基于关键词权重的文献推荐算法及模型研究

**汪晓锋1 罗逸杰1 容天泽1（1 上海交通大学）**

摘要：随着互联网技术的发展，各学科邻域的不断推进，学术科研资源呈现出爆炸性增长，信息过载问题越来越凸显。如何快速从众多各类文献资料中找出最适合自己的部分成为了急需解决的难题。而基于用户数据的个性化推荐系统可以提供一个很好的解决方案。

传统的基于内容的推荐算法在内容特征提取上都很困难，而大部分的文献都已经附带关键词，论文提出一种基于关键词权重的推荐算法，利用基于机器学习的关键词提取模型完善文献的关键词，然后直接利用关键词作为文献的内容信息，对用户的浏览和下载纪录进行建模，由关键词出现频次和时间顺序以及数据库中关键词的相关信息得到用户关键词权重，构建关键词向量，推断出用户的兴趣偏好，以此对相关文献进行打分，找到关键词信息最为匹配的文献推荐给用户。通过实验发现，基于关键词权重的推荐系统在文献推荐上效果显著。

关键词：推荐系统 关键词提取 相似度测量 关键词权重

**Research on the Modeling and Related Algorithms of Keyword-Weight Rating Based Literature Recommendation System**

Abstract: With the development of Internet technology and the continuous advancement of various disciplines, academic and scientific research resources show an explosive growth, and the problem of information overload is becoming more and more prominent. How to quickly find the most suitable one from numerous documents has become an urgent problem to be solved. In modern time, the personalized recommendation system based on user data appears to be a good solution.

The traditional content-based recommendation algorithms are difficult to extract content features. Since most of the documents have their keywords attached, our paper proposes a recommendation algorithm based on keyword weight, which directly uses keywords as the content information to model users' browsing and downloading records. The user's keyword weight is obtained from the keyword's frequency of occurrence and its time sequence, based on which the keyword vector is constructed to infer the users' preference. Meanwhile, the algorithm will score the relevant documents according to the similarity between the keywords and recommends the users with the one that have the best-matched keywords. The recommendation algorithm based on keyword weight effect well through experiments.

Keywords: Recommendation System, keyword extraction, Similarity Measure, Keyword-weight Rating

**0 引言**

近年来网络技术飞速发展，各类数据资源在网络上都呈现爆炸式增长，随着网络的普及，用户数量也越来越多，对各类数据的需求也越来越多。从前我们上街买衣服只能选择一条街上为数不多的服装店，如今淘宝和京东商城的商品数量却能以亿的单位来计算，在任意一个平台搜索自己想要的信息，上千条链接一字排开。人们在庞大的信息量中很容易被无关因素干扰，陷入到信息过载的漩涡之中，无法及时获取信息。而万方数据知识服务平台包含了上亿篇文献，推荐系统能够帮助我们解决部分难题，为用户提供一个快速获取所需资料的方法。[1]

而当前的主流推荐算法主要有基于内容[2-3]和协同过滤[4-5]的算法。基于内容的算法分析物品的特征，然后根据用户历史纪录中的物品特征，推荐相似的物品。而协同过滤的推荐算法又分为基于用户和基于物品两种方式：前者通过分析用户对物品的行为的相似性，对指定用户进行物品推荐。后者通过相同用户对不同物品的行为找出相似的物品，然后以此为指定用户进行推荐。

这两种算法目前的运用最为广泛，然而依然存在很多缺陷。基于内容的推荐算法难以扩展导致推荐内容局限，且在很多情况下特征难以抽取，协同过滤的推荐算法在存在有冷启动问题，而且在给用户推荐物品时无法给出令人信服的理由，可解释性差，很难满足特定用户的需求。

除此之外还有基于标签推荐算法[6-7]也运用广泛，通过分析用户的标签喜好以及物品的标签特征，利用标签特征的相似性对用户进行推荐。这种算法依然具有冷启动问题，同样也依赖于用户的历史行为数据。

