**AI模块**

1. **新闻分类部分**
   1. 基于**Transformer**的多标签文本分类

2017年Google提出了网络结构Transformer}39}，在这之前传统的编码器、解码器模型结

构通常需要结合循环神经网络或者是卷积神经网络，Transformer打破了这种固有模式，仅仅采用了注意力机制来执行自然语言处理的任务。

Chang等人针对极端多标签文本分类中存在的输出空间大和标签稀疏性问题，在传统Transformer模型的基础上提出了X-Transformer模型，这是针对极端多标签文本分类问题的第一个微调Transformer模型的可扩展方法。考虑到多标签文本分类标签之间的相关性，Gong等人提出了一种基于层次图Trnasformer的深度学习模型，首先将文本以图形结构的方式进行建模，该模型结构可以体现文本的不同语义以及它们之间的联系。然后在单词、句子和图形层次上通过Trnasformer结构充分捕捉文本的特征，并分析各个部分的重要性，并且在标签表示部分充分考虑了标签之间的层次关系，并根据标签之间的语义距离设计一个加权损失函数。Jiang等人研究认为X-Transformer模型在训练标签排序模型的过程中静态采样负标签，降低了模型的效率和准确性。并提出了LightXML模型，该模型采用端到端训练和态负标签采样的方法，在模型中使用生成式合作网络对标签进行回收和排序，其中回收标签的部分生成否定标签和肯定标签，而排序标签的部分则将肯定标签与这些标签区分开来，这样在标签排序部分训练过程中，通过输入相同的文本表示来动态采样负面标签。Pujari等人为了解决多标签分层文本分类中存在的内存限制问题，第一次提出了将分层算法和Transformer结合，设计出一种基于Transformer的多任务模型，该模型使用多任务设置并且共享一个底层语言模型，为每个节点训练一个分类器。在大型多标签文本分类场景中，Shaheen等人结合生成式预训练、逐步解冻和区别学习等策略，研究了各种基于Transformer改进模型的性能，使最后分类器的性能具有竞争性。

2018年Devlin基于Transformer提出的预训练模型Bert更是在自然语言处理领域取得了突破性的进展。大批学者将Bert应用到多标签文本分类任务中，并取得了理想的效果。Chang等人第一次调整Bert模型并尝试应用于大型多标签文本分类场景，提出了X-Bert模型，该模型利用标签和文档文本来构建标签表示，从而引入语义标签集群，以便更好地建模标签依赖关系。X-B ert的核心是对Bert模型进行微调，最后将不同的标签簇上训练的不同Bert模型进行集成，得到最佳的最终模型。Leheka等人同样将Bert模型运用在了大规模多标签文本分类任务中，在Bert模型的基础上提出了一种池化层结构，它利用标准[CLS]令牌的信息结合池化序列输出来提高分类质量。Zhang等人将增强层次信息的模型和Bert模型相结合提出了一种新型的Hie-Bert模型，该模型首先通过Bert模型获取电子病历信息的特征表示，之后对获取的特征表示中不需要的信息进行过滤，并与诊断内容中的关键信息进行融合，再经过数据增强之后，执行分类操作;该模型与传统Bert模型相比，有着更好的分类性能。Jin等人提出了一种基于Bert和改进TF-IDF的多标签文本分类模型。在文本特征提取阶段，提出了一种改进的TF-IDF方法，通过计算单个类别的不同权重，更好地反映单词在多标签分类问题中的重要性。将Bert与改进的TF-IDF相结合，生成一种新的特征用于分类任务。然后将特征输入到全连接层中，对Bert模型进行微调，从而实现多标签分类任务。Huang等人将包含每个单词和目标实体之间的依赖关系的高级语法特征，合并到预训练的语言模型中，还利用Bert的中间层来获取不同层次的语义信息，并设计多粒度特征来实现最终的关系分类。Kim等人提出了一种基于CO Bert的多标签文本分类模型EnvBert，利用了数据过采样技术解决了多标签数据的不平衡特性，该模型对不平衡的、有噪声的多标签环境新闻数据具有良好的预测性能。Yarullin等人针对多标签文本分类问题，研究了Bert语言表示模型和带有Bert编码器的Seq2Seq模型，证明了带有Bert编码器的Seq2Seq模型与标准Bert模型相比，在更少的训练时间内获得了不错的结果。

