figure 1: Illustration of KG-aware recommendation in the music domain. The dashed lines between entities are the corresponding relations, while the sold lines are the user-item interactions.

Extra user-item **connectivity** information derived from KG endows recommender systems the ability of **reasoning** and **explainability**. Taking music recommendation as an example (Figure 1), a user is connected to *I See Fire* since she likes *Shape of You* sung by the same singer *Ed Sheeran*. Such connectivity helps to **reason** about unseen user-item interactions (*i.e.*, a potential recommendation) by synthesizing information from paths.

Running Example: (*Alice, Interact, Shape of You*)∧(*Shape of You, SungBy, Ed Sheeran*)∧(*Ed Sheeran, IsSingerOf, I See Fire*)⇒(*Alice, Interact, I See Fire*).

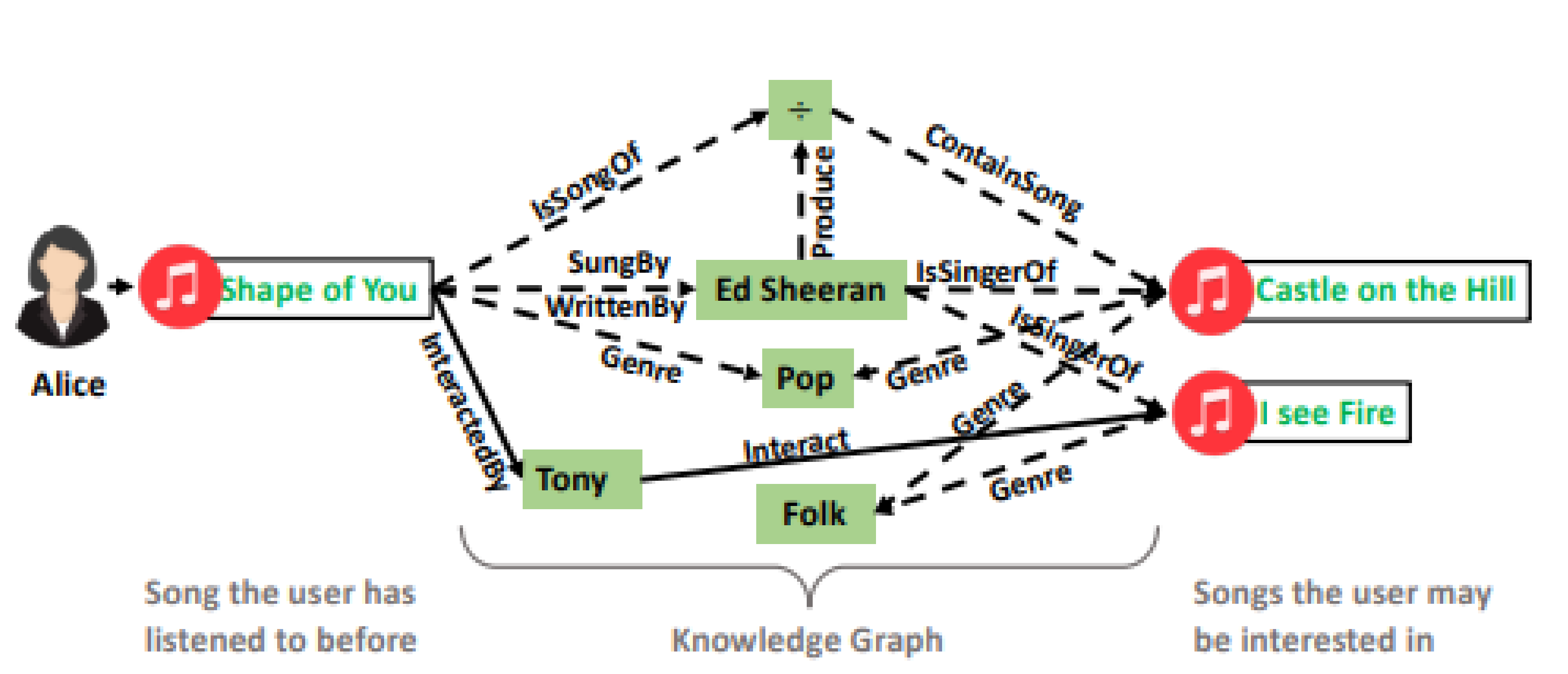
Clearly, the reasoning unveils the possible user intents behind an interaction, offering **explanations** behind a recommendation. How to model such connectivity in KGs, hence, is of critical importance to inject knowledge into a recommender systems.

