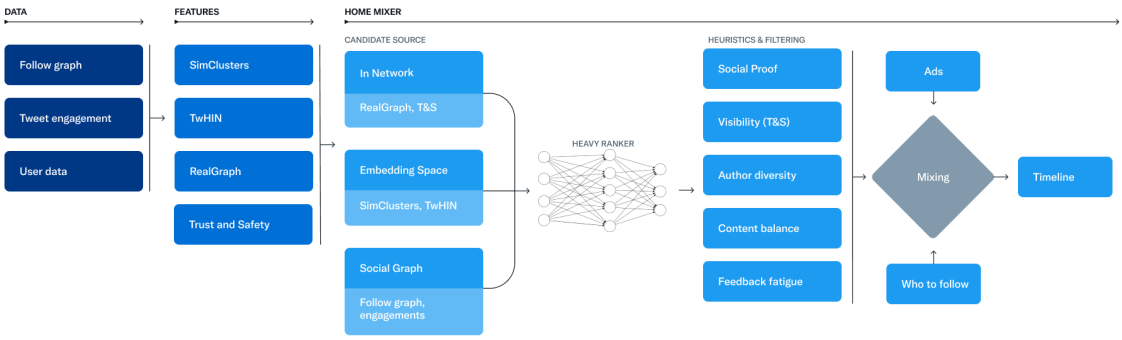
Twitter推荐算法

这个推荐流程包括三个主要阶段，这些阶段使用以下特征来进行处理：

1. **获取候选推文（Candidate Sourcing）**： 在这个阶段，系统会从不同的推荐源中获取最佳的推文。这些推荐源可以包括关注的用户、热门话题、趋势话题、推荐算法等等。主要任务是从多个来源中筛选出一组候选推文，以便在后续的阶段中对它们进行排名和进一步处理。
2. **推文排名（Tweet Ranking）**： 在这个阶段，系统会使用机器学习模型对每一条候选推文进行排名。排名的目的是根据用户的兴趣和行为，确定哪些推文最有可能吸引用户的注意力。这个机器学习模型会考虑一系列特征，这些特征可能包括用户的历史行为、推文的内容、互动数据等等。通过排名，系统能够为用户提供个性化的推荐内容，以增加用户的满意度和互动。
3. **应用启发式规则和过滤器**： 在最后一个阶段，系统会应用启发式规则和过滤器，以进一步精细调整推文的选择。这些规则和过滤器可以包括以下内容：
    - 过滤掉来自用户已屏蔽的账号的推文，以确保用户不会看到他们不感兴趣的内容。
    - 过滤掉不适宜内容，如色情、暴力等敏感内容，以维护平台的社区规范。
    - 检查用户是否已经看过某些推文，以避免在时间线上显示重复内容。



1. 推文候选源 Candidate Source

Twitter has several Candidate Sources that we use to retrieve recent and relevant Tweets for a user. For each request, we attempt to extract the best 1500 Tweets from a pool of hundreds of millions through these sources. We find candidates from people you follow (**In-Network**) and from people you don't follow (**Out-of-Network**). Today, the For You timeline consists of 50% In-Network Tweets and 50% Out-of-Network Tweets on average, though this may vary from user to user.

1. **候选源**：这是用于构建用户时间线的来源，用于提供可能的推文选择。候选源包括用户关注的人（In-Network）和用户不关注的人（Out-of-Network）。
2. **1500条推文**：每次请求尝试从数以亿计的推文中选择最佳的1500条。这是一个筛选过程，旨在确保用户看到最相关和有趣的内容。
3. **In-Network和Out-of-Network**：In-Network推文来自用户关注的人，Out-of-Network推文来自用户不关注的人。这种组合可以帮助确保用户不仅看到自己熟悉的内容，还能够发现来自更广泛社交网络的新内容。
4. **时间线构成**：用户的时间线通常由50%的In-Network推文和50%的Out-of-Network推文组成，平均分配。这种平均分配旨在提供平衡，让用户既能与已知关注者的互动，又能接触到新的、未知的内容。