文献作为一种特殊的物品，很多文献本身就携带了关键词，标签信息获取比较方便，本文在基于标签的推荐算法上，加入关键词权重以充分利用已知信息，以提高推荐的精度，提出一种基于关键词权重的推荐算法,主要有如下两个创新点：

(1)基于深度学习的关键词提取模型的构建，本文尝试了多种方法，对关键词的提取基本能做到准确且有一定的概括性，对完善文献关键词信息有很大作用。

(2)关键词权重的引入，构建了一个同时考虑用户和文献的权重计算办法，且在当中加入时间权重，大大提高了模型的精度。

**1 相关工作**

1.1 数据集分析

本文选择了万方数据知识服务平台期刊文献用户行为日志+这一数据库，数据库中包含了2020年全年万方数据知识服务平台中用户的浏览纪录、下载纪录和检索纪录。其中浏览纪录和下载纪录中包含的每条数据信息有用户ID、用户类型、文献ID、文献标题、浏览时间、省份、文章关键词和作者，经分析本文发现每条信息均包含用户ID、文献ID和文献标题，而有一部分信息缺乏文章关键词等其他内容。本文想要借助数据库中的这些信息搭建一个个性化推荐系统。

1.2 传统推荐算法

1.2.1 基于内容的推荐算法[2-3]

基于内容+的推荐算法的流程是为首先分析一个用户的使用纪录中的物品的共同特点，然后根据这些共同特点去数据库中找到具有同样特点的物品推荐给用户。比如一个用户看过许多推荐算法方面的文献，那么该系统就会判定推荐算法是该用户所喜好的，于是便大量推荐有关推荐算法的文章。目前来看，比较典型的基于内容的推荐系统有 Uneset 新闻基于内容的推荐算法推荐系统和 Fab 网页推荐系统。这样的推荐算法简洁直接而且可解释性强，不存在冷启动问题，然而局限性也很明显：

(1)大部分物品的特征难以提取。

(2)没有具体的评分系统，推荐的精度较低。

(3)多样性差，在推荐内容上难以有更多的拓展。

考虑到文献的特征同样难以提取，本文没有考虑基于内容的推荐算法。

1.2.2 协同过滤+推荐算法[4-5]

作为目前使用最为广泛的推荐算法之一，协同过滤的推荐算法被广泛应用于各种平台。引言中以及提到协同过滤推荐算法主要分为两种：基于用户的协同过滤推荐算法和基于物品的协同过滤推荐算法。前者为通过寻找相似的用户再为其推荐相似用户使用过的物品。比较典型的有研究发现很多中年男性会同时购买纸尿片和啤酒，于是很多商场便会将这两个看起来毫不相干的物品摆放在一起，为购买纸尿片的男性同时推荐了啤酒，从而大大提高了销量。而基于物品的推荐算法同理，通过寻找相似的物品然后向相似物品的用户推荐目标物品。这样的推荐算法拓展性很强，可以将用户的目光发散出去，推荐用户的潜在偏好，然而协同过滤推荐算法也存在以下缺点：

1. 缺乏可解释性。
2. 冷启动问题无法解决，过度依赖历史数据。

(3)对稀疏数据缺乏解决方案。

由于数据的稀疏性以及考虑到文献这类物品的特殊特性对发散性的需求不高，协同过滤的推荐算法并不适用本文研究场景。

1.2.3 基于标签+的推荐算法[6-7]

基于标签的推荐算法通常为统计一个用户最长用的标签和每个物品最长被打上的标签然后通过一定的关系将二者配对从而进行推荐。比较经典的是豆瓣网的推荐系统，通过大量使用标签信息来进行个性化推荐。其本质为通过引入标签形成一个用户-标签-物品的三元关系。这样的推荐算法与基于内容的推荐算法原理上相似，但算法更加简便。然而这样的推荐算法对标签的分析具有局限，算法精度方面无法保证。