Prior efforts on knowledge-aware recommendation are roughly categorized into path and embedding fashion. Path-based methods (Yu et al. 2014; 2013; Gao et al. 2018) introduce *meta-paths* to refine the similarities between users and items. However, we argue that meta-path is inefficient in reasoning over KGs, owing to the following limitations: 1) As relations are usually excluded from meta-paths, they hardly specify the holistic semantics of paths, especially when similar entities but different relations are involved in a meta-path; and 2) They fail to automatically uncover and reason on unseen connectivity patterns, since meta-paths

相關研究

Knowledge Graph 上面有許多的 Entity，包含多位使用者、多首歌曲以及一些專有名詞 (如：創作者、專輯名稱)，這些節點依照他們之間的關係互相連接，以此構成知識圖。

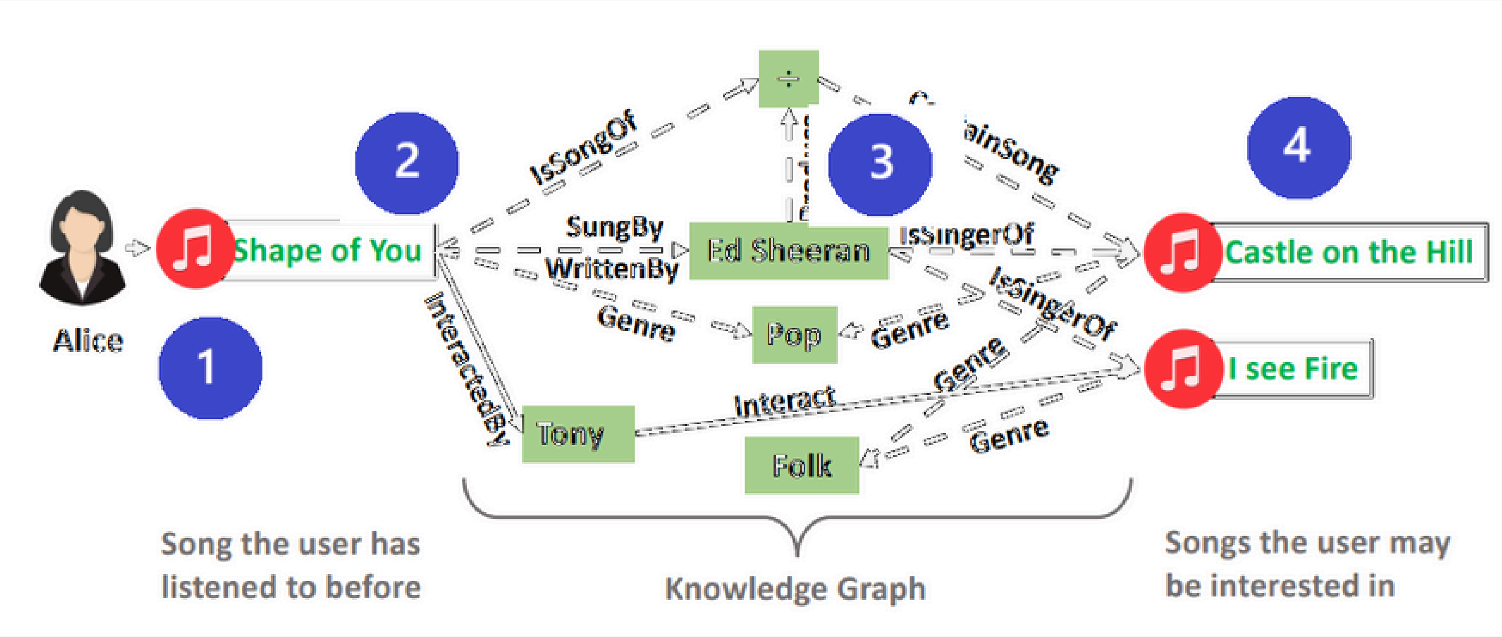
接著找出該位使用者連接到各首歌曲的路徑(Path)，經由 LSTM Model 計算這些路徑的推薦機率，並推薦機率較高的路徑末端的歌曲。