2. 社交网络内部来源

网络内来源是最大的候选来源，旨在提供您关注的用户最相关的最新推文。它使用逻辑回归模型根据相关性对您关注的推文进行有效排名。然后，热门推文将被发送到下一阶段。

网络内推文排名中最重要的组成部分是真实图。真实图是一个预测两个用户之间参与可能性的模型。您和推文作者之间的真实图表得分越高，我们将包含的推文就越多。

3. 网络外来源

在用户的社交网络之外寻找相关的推文是一个复杂的问题，因为用户可能没有关注这些作者，因此我们需要采用不同的方法来判断某条推文是否与用户相关。Twitter采用了两种方法来解决这个问题：

社交图谱

Twitter的第一种方法是通过分析用户关注的人或具有相似兴趣的人的参与度来估计用户可能认为相关的内容。这个方法包括以下步骤：

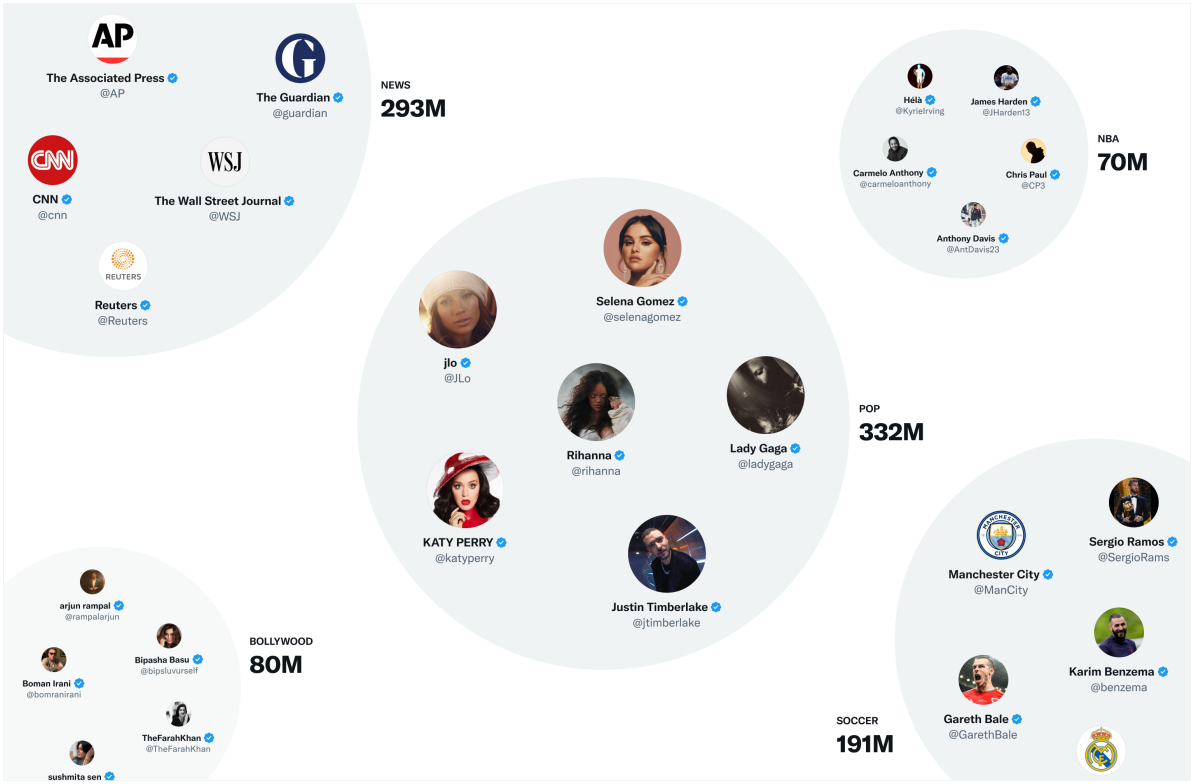
- 我关注的人最近与哪些推文互动？
- 那些喜欢与我相似的推文的人，最近还喜欢哪些内容？