在综合考虑数据库的特征之后，本文将关键词作为文献的标签，在基于标签的推荐算法的基础上提出基于关键词权重的推荐算法。

**2 基于关键词权重的推荐算法**

2.1 推荐系统

2.1.1设计思想与框架

同于基于标签的推荐算法，我们意图构建一个用户-关键词-文献的三元关系，如图所示：

用户关键词向量

对应关系

Top-N推荐

图2.1.1.1 三元关系

每两层之间的关系用算法来实现，构建出一个基于关键词权重的推荐系统，系统体系架构如图2.1.1.2所示：

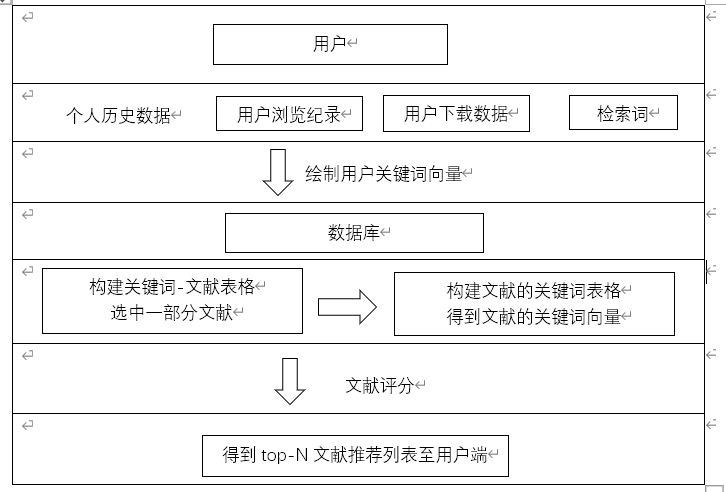


图2.1.1.2 系统体系架构

传统的基于内容的推荐算法主要基于物品与物品之间的相似度。“内容”是构成物品的要素，即用来描述项目内容的属性特征，但大部分物品的属性特征很难提取，且精确度得不到保证，用户对不同特征的偏好可能有差异，如果一件物品附带有两个特征，那么很有可能其中一个特征是用户所偏好的，而另外一个特征只是处于一个恰好被提及的地位。在基于关键词权重的推荐算法中，将关键词看作是文献的特征，然后通过频率关系计算关键词权重将不同的关键词区别开来，在减小计算量的同时提高系统的精度。基于关键词权重的推荐算法有如下几个优点：

(1)可解释性强，本质上是基于内容的推荐算法，都是利用用户的使用纪录提取偏好特征，可解释性强，用户能够接受。

(2)没有冷启动问题，一旦有新的论文分布，本身会附带关键词，可以和匹配给相关用户，新旧文献在数据库中没有区别。

(3)对偏好的描述更加精确，权重的引入让文献的推荐更加具有针对性，有利于用户找到更适合自己的论文。

2.1.2 用户关键词向量的构建

通过计算关键词权重，可以将用户的下载纪录和浏览纪录构建成一个关键词向量[8]：

其中表示关键词，为该关键词所对应的权重，在有检索词的情况下，可以从检索词中提取关键词，然后将这部分关键词加到关键词向量中，得到一个个性化的检索系统。再次基础上，目前的模型存在有关键词难以拓展的缺点，将一些与已有关键词余弦相似度较高的关键词在加关键词向量中，可以使推荐又具有一定的发散性。

2.1.3 关键词-文献对应关系的构建

数据库中有上亿篇文献，我们先依据用户关键词向量中已经出现的关键词圈定一部分文献，然后再分析这部分文献的关键词进行打分。

2.1.4 获取Top-N 推荐列表

前面提到了用户关键词向量以及文献的关键词已经获取，接下来需要得到圈定文献的评分，在向量的余弦相似度[9]的基础上，定义一篇文献对于用户的评分为：

选取评分最高的N篇文献得到推荐列表。

2.2关键词的提取

2.2.1 基于深度学习的分词模型

前面讲到有部分文献关键词信息缺失，为了完善关键词信息，本文采用了RoBERTa[10]中文深度学习预训练模型作为关键词提取的模型。且使用了多种语言模型并根据经验比较了其性能。