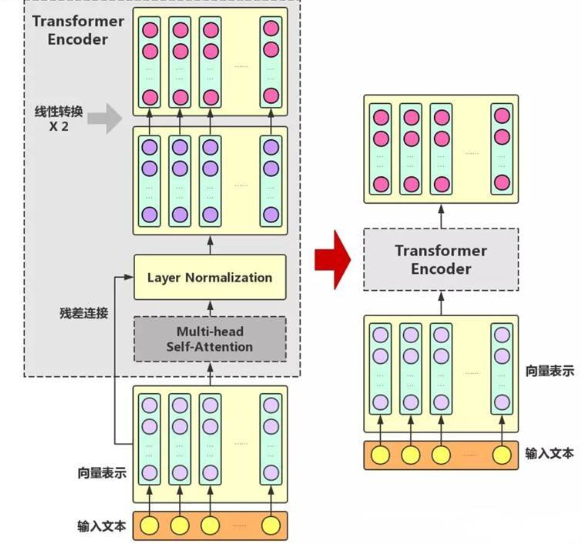


图 Transformer原理解析图

* 1. 算法设计目标

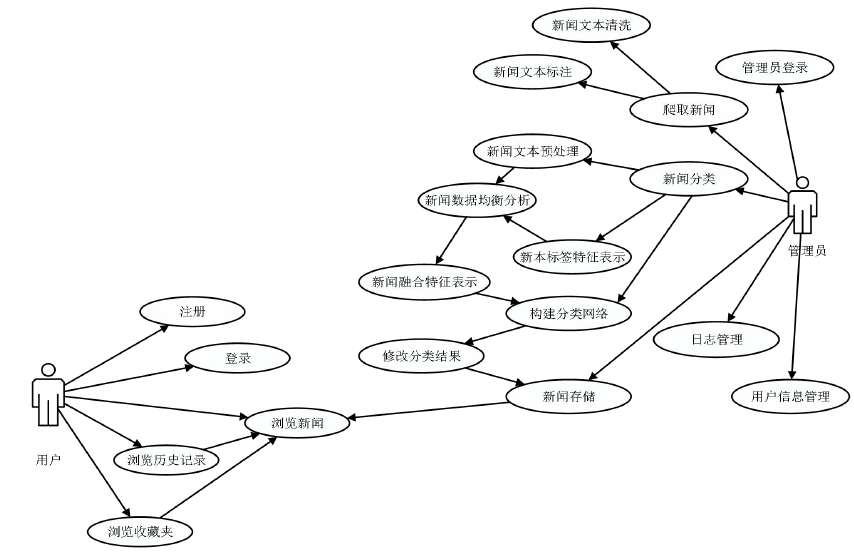


图2 设计目标图

* 1. 基于注意力机制的**FAA-MLSMOTE**算法
     1. 详细设计

1. 基于标签森林的分层样本扩充

针对网络新闻文本的分类形式，数据的不均衡主要体现的父标签之间不平衡以及父标签内部的子标签分类不均衡这两种情况。本小节提出一种基于标签树的分层样本扩充方法，首先需要筛选出高层次的少数标签对应的样本并扩充，之后再去下一级层次进行筛选并扩充。为了实现这样的分层扩充方法，首先要搜集所有标签并构建一片标签森林，标签森林中的每一棵标签树都层次鲜明地显示着每个标签之间以及标签内部的层次关系，如图3所示，之后对从高层开始逐层进行样本的选择和扩充。

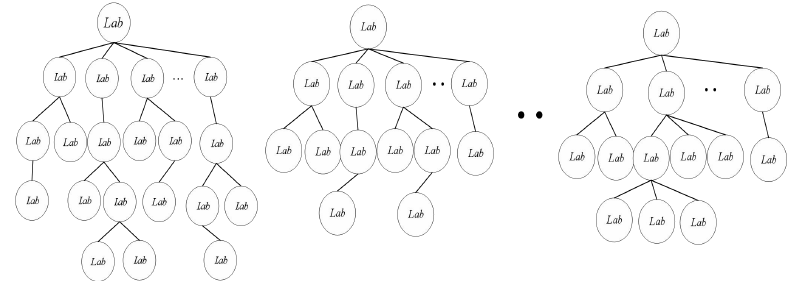
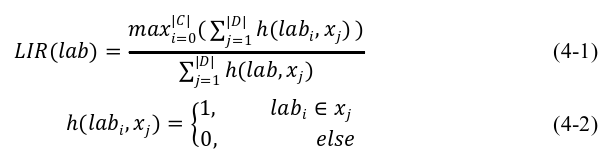


图3 标签森林图

图中每一棵标签树的根节点代表着一个最外层的父标签。

步骤1：首先遍历标签森林，从每一颗标签树的最高层开始，筛选出少数标签样本。

步骤2：将该层的标签节点按照兄弟关系划分集合。首先计算每一个兄弟集合中的标签节点的失衡比例，如公式(4-1)、(4-2)所示：



其中labi表示标签，xj，表示样本**|**C**|**表示兄弟集合中所有标签个数，**|**D**|**表示所有样本个数，h(labi, xj)函数表示若标签i属于样本j,则值为1。每个标签的失衡比例实际上是集合中出现次数最多的标签与检测标签出现的次数的比值。