根据这些问题的答案，Twitter生成候选推文，并使用逻辑回归模型对这些候选推文进行排名。这种图遍历方法对于网络外推荐至关重要，Twitter开发了一个图形处理引擎称为GraphJet，它实时维护用户和推文之间的互动图，以执行这些遍历操作。虽然这种启发式方法已被证明对搜索Twitter参与度和关注网络非常有用（目前占家庭时间线推文的约15%），但嵌入空间方法也成为网络外推文的重要来源。

嵌入空间：

嵌入空间方法旨在回答关于内容相似性的更普遍问题，即哪些推文和用户与我的兴趣相似。这个方法使用数字表示生成用户兴趣和推文内容，然后计算这些嵌入空间中的任意两个用户、推文或用户推文对之间的相似度。如果生成准确的嵌入空间，可以将这种相似性视为相关性的替代。

Twitter最有用的嵌入空间之一是SimClusters。SimClusters使用自定义矩阵分解算法发现由一群有影响力的用户锚定的社区。这些社区每三周更新一次，总共有145,000个社区。用户和推文在社区空间中表示，并且可以属于多个社区，这些社区的规模不同，从个人朋友组到数亿个新闻或流行文化的用户不等。Twitter可以通过查看推文在每个社区中的当前受欢迎程度来将推文嵌入到这些社区中，如果来自社区的用户喜欢推文的数量越多，推文与该社区的关联度就越高。



4. 启发式规则、过滤器和产品特性

在排名阶段之后，我们应用启发式规则和过滤器来实现各种产品特性。这些特性共同协作，创建了一个平衡和多样化的信息流。一些示例包括：

1. **可见性过滤**：基于推文内容和您的偏好来过滤推文。例如，删除来自您屏蔽或静音的帐户的推文。
2. **作者多样性**：避免来自同一作者的过多连续推文。
3. **内容平衡**：确保我们提供了平衡的社交网络内部和社交网络外部推文。
4. **基于反馈的疲劳**：如果观看者对某些推文提供了负面反馈，降低这些推文的分数。
5. **社交证明**：排除与推文没有二度关联的社交网络外部推文，以确保您关注的人与该推文作者进行了互动或关注了该推文的作者。
6. **对话**：通过将回复与原始推文串联在一起，为回复提供更多上下文。
7. **编辑推文**：确定设备上的推文是否过时，并发送指令以用编辑版本替换它们。

数据集解读

user

```
1::F::1::10::48067
2::M::56::16::70072
3::M::25::15::55117
4::M::45::7::02460
5::M::25::20::55455
6::F::50::9::55117
7::M::35::1::06810
8::M::25::12::11413
9::M::25::17::61614
10::F::35::1::95370
11::F::25::1::04093
12::M::25::12::32793
13::M::45::1::93304
```

UserID::Gender::Age::Occupation::Zip-code

- UserID范围是1~6040，代表了6040个MovieLens用户
- Gender 'M'代表男性，'F'代表女性
- Age 年龄的范围如下
    - 1： 18岁以下
    - 18： 18~24岁
    - 25： 25~34岁
    - 35： 35~44岁
    - 45： 45~49岁
    - 50： 50~55岁
    - 56： 56岁以上
- Occupation 的范围如下
    - 0：其他或者未指定
    - 1： 学者/教育行业
    - 2：艺术家
    - 3： 办事员/行政人员
    - 4： 大学生/研究生
    - 5：服务业
    - 6：医疗医护业
    - 7：执行官/管理者
    - 8：农民
    - 9：家庭主妇
    - 10：中小学生
    - 11：律师
    - 12：程序员
    - 13：退休人员
    - 14：销售人员/市场人员
    - 15：科学家
    - 16：自主创业
    - 17：技术人员/工程师
    - 18：商人/手工工作者
    - 19：失业
    - 20：作家
- Zip-dode 邮政编码