(1)n-gram语言模型[11]

考虑一个由个中文字符组成的序列，根据全概率公式我们有：

如果我们假设每个位置上的字符仅与它前n个词有关，上述公式即可简化为：

根据贝叶斯公式，上述公式右边的每一项可以写作：

这样我们根据n-gram模型（n=2或3）可以计算出语料中相邻词语的条件概率。而经验2告诉我们，属于同一个语词的字符的出现的相关性是相对高的，也即上述的条件概率值更高。于是，选择合适的阈值可以较好地完成分词任务。

(2)MLM(Masked Language Model掩码语言模型)

掩码语言模型是在BERT[12]（Bidirectional Encoder Representations from Transformers，由Transformers模型而来的双向编码表征模型）预训练过程中采用的一种无监督算法。标准的RNN（recursive neural network递归神经网路）模型或者LSTM（long short term memory）模型只允许单向的训练，因而模型只具备表征单向的语义的功能。而MLM模型随机的将一定百分比的词替换成[MASK]，然后预测这些被掩码覆盖掉的字符，这样MLM就允许每个词间接地“看见自己”。

在分词任务中，我们沿用了MLM模型。我们定义第个词对第个词的依赖度为：

而由于语词之间的相关度应具有对称性，因此定义第个词和第个词的相关度为：

将文本逐词按相关度大小排序，选择合适的阈值可以较好地完成分词任务。

(3)jieba分词[13]

Jieba分词是一款开源的python中文分词库，是传统的中文文本处理任务的分词库选择。（1）和（2）的方法都依赖于深度学习模型，需要一定的模型加载空间和时间。此外，两种方法都依赖于一个阈值超参数的选择，而在试验中这个超参数与文本的具体内容相关，因此分词的效果相对不理想。而相比于前两种方法，(3)具有封装性好，计算性能相对优秀，最终效果可靠的优点。因此我们最终选择使用jieba库实现分词。

2.2.2 基于深度学习的关键词提取算法

完成了分词任务后，我们要从中找出能最好地代表文本意思的一个或几个语词作为文本的关键词。这些关键词将作为用户搜索输入的特征表示，进入下游的推荐系统。

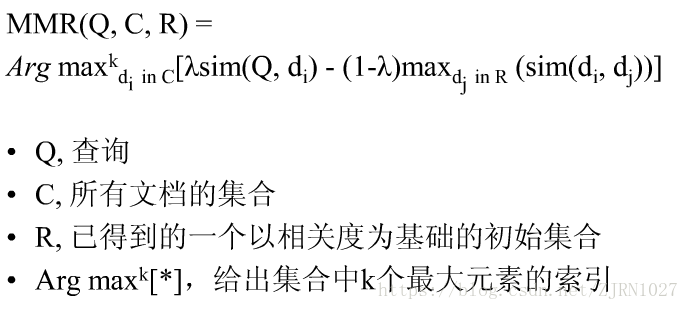
(1)基线算法

我们采用的基线算法为简单的余弦相似度排序。关键词应要能够涵盖一个字符中语义最重要的部分，因而关键词与字符之间的相似度要尽量高。考虑到经过BERT模型后每个字符都对应一个768维的高维向量，采用余弦相似度的准确性更高。

但是基线算法面临一个问题：它将平等对待极相似的关键词，因而结果中很可能出现几个意义相同但多字少字的关键词，从而影响关键词提取的效果，乃至之后的推荐系统性能。

(2)改进方法[14]

我们考虑引入惩罚项来限制意义相近词的同时出现。具体有两种算法：一为MSS(maximum sum similarity最大相似度和)，二为MMR(maximum marginal relevance最大边际相关度)。前者将关键词之间的余弦相似度相加，更加不同的关键词组合就会增大这个和。后者根据公式：



