信息检索系统的畅想

----基于目前网络搜索的个性化推荐设想

问题发现

基于我自己使用搜索引擎的感受来说，在有些时候搜索结果并不能让我很满意。例如当我想查找一些概念性知识，比较想要的是一些高质量博主（关注的，或者经常访问的博主）的详细解答，但是往往排在前面的并不是我最想要的，我需要多点一些链接，或者换一些搜索词来达到我的目的，还有就是在阅读英文文献时，有些东西看不懂，想要搜索一些中文的分析，但是当我输入这篇论文的英文题目去找的时候，大概率返回结果也都是英文的，也不会有中文意义的近义词出现，有时又想看一下某个方面外国一些人的解读会是怎样，但是输入中文大概率使中文内容的输出，感觉跨语言功能有些缺失。这些缺失使我的搜索体验感大大降低，如果可以实现一些基于信息提供方和跨语言的推荐的话，体验感应该会更好。

背景调研

经过查找资料，找到了目前的一些信息检索系统关于个性化推荐使用的方法，目前主流的推荐算法基本包括以下几种：

- **协同过滤系统**

协同过滤系统核心思想是：利用用户的历史信息计算用户之间的相似性——>利用与目标用户相似性较高的用户对其他产品的评价来预测目标用户对特定产品的喜好程度——>根据喜好程度来对目标用户进行推荐。在计算用户之间相似度时，大部分都是基于用户对共同喜好产品的打分。最常用的方法是Pearson 相关性和夹角余弦。协同过滤推荐系统的算法可以分为两类，基于记忆(memory-based)&基于模型(model-based)，前者是根据系统中所有 被打过分的产品信息进行预测，注重于预测用户的相对偏好面不是评分绝对值:后者是收集打分数据进行学习并推断用户行为模型，再对某个产品进行预测打分。

- **基于内容的推荐系统**

基于内容的推荐系统是协同过滤技术的延续与发展。其核心思想:分别对用户和产品建立配置文件——>比较用户与产品配置文件的相似度——>推荐与其配置文件最相似的产品。例如，在电影推荐中， 基于内容的系统首先分析用户已经看过的打分较高的电影的共性(演员、 导演、风格等)， 再推荐与这些用户感兴趣的电影内容相似度很高的其他电影。基于内容的推荐算法根本在于信息获取和信息过滤。因为在文本信息获取与过滤方面的研究较为成熟，现有很多基于内容的推荐系统都是通过分析产品的文本信息进行推荐。在信息获取中，最常用的是TF-IDF方法。

- **基于网络结构的推荐算法**

基于网络结构的推荐算法仅仅啊用户和产品的内容特征看成抽象的节点，所有算法利用的信息都藏在用户和产品的选择关系中。其核心思想是：建立用户---产品二部图关联网络。对于任意目标用户i，假设i选择过所有的产品，每种产品都具有向i推荐其他产品的能力，把所有i没有选择过的产品按照他喜欢的程度进行排序，把排名靠前的推荐给i。在同样的用户爱好程度下，推荐冷门的产品要比推荐热门的产品意义更大。此算法开辟了推荐算法研究的新方向。

- **混合推荐**

将上述几种推荐方法有机结合，实际的推荐系统中最常见的是基于协同过滤和基于内容的。在协同过滤系统中加入基于内容的算法，利用用户的配置文件进行传统的协同过滤计算，用户的相似度通过基于内容的配置文件计算得出, 而非共同打过分的产品的信息。这样可以克服协同过滤系统中的稀疏性问题，另外 不仅仅是当产品被配置文件相似的用户打了分才能被推荐，如果产品与用户的配置文件很相似也会被直接推荐。

经过搜索发现，其实现有的推荐算法更多是基于内容的，我认为应该也可以加入信息来源方进去考虑，例如用户经常访问的网址，用户经常访问的用户等，提高信息来源的特征的比重，可能会得到用户更想要的信息，还可以加入一些用户更多的行为，例如根据人体记忆曲线等设计一些动态信息检索推荐算法，例如加入鼠标光标行为进行预测。

功能设计

补充信息来源和用户行为

根据基于内容的推荐算法的灵感来源，为实现基于信息来源的个性化推荐，首先在采集用户信息时，就需要采集用户的一些关注者，以及一些账户信息，但是可能一些网站涉及隐私问题，会阻止获取用户的个人信息，信息检索引擎可以根据一些日志记录，记录用户常访问的网址，根据网址信息来进行推荐，将网址的信息也加入推荐算法中，并且基于用户的访问率，可以适当提高某些信息来源的占比。对于一些能拿到用户账户信息的网站，检索系统在做搜索的过程中，加入用户感兴趣的作者，有限搜索该作者下的内容，是否有与搜索内容相匹配的结果，这样用户或许能得到比较中意的结果。

第二次作业中读过的内容是关于用户鼠标光标行为与检测的预测的，或许在个性化推荐中，我们也可以加入这个来源，以更好的揣测用户心理。鼠标光标可以模拟人眼的行为，根据鼠标的移动序列，在一定程度上可以预测用户的感兴趣的东西，我们可以加入鼠标进行计算，对于用户感兴趣的东西进行个性化推荐，如果推荐的内容用户进行了点击，我们可以做些标记，用户确实对这方面感兴趣，反之，不会增加类似推荐，不过加入这个因素，需要控制好度，因为鼠标的不稳定性因素有点大，干扰信息较多，或许会降低准确度。