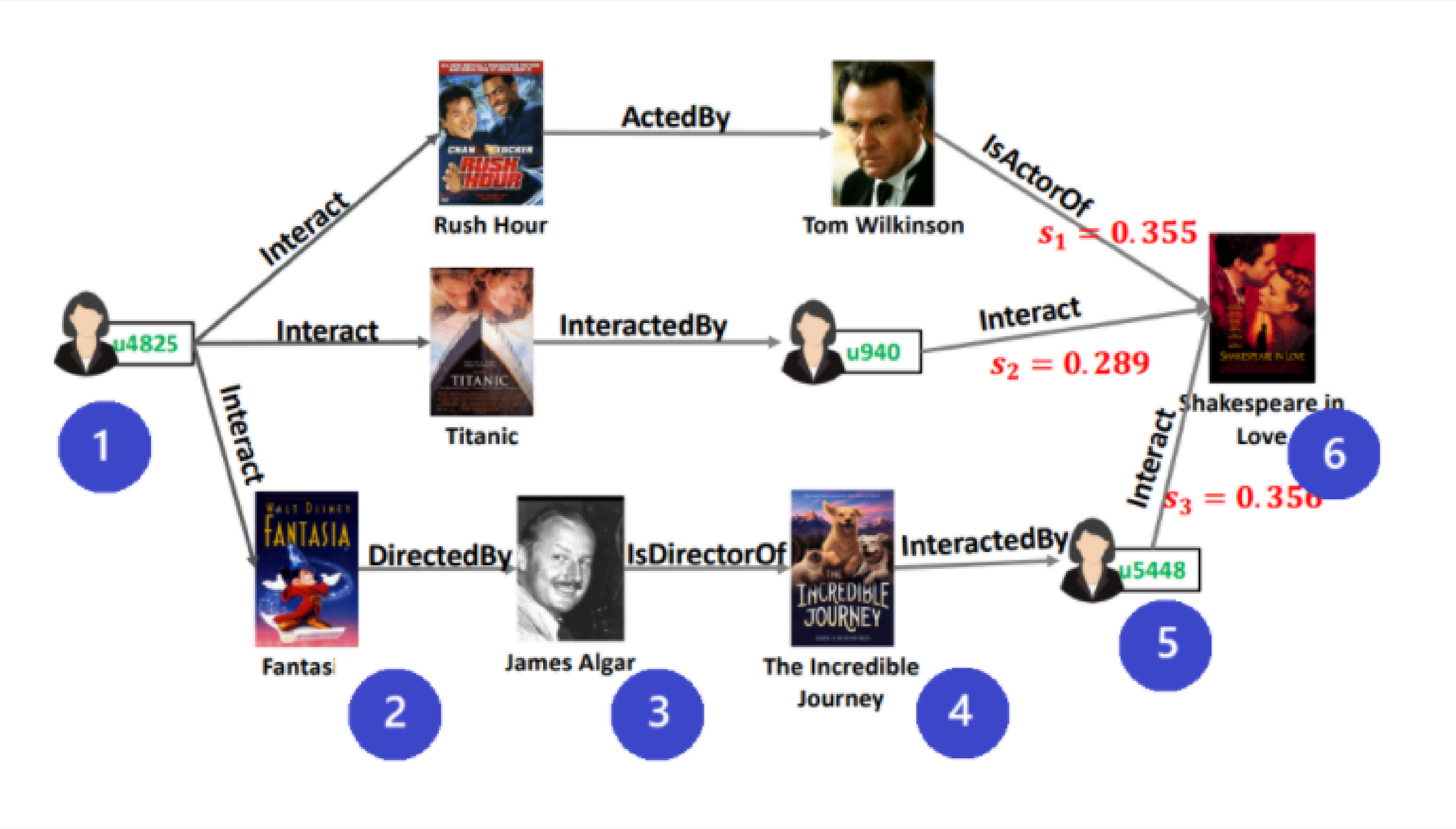


相關研究

經過該篇論文實驗， Path 的長度為 4 ~ 6 時會有較佳的效果。

因此我們的 Model 也是在這個範圍做 Grid Search



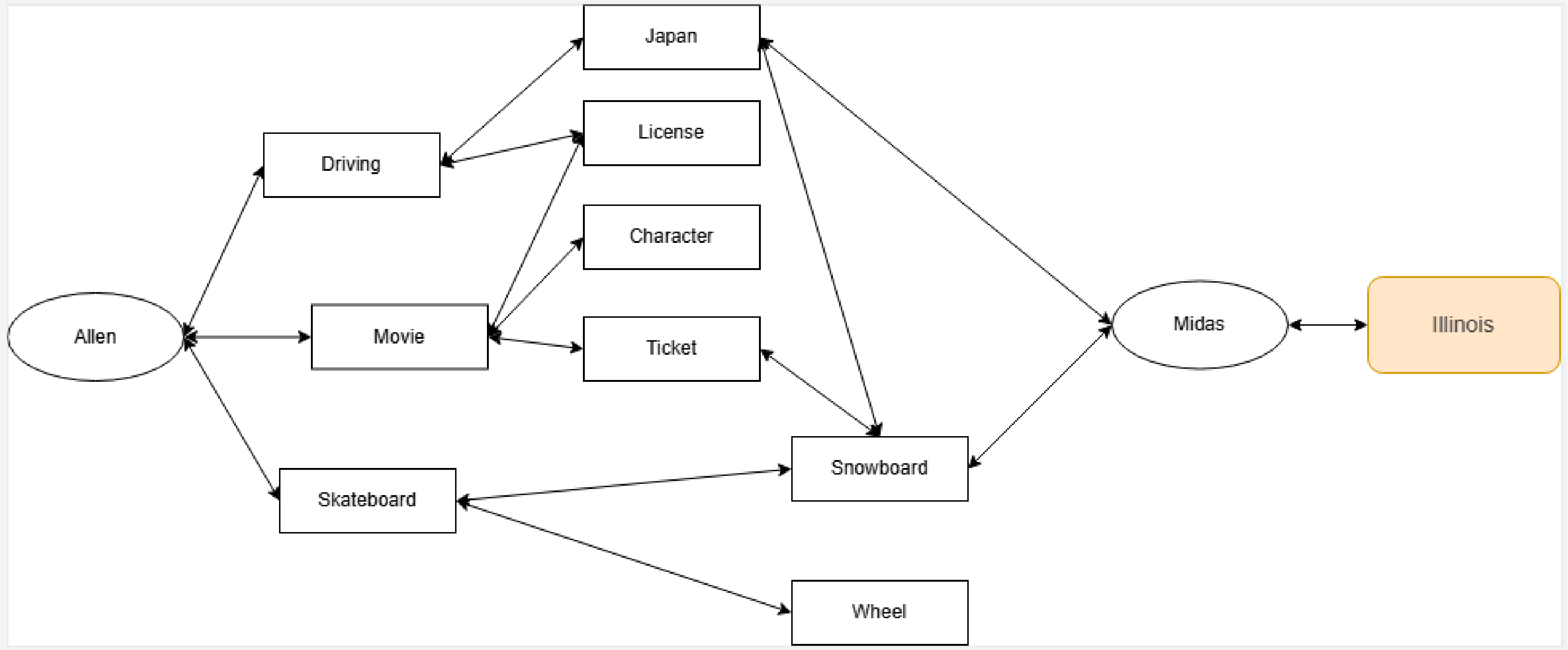


研究方法 -- Knowledge Graph

我們將從使用者的推文推算使用者輪廓，這邊以推測使用者所屬的美國州別說明。