这个算法同时考虑了关键词与文本的相关度和关键词之间的关联度，是控制关键词多样性的超参数。下图展示了加入惩罚项前后的关键词提取效果的区别：

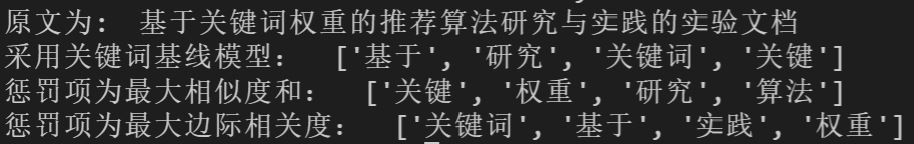


图2.2.2 关键词提取效果对比

2.3 关键词权重的获取

2.3.1 关键词权重的定义

引言中提到大部分文献本身具有关键词，这些关键词对文献有一定的概括性。由于用户对关键词的选择具有一定的主观性，不同关键词对文章的概括程度有所不同，比如对于一篇论文：“基于互联网大数据的检测服务平台的实现”，关键词为“互联网技术;创新;检测服务;大数据”，很明显其中“创新”和“互联网技术”这两个关键词相比另外两个比较宽泛，概括性较小。通常情况下，包含相关关键词的文献数目越多，则该关键词的概括性越强，因此我们可以根据关键词在文献库中的出现次数来对关键词加权，这就是文献的关键词权重。而根据用户与关键词的关系，我们可以得到用户的关键词权重。用户对某一关键词的使用频率越高，则权重越大，同时考虑到用户兴趣偏向会随着时间而改变，对关键词的每一次使用可以按照时间的远近关系进行加权，以降低早期使用纪录对推荐结果的影响，然后再计算使用频率，即得到文献的关键词权重。将二者结合，我们可以得到关键词权重。

2.3.2 关键词权重的计算

如上文所述，我们将其浏览及下载纪录表示成一个关键词向量：

表示关键词，为该关键词所对应的权重，利用TF-IDF公式[15]，TF为关键词频率，这里用来表示用户的关键词权重，IDF为逆向文件频率，用来表示文献的关键词权重，其计算方式为：

其中表示关键词在用户纪录中的出现次数，表示用户纪录中所有关键词出现的次数的总和，表示文献总数,表示出现关键词的文献总数。

在此基础上本文意图引入时间元素。传统的推荐算法大多会忽略时间因素，然而在现实中，科研人员在不同的时间段所需要的文献很有可能是不完全相同的，甚至有一些会面临研究大方向的转变，所以时间权重的引入便十分有必要。考虑到用户使用关键词时的时间，当下时间为,我们对此次使用的时间权重定义为：

其中，为时间因子，可以调整时间远近的影响程度，越大则随时间变化越快，可以依据不同资源的发展速度改变的值。此时关键词出现次数被重新定义为：

当时间比较接近时，为了减小运算的复杂度，我们可以近似将时间权重视为1.

**3 数据集与实验设计**

第2节展示了基于关键词权重的推荐系统的基本原理和流程，接下来进行实验设计和模型呈现。

3.1数据集

本文实验的数据集为万方知识数据平台2020年的部分用户浏览纪录和下载纪录，使用纪录2843595条，涉及的用户395266个，文献1558865篇。原始数据包含用户ID、用户类型、文献ID、文献标题、浏览时间、省份、文章关键词、作者等信息，如下图所示：