步骤3:当集合中每个标签的失衡比例计算完成之后，计算该集合平均标签失衡比例。如公式((4-3)所示。



步骤4:当集合中的某一个标签的失衡比例大于该集合的平均失衡比例，那么该标签就本加入到待扩充的集合L-Collection中。

步骤5:遍历L-Collection集合，获取少数标签。

步骤6:遍历样本集合，获取与步骤5得到的标签相关联的样本，并将其加入到扩充样本集合，并得到集合中每个样本在特征空间的分布情况。

步骤7:遍历特征空间的每一个样本，并计算每一个样本之间的欧式距离，具体公式如(4-4)所示，在特征空间中选择5个距离该样本最近的其他样本。



其中N代表样本特征表示的维度。

步骤8:随机选择5个其他样本中的其中1个样本，并计算该样本和需要被扩充样本各个维度之间的差距，得到相对差距向量Vecdif，如公式(4-5)所示。



其中Vecsld表示选中的样本的特征向量，Vexrdm表示需要被扩充样本的特征向量。

步骤9:随机生成距离因子fac，该因子值的范围在0-1，将距离因子结合到步骤8计算得到的差距向量中，旧向量与差距向量进行融合得到新的样本向量Vecnew，如公式(4-6)所示。并对每一新样本生成对应的标签集合。



步骤10:根据少数样本需要扩充的倍率N，对步骤8步骤9的操作循环N次。

步骤11:一层样本扩充完毕之后，回到上一层，继续执行步骤2到步骤10的操作。

步骤12:当所有标签树扩充完毕之后，最后对整个标签森林以树为单位进行样本扩充，以此完成对所有少数标签样本的扩充。

以这样层次化的结构对标签森林中的每一棵标签树，标签树中的每一层进行样本的选择和扩充，不仅考虑到了每一个标签之间的分布不均衡，而且对每一个标签内部的不均衡分布进行了详细的分析。完整的样本生成策略伪代码如图4所示。