首先，我們先將使用者的貼文做 TF-IDF 取得貼文中的關鍵字詞。

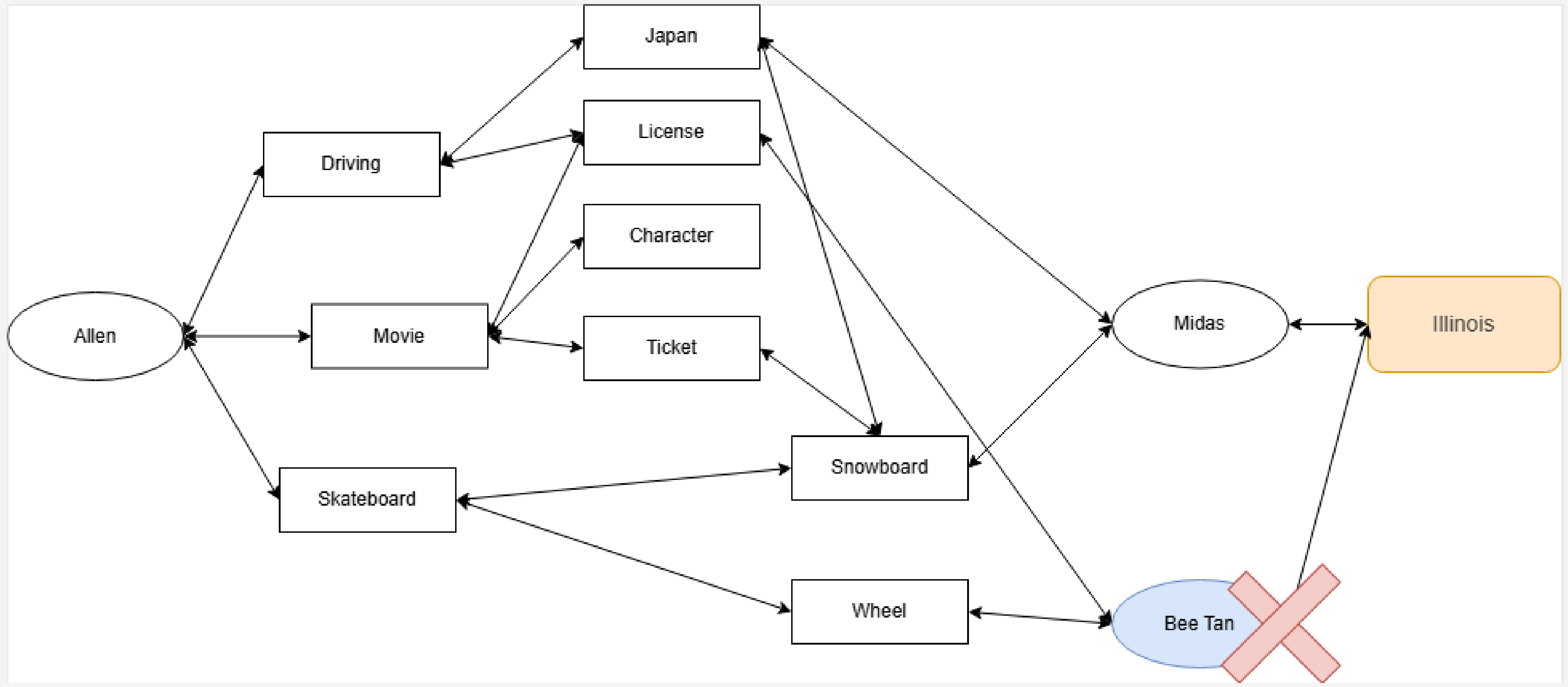
接著建立 Knowledge Graph 。 Entity 有使用者、字詞以及目標 (50個州別)，其中字詞與字詞之間的 Edge 我們使用 Bert 算出與某詞彙最相近的 K 個字彙並進行連接。 (K Nearest Neighbors)



研究方法 -- 找尋路徑(Path)

訓練時，我們會將 Testing Data 使用者與其目標從 Knowledge Graph 移除，接著以1:4的數量建立正樣本與負樣本。正樣本為使用者連結到正確所屬州的Path;而負樣本則連結到非所屬州的 Path 。

測試時則會再將 Testing Data 中的使用者與其目標放回 Knowledge Graph 中。我們會找出該位使用者連結到50州各自的機率，並從中判別其所屬州。