图3.1.1 部分用户使用纪录

在本文中，针对实验数据集的数据预处理包括如下操作：

(1)删除数据中的无用字段，留下用户ID、文献标题和关键词以减小操作量。

(2)清除表中的无效数据，例如缺少文献标题的使用数据。

(3)将关键词拆分成列表，原表的关键词合并展示，本文将其分隔开并统一格式以方便后续操作。

(4)将数据按ID + title + keywords的新式存储到txt文件中，如下图所示：

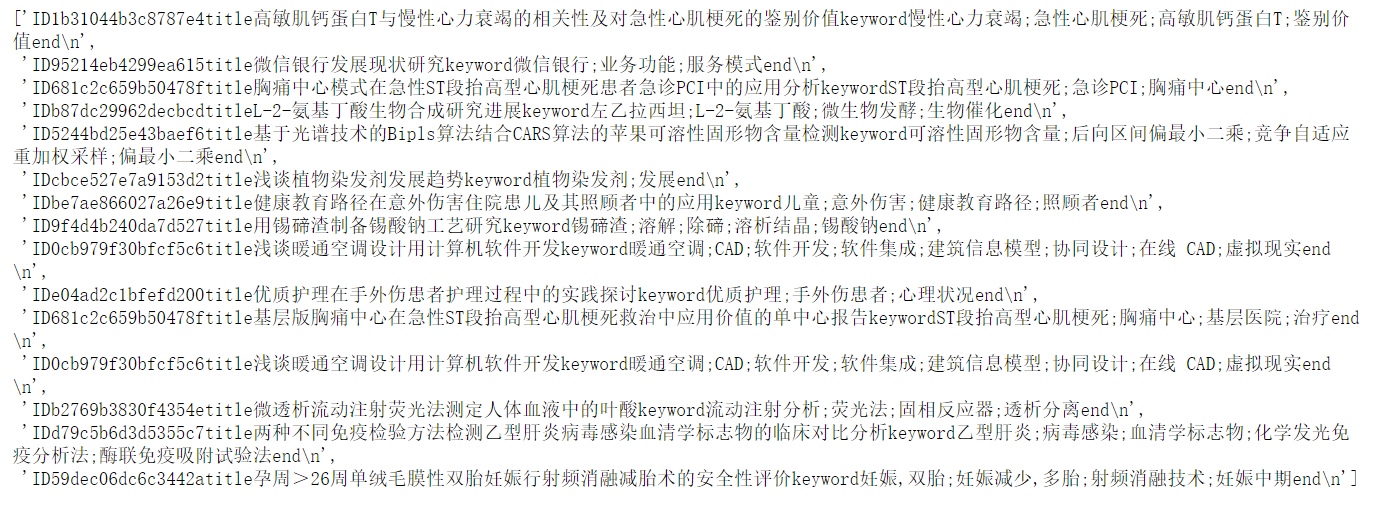


图3.1.2 数据存储形式

随后，本文随机选取一部分用户，对单个用户的使用纪录进行整合并随机分成两组，其中一组用来得出推荐列表，第二组数据用来判断模型的性能。

3.2 获取top-N推荐列表

在获取了单个用户使用纪录数据集(图4.2.1)后，将每条数据的关键词提取在一个列表中，考虑到重复浏览可能代表对文献的兴趣越大，本文对重复的浏览数据采用同样的处理方法。由于数据集中只包含了2020年的数据且数据完整性不足，本文未将时间权重考虑在内。由此得到用户关键词频率向量，再在数据库中查找每个关键词所带有的文献数量，将二者结合使用第一节中关键词权重的计算公式，得到用户的关键词向量，该过程如图所示：



图3.2.1 提取后某一用户的历史数据

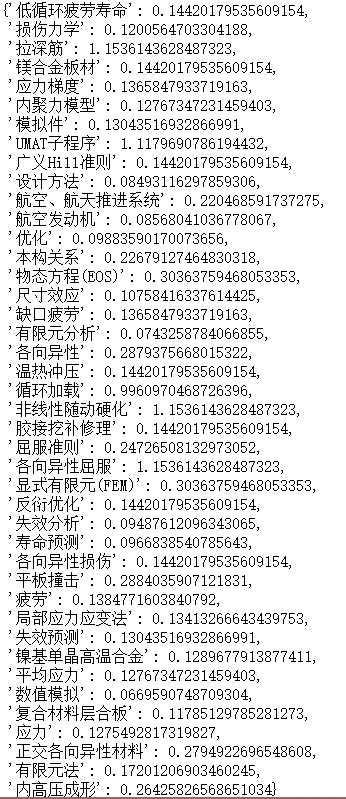


图3.2.2获取的关键词向量

得到关键词向量后，我们将所有关键词所对应的所有文献标题集中在一个表格中，然后利用2.1.4中的评分计算公式对每一篇论文进行打分，然后选取评分最高的N项呈现出来，当N为10时，我们得到的推荐列表(范例)如下：