这个方面的涉及其实跟之前的算法基本相同，或许对基于内容的推荐算法加一些修改，这些想法就可以实现，感觉这些的想法与基于内容的推荐算法高度重合。

加入遗忘因素

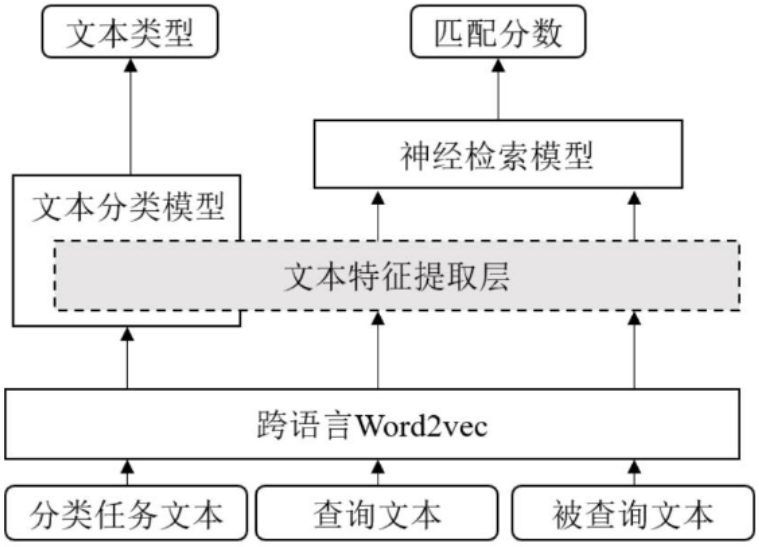
德国心理学家艾宾浩斯研究发现了遗忘曲线，揭示了人类大脑对新事物记忆随着时间逐渐降低的过程。推荐系统处于一个动态的环境中，用户的兴趣和物品种类都在不停变化。传统的推荐算法只考虑用户的相似性或者项目间的相似性，忽略用户兴趣的动态变化，导致推荐精度会随时间推移而下降。关于推荐模型很有必要根据人的思维模型的进行相应调整。关于这块模型的设计，在收集到的用户信息加入推荐算法运算时，或许可以引入神经网络，长短时记忆LSTM或许可以加入运算，LSTM是RNN的一种，不过LSTM可以有将之前的信息加入模型计算。在进行信息使用时，我们应当减小较早的用户兴趣占比，这样我们就可以通过LSTM中的遗忘门来达到这个效果。

在个性化推荐的设计中，更多的考虑人的思考特点，可以提交推荐的一些精准性，这里只提出了遗忘因素，或许可以有更多的可以加进来。对于遗忘因素这一方面，已经有人进行了初步实现，他们提出了TIME-GAN模型：利用遗忘时间函数模型对原始用户评分数据进行加权处理，缓解因时间产生用户兴趣漂移而带来的影响；基于深度因子分解机DeepFM模型形成Time-GAN的生成模型，采用Pearson相似度计算生成推荐列表与用户真实列表之间相似性。通过对抗网络框架完成推荐列表对抗训练，利用MovieLens-100K数据集对模型进行了实验，实验结果优于传统推荐系统。

实现跨语言检索推荐

随着互联网的发展与全球化进程的推进，信息的数量飞速增加，用户在非母语条件下进行检索的需求也逐渐提高，但是目前搜索引擎中跨语言这个搜索功能我并没有太体会到，我如果搜索英文很少会有基于翻译后的中文相关搜索结果出现，目前关于此方面有一个搜索技术叫跨语言信息检索（cross-language information retrieval, CLIR）技术，用户可以使用母语直接检索多种其他语言的信息，或许在个性化推荐系统中，我们也可以加入这部分来做。在用户进行搜索信息过程中，搜索引擎不只提供与母语相关内容，可以提供其他语言的一些相关推荐，也可以根据用户的一些语言使用习惯，设置一些推荐语言比例，这样用户能得到更加多样化的搜索结果。为了提高结果的精度，有研究人员使用神经网络等加入跨语言嵌入的研究。

跨语言信息检索的任务流程通常分为3步：统一查询和文档的语种、提取文本特征、执行检索。关于这个部分，对于文本的研究更多些，或许需要更多的自然语言处理过程，有一些学者提出多任务的跨语言信息检索方法，这块跟神经网络联系比较紧密，一些学者提出如下模型。



在个性化推荐的设计过程中，对于用户信息的处理，我们应该也可一借鉴此模型，将用户的一些常访问的文件进行跨语言嵌入操作，获取其他语言的一些相关推荐，得到更丰富的搜索结果。

总结

信息检索的个性化推荐其实技术已经比较成熟了，之前背景调研中提到的那些推荐算法也已经成了体系，但是仍有些时候推荐结果总是不太让人满意，所以个性化推荐是需要根据信息检索领域的一些技术的改进进行升级的，同时个性化推荐领域又有一个比较矛盾的点就是，为了提高推荐内容的针对性和准确性，我们需要多采集用户的一些信息，但是这又会涉及到一些隐私问题，所以在保护隐私的前提下提高个性化推荐的精度和满意度是一个比较难的问题，而本文中所提到的一些新加入个性化推荐的功能，或许可以有利于用户体验感的提升，目前还处于概念阶段，具体的效果只有实现了才能知道，理论阶段就是以上的内容。