rating

```
1::1193::5::978300760
1::661::3::978302109
1::914::3::978301968
1::3408::4::978300275
1::2355::5::978824291
1::1197::3::978302268
1::1287::5::978302039
1::2804::5::978300719
1::594::4::978302268
1::919::4::978301368
1::595::5::978824268
1::938::4::978301752
1::2398::4::978302281
1::2918::4::978302124
1::1035::5::978301753
```

```
1::2791::4::978302188
1::2687::3::978824268
1::2018::4::978301777
1::3105::5::978301713
1::2797::4::978302039
1::2321::3::978302205
1::720::3::978300760
1::1270::5::978300055
1::527::5::978824195
1::2340::3::978300103
1::48::5::978824351
1::1097::4::978301953
1::1721::4::978300055
1::1545::4::978824139
1::745::3::978824268
1::2294::4::978824291
1::3186::4::978300019
1::1566::4::978824330
1::588::4::978824268
1::1907::4::978824330
1::783::4::978824291
1::1836::5::978300172
1::1022::5::978300055
1::2762::4::978302091
1::150::5::978301777
1::1::5::978824268
1::1961::5::978301590
1::1962::4::978301753
1::2692::4::978301570
1::260::4::978300760
1::1028::5::978301777
1::1029::5::978302205
1::1207::4::978300719
1::2028::5::978301619
1::531::4::978302149
1::3114::4::978302174
1::608::4::978301398
1::1246::4::978302091
2::1357::5::978298709
2::3068::4::978299000
2::1537::4::978299620
2::647::3::978299351
2::2194::4::978299297
2::648::4::978299913
2::2268::5::978299297
2::2628::3::978300051
2::1103::3::978298905
2::2916::3::978299809
2::3468::5::978298542
2::1210::4::978298151
2::1792::3::978299941
2::1687::3::978300174
2::1213::2::978298458
2::3578::5::978298958
2::2881::3::978300002
2::3030::4::978298434
2::1217::3::978298151
2::3105::4::978298673
```

```
2::434::2::978300174
2::2126::3::978300123
2::3107::2::978300002
2::3108::3::978299712
2::3035::4::978298625
2::1253::3::978299120
2::1610::5::978299809
2::292::3::978300123
2::2236::5::978299220
2::3071::4::978299120
2::902::2::978298905
2::368::4::978300002
2::1259::5::978298841
2::3147::5::978298652
```

其文件格式为:
UserID::MovieID::Rating::Timestamp

- UserID范围是1~6040，代表了6040个MovieLens用户
- MovieID范围是1到3952，代表了3952部电影
- Rating范围是1到5，代表了用户对电影的评级，最高5颗星，不允许半颗星存在
- Timestamp是以秒为单位的时间戳

注意： 每个用户至少会对20部电影进行评级。

Movie

```
3868::Naked Gun: From the Files of Police Squad!, The (1988)::Comedy
3869::Naked Gun 2 1/2: The Smell of Fear, The (1991)::Comedy
3870::Our Town (1940)::Drama
3871::Shane (1953)::Drama|Western
3872::Suddenly, Last Summer (1959)::Drama
3873::Cat Ballou (1965)::Comedy|Western
3874::Couch in New York, A (1996)::Comedy|Romance
3875::Devil Rides Out, The (1968)::Horror
3876::Jerry & Tom (1998)::Drama
3877::Supergirl (1984)::Action|Adventure|Fantasy
3878::X: The Unknown (1956)::Sci-Fi
3879::Art of War, The (2000)::Action
3880::Ballad of Ramblin' Jack, The (2000)::Documentary
3881::Bittersweet Motel (2000)::Documentary
3882::Bring It On (2000)::Comedy
3883::Catfish in Black Bean Sauce (2000)::Comedy|Drama
```

MovieID 范围是1到3952，代表了3952部电影

Title 是电影名称，由IMDB提供，包括了发行年份

Genres 电影题材由竖线分开，从以下类别中选取

- Action
- Adventure
- Animation
- Children's
- Comedy
- Crime
- Documentary

- Drama
- Fantasy
- Film-Noir
- Horror
- Musical
- Mystery
- Romance
- Sci-Fi
- Thriller
- War
- Western