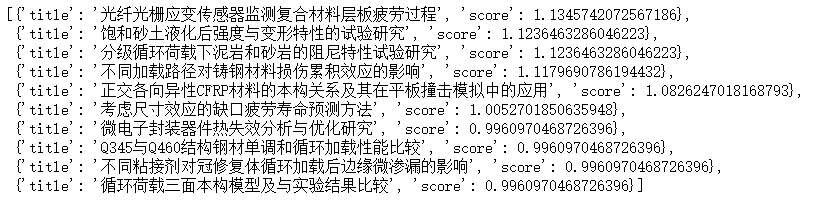


图3.2.3 推荐列表及其内文献的评分

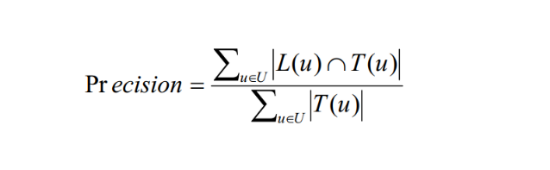
**4 模型评估**

4.1 评价指标

推荐算法的性能衡量指标有多样性、满意度、准确率、实时性等，考虑到数据的特征，本文选择三种常用的推荐系统评测指标有准确率、召回率和F值。[16]

4.1.1 准确率(Precision)

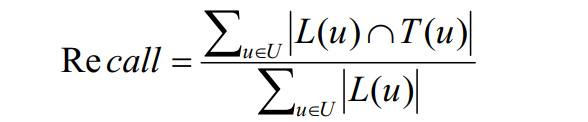
准确率用来衡量推荐列表的正确率，通常情况下为用户看到推荐列表L(u)后所做出反应的比例，本文用视测试集内的数据Y(u)为用户的未至反应，计算方法如下：



U为用于测试的用户集合。

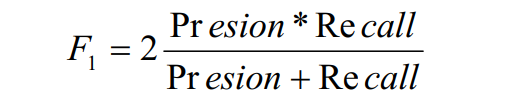
4.1.2 召回率(Recall)

召回率衡量推荐列表的完整度，为推荐正确的文献书占据所有合适的文献(Y(u))数的比例，计算方法如下(推荐列表为L(u))：



4.1.3 F值( score)

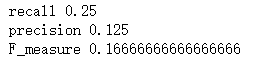
推荐系统的预测成分很大，很难做到两者都很高，通常情况下准确率高则召回率低，召回率高则准确率低。而F值则兼顾召回率和准确率，对模型的综合特性的评估更加全面，其计算方法为：



这三者越接近1则模型的性能越好。

4.2 结果分析

本文首先对图3.2.3中从单个用户数据中获取的推荐列表进行评估，得到再N为10时，召回率、正确率和F-值分别如下图所示：



由于推荐列表数目有限而测试集庞大，数值较低，通过改变N值的大小观察评价指标的变化，得到折线图如下：

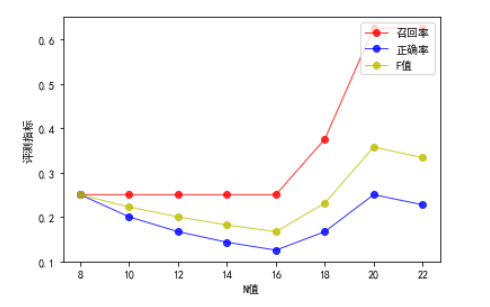


图4.2 不同N值下模型的精度值

可知当N取20时，即推荐列表中文献数目为20篇时，推荐结果最好，召回率超过了0.6，正确率接近0.3。在用户未看到推荐列表的情况下能仅根据个人历史数据的拆分出去一个测试集进行评估，且在文献数量十分庞大的前提下，达到这样的结果表明基于关键词的推荐系统效果可观，具有一定的可行性。