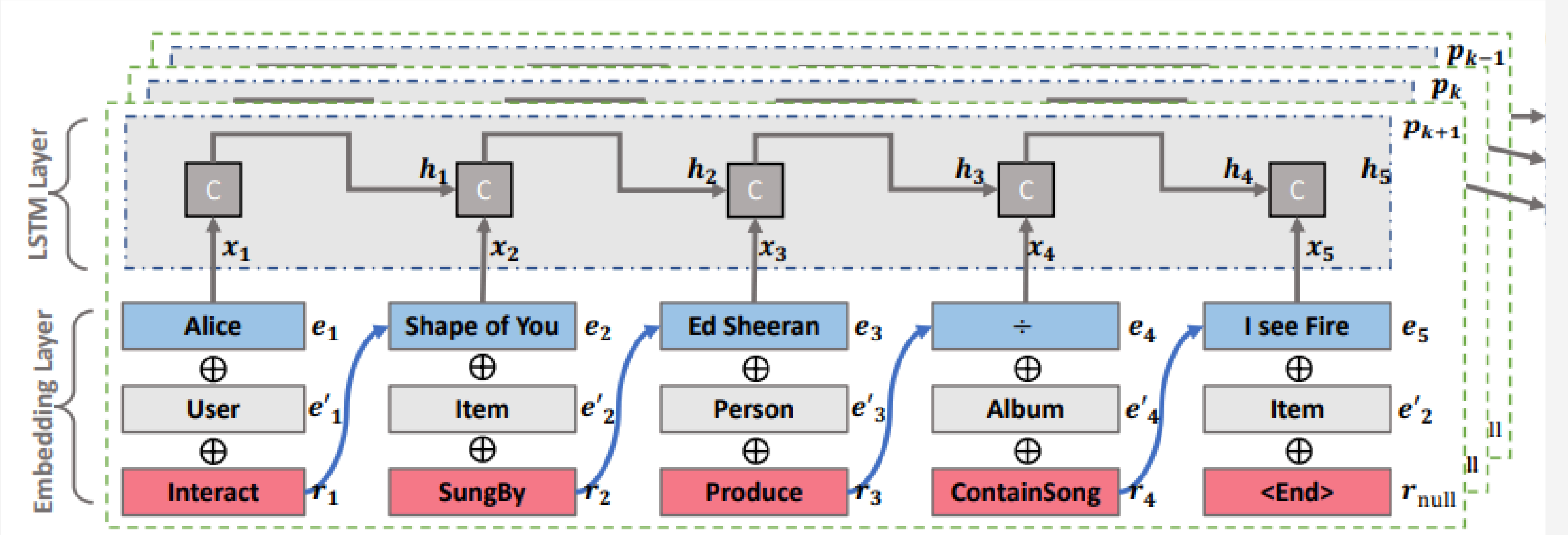
研究方法 -- 超參數與 Model

尋找 Path 時，可以決定超參數 M 決定接下來要前往幾個 Entity ，若是 M 越大則找到的 Path 數量也會增加。

另外 Bert 的 K Nearest Neighbors 也可以設定 K 的值，我們在 5, 7 與 10 之間做 Grid Search 。K 越大則知識圖連接地越密集，但可能也包含了不必要的資訊。

研究方法 -- 超參數與 Model

Model 我們採用參考文獻所使用的 LSTM。 Loss Function 選擇使用 Cross Entropy，預測州別時因為各州的使用者數量懸殊，因此我們有使用 Weighted Loss
Optimizer 則使用 Adam 進行模型權重更新。



實驗(A)

預測性別

- 數據組成：
  - 男性：1653 筆
    - 最多詞彙數量: 52062 個
    - 最少詞彙數量: 1 個
    - 平均詞彙數量: 645.97 個
  - 女性：1098 筆
    - 最多詞彙數量: 52062 個
    - 最少詞彙數量: 1 個
    - 平均詞彙數量: 254.49 個

實驗(A)
預測性別

- 準確率：

  K 值不論是選擇 5，7，10 ，抑或是 M 值或 Learning Rate 任意 Grid Search，Testing Data 的預測準確率皆為100%

- 觀察與結論：

  相較於推薦成千上萬的歌曲或是電影，或許將使用者分類為兩類性別的任務過於簡單，因此讓 Model 輕易完成任務

實驗(B)
預測地區

- 數據組成：
  - 共有2770筆地區資料

- 準確率：

K	5	7	10
Best Accuracy (%)	55 %	62 %	47 %

其中 K = 7 時 M = 60, Epoch 為 120, Learning Rate 為 0.0002, Batch Size 為 16, Regularization Term則為0.0001 ，而經過 Ensemble 5 個模型後，準確率上升至 71%