

# Data Mining Final Project

題目：Beyond Query: Interactive User Intention Understanding

組別：第六組

組員：

M10515010 黃健倫

M10515048 詹雨時

M10515028 曾也晏

M10515038 唐毅

M10515076 羅偉芯

Github: [https://github.com/z0012420014/DM\\_Final/tree/master](https://github.com/z0012420014/DM_Final/tree/master)

程式使用方式(Python)：

執行 GUI 的 code，按 GUI 上的 button 即可，也可以直接編譯 python 或是 java 的演算法執行，三個分別獨立皆可執行運作。

程式使用方式(JAVA)：

程式說明：

以 Java 語言實作，主要包含以下四個類別，

InteractiveHeuristicSearch\_Greedy 以 Greedy 方法產生所要問的問題

InteractiveHeuristicSearch\_MMAS 以 Heuristic 方法產生所要問的問題

InteractiveRecommendation 實作論文所提出的方法，用上兩個類別產生問題

RandomDataGenerator 對 Random.資料集進行論文中的實驗

Enter!										
uer/pob	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0.00	21	12	10	9	8	8	8	8	8	8
0.02	46	32	31	24	23	24	22	26	23	25
0.04	58	46	42	46	40	36	39	35	36	33
0.06	70	53	53	51	48	50	48	44	45	48
0.08	78	65	61	63	64	63	56	54	54	56
0.10	80	74	65	69	66	67	67	62	65	58
uer/pob 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50										
0.00	21	13	10	9	8	8	8	8	8	7
0.02	46	30	28	26	26	23	25	23	23	28
0.04	62	46	41	35	37	39	37	36	38	36
0.06	76	60	53	51	46	46	48	47	42	47
0.08	80	68	61	58	56	58	58	58	53	60
0.10	84	75	71	67	64	63	61	70	62	64
uer/pob 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50										
0.00	22	12	10	9	8	8	7	7	7	7
0.02	47	36	32	26	21	25	20	22	22	31
0.04	62	50	45	42	44	40	42	39	37	38
0.06	72	60	56	54	51	48	47	55	52	49
0.08	78	71	65	67	63	59	59	61	55	57
0.10	87	79	77	68	71	67	63	59	65	66
Done!										

RandomDataGenerator 類別的執行結果如上圖，黑色、紅色、藍色框所框住地的部分，分別代表 Greedy、IHS\_MMAS、SHS\_MMAS 的實驗結果，uer 是使用者回答錯誤率，從 0.0~0.1；pob 是在產生 item to tag 的 0/1 矩陣時出現 1 的機率，從 0.05~0.50，其餘數值表示找到目標物所需要的平均問題數量。

資料集說明：

我們從 RO 透視鏡(<http://gametsg.techbang.com/ro/>)這個網站去撈資料，撈出各個怪物會掉落的物品以及出沒的地點，把這兩個東西當成 feature 去猜怪物。

# 成員分工

黃健倫：

Paper presentation、PPT 製作、製作 GUI 與整合 python code

曾也晏：

Code review(演算法)、演算法撰寫(Python Greedy, SHS\_SA, IHS\_SA)、實驗結果

唐毅：

Code review(演算法)、演算法撰寫(Java Greedy, IHS\_MMAS, SHS\_MMAS)、實驗結果

詹雨時：

Code review(資料收集)、資料收集與整理

羅偉芯：

資料收集與整理

# 心得

黃健倫：

這次的期末專案我負責的部分是 GUI，這次我第一次使用 Python 撰寫 GUI，此外我還整合了專案中 Python 的兩種演算法的 code，如 Greedy 以及 IHS\_SA(Simulating Annealing)的 code 至 GUI 中並且作些許的修改。

此外，我還負責論文報告的部分，透過報告論文，我更清楚地瞭解論文的演算法用意，使得在整合 code 更加有效率。

詹雨時：

這次是我第一次使用 Python 去撈網頁上的資料，我覺得這個網站寫得不太好，許多 class 的命名都重複，因此我在撈資料上遇到了一些問題，不過最後也都順利的解決了，經由這次期末專案，使我對資料探勘這個領域有了更深入的認識。

曾也晏：

以 Python 實作論文中的三種演算法，包含 Greedy、IHS 與 SHS，其中，Heuristic search 採用 Simulated Annealing 並實驗。使用本學期資料分析課程學到的幾項技術完成，雖實驗結果稍有差距，但總體值域相同，也對 SA 有所認識。

唐毅：

做到後面發現論文有些地方沒有講清楚，例如  $\gamma$  參數的設定，所使用的 Heuristic 演算法還有向後考慮的問題數量，且因論文所提供的原始碼連結已經失效，使得我們只好自己去猜。

初步得到的結論  $\gamma$  基本等於 0，將  $\gamma$  從 0.1~0.9、0.01~0.05、0.001~0.005 進行測試， $\gamma$  只要不為 0 都跟論文實驗所得到的結果差異過大，平均所需要的問題數量會相差兩倍左右。這樣的設定其實還蠻詭異的， $\gamma$  如果為 0，使用者只要回答錯誤，目標物的權重即變為 0，這代表目標物永遠不可能被找到。

再來是 Heuristic 演算法，目前所做的雖然和論文的結果還差一截，但基本可與 Greedy 方法相當，少數時候比 Greedy 方法要好。以執行時間來說，論文中對 Random 資料集的執行時間 IHS 方法大概是 Greedy 方法的 4 倍，而所實作的 MMAS 演算法即使是使用很小的族群數量與迭代數，也會是 Greedy 方法的十倍有餘，因此推斷論文所使用的方法比較可能是 Simulated Annealing、Tabu Search 一類的單體搜尋式演算法，而非 ACO、PSO 一類的族群搜尋式演算法。

最後對於向後考慮的問題數量則沒有進一步的測試，目前是設為最小的 2。

羅偉芯：

這是我第一次修資料探勘相關課程，以前都只有大略聽過同學或者從一些文章中得知資料探勘的毛皮，在真正學習這門課前，覺得這些東西是有點遙不可及的，但在了解之後才會發現，其實在生活中料探勘的應用與我們息息相關。在實作 Project 的時候，第一次嘗試如何去把收集來的資料進行篩選還有整理，也了解到這些前置作業其實沒有想像中的複雜但也不是如此簡單，要去過濾掉甚麼需要的資料，或者最後要呈現什麼樣的格式，每個步驟其實都必須思考清楚並且多家嘗試，很高興能有此經驗。