5 总结

本文提出了一种基于关键词权重的的推荐系统，在系统的搭建过程中为了解决部分文献缺乏关键词以及关键词不够准确的问题又提出了一个基于深度学习的分词模型，为系统正常实际运行提供保障。基于关键词权重的推荐系统利用了文献的特殊性(具有关键词)，相比于基于内容的推荐算法省去了很多物品特征提取的步骤，相比于协同过滤的推荐算法解决了冷启动问题，且具有可观的推荐结果。不足的是推荐的效率比较低，未来拟对系统算法进行优化，使其能在现实语境中更好的发挥作用。

参考文献：

[1] R. Eric Hostler，Victoria Y. Yoon，Tor Guimaraes. Recommendation agent impact on consumer online shopping：The Movie Magic case study［J］. Expert Systems With Applications，2012，39（3）

[2] Balabanovic M, Shoham Y. Fab: Content-based, collaborative recommendation//Proceedings of the Communications of the ACM. Zurich, Switzerland, 1997: 66-72

[3] Mooney R J, Roy L. Content-based book recommending using learning for text categorization//Proceedings of the 5th ACM Conference on Digital Libraries. San Antonio, USA, 2000: 195-204

[4] Salakhutdinov R, Mnih A. Probabilistic matrix factorization. Advances in Neural Information Processing Systems, 2008:1257-1264

[5] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms//Proceedings of the10th International Conference on World Wide Web. Hong Kong, China, 2001 : 285-295

[6] Tso-Sutter K H L, Marinho L B, Schmidt-Thieme L. Tag-aware recommender systems by fusion of collaborative filtering algorithms//Proceedings of the 2008 ACM Symposium on Applied Computing. Fortaleza, Brazil, 2008: 1995-1999

[7] Zhang Z K, ZhouT, Zhang Y C.Tag-aware recommender systems: Astate-of-the-art survey. Journal of Computer Science & Technology, 2011 , 26(5): 767-777

[8] 刘宇,朱文浩. 基于内容和标签权重的混合推荐算法[J]. 计算机与数字工程,2020,48(4):773-777. DOI:10.3969/j.issn.1672-9722.2020.04.008.

[9] Park,Hong,Kim. A Methodology Combining Cosine Similarity with Classifier for Text  
Classification[J]. Applied Artificial Intelligence,2020,34(5):45-52.

[10] Facebook AI, RoBERTa: A Robustly Optimized BERT Pretraining Approach. [J/OL] arxiv:1907.11692

[11] 关毅. 2007. 统计自然语言处理

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J/OL].https: //arxiv. org/abs/1810.04805,2018-10-11

[13] “结巴”中文分词：做最好的 Python 中文分词组<https://github.com/fxsjy/jieba>

# [14] 刘欢,张智雄,王宇飞. BERT模型的主要优化改进方法研究综述[J]. 数据分析与知识发现,2021,5(1):3-15. DOI:10.11925/infotech.2096-3467.2020.0965.

[15] 唐明,朱磊,邹显春. 基于Word2Vec的一种文档向量表示[J]. 计算机科学,2016,43(6):214-217,269. DOI:10.11896/j.issn.1002-137X.2016.6.043.

[16] 甘丽新,万常选,刘德喜,等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展,2016,53(2):284-302. DOI:10.7544/issn1000-1239.2016.20150842.

汪晓锋：提出了算法模型，负责论文的编排和数据预处理，并对模型进行实验测试。

罗逸杰：提出了关键词提取模型，并进行算法实现。

容天泽：负责海报制作及论文英文部分的写作。