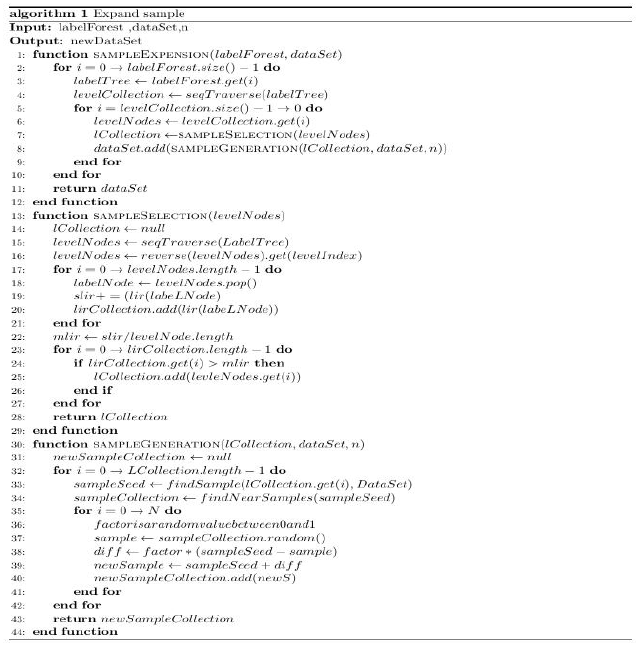


图4 标签集生成伪代码

1. **FAA-MLSMOTE**算法流程描述

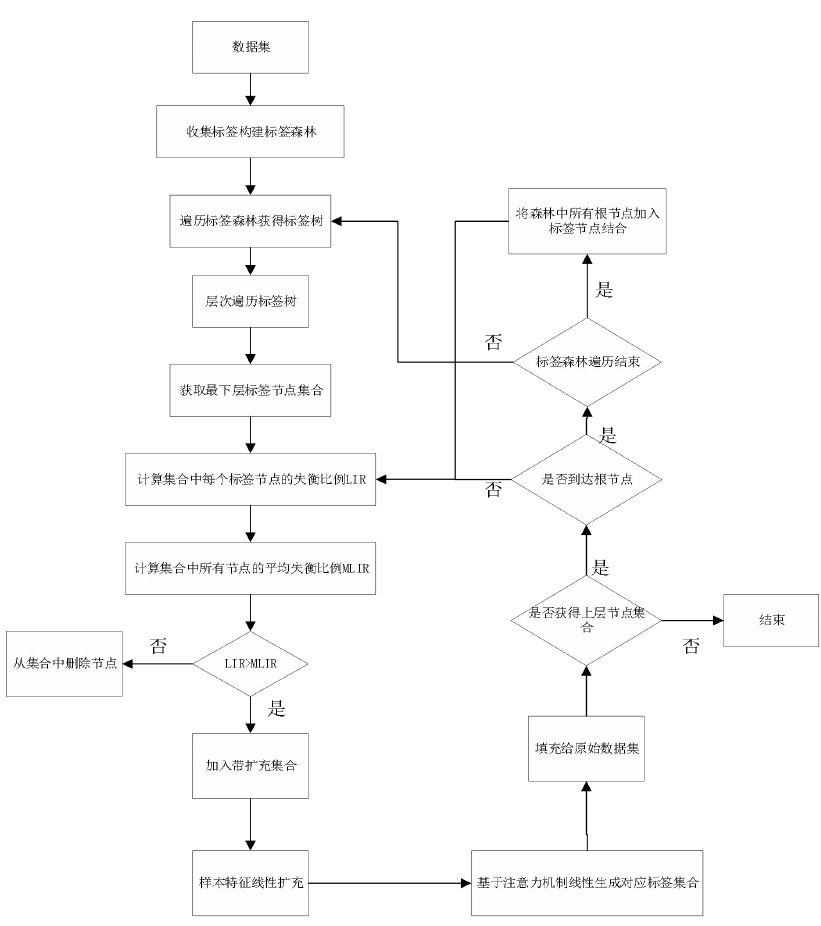


图5 FAA-MLSMOTE算法流程图

1. **新闻摘要模块**
   1. 基于**MMR** 的改进模型

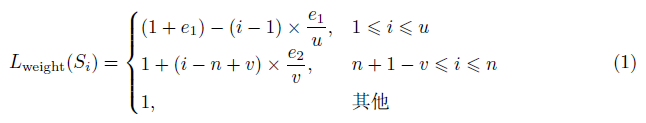
在实现该模型时，首先需要综合考虑影响新闻摘要的各个因素，计算新闻文本句子的初始权重。然后，通过改进的MMR 算法对句子的初始权重进行迭代，得到这些句子的最终权重。最后，按最终权重的大小对句子进行排序选出一定比例的句子，进而按照句子在原文中的顺序输出所抽取出的句子得到摘要集。

* + 1. 影响新闻摘要的因素

根据已有的抽取式文本摘要相关研究成果、对人工标注文本摘要过程的分析以及对新闻文本特点的归纳，本文总结出了以下4 个影响新闻摘要的因素。

1. 句子的位置权重

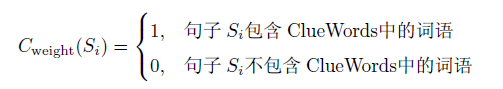
句子的位置与新闻文本的主题存在相关关系，一般来说，文本的第1段或最后1 段通常是对整个报道的总结，具有高度的概括性。设新闻文本的第1 段由u个句子组成，最后1 段由v 个句子组成，文本中句子总数为n，则可设置句子Si 的位置权重Lweight(Si) 为



该式保证了在第1 段话中距离第1 句越远的句子，其权重越小；而在最后1 段话中，距离最后1 句越近的句子，其权重越大。式中，e1 和e2 分别用来控制第1 段和最后1 段所有句子的相对初始权重，一般在0 到1 之间。具体值可以根据待处理新闻文本的特征进行设置。如果新闻中第1 段的概括性更强，则将e1 设为比e2 更大的数，例如e1 为0.5，e2 为0.1。

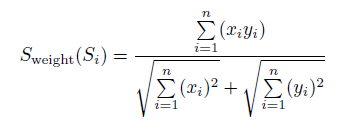
1. 线索词与转折词

线索词和转折词通常引出具有总结性或强调性的句子。包含线索词和转折词的句子往往比不包含该类词的句子更能表达新闻主旨。设ClueWords 代表线索词和转折词，Cweight(Si)表示句子Si 的线索词和转折词权重，则该权重的计算公式为



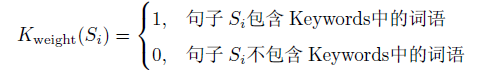
1. 标题相似度

标题往往是新闻内容的高度凝练，因此与标题相似度高的句子应具有更高的权重。本文使用预训练BERT 模型对句子和标题进行向量化，然后将两个向量的余弦相似度作为句子和标题的相似度。设句子Si 和标题T的句向量分别为(x1,…,xn) 和(y1,…,yn)，则句子Si的标题相似度权重Sweight(Si) 的计算公式为



1. 关键词权重

含有文本关键词的句子通常比其他句子具有更多的文本有效信息。本文通过TF-IDF 算法来抽取新闻文本和新闻标题中10个关键词作为关键词表，使用Keywords表示关键词集合。如果句子中含有关键词，则关键词权重值Kweight(Si) 赋值为1，即



* + 1. 初始权重计算

得到上述影响新闻摘要的4个权重值后，需要设计加权算法以计算最终的初始权重。首先由Lweight、Cweight、Kweight加权得到中间权重wmid，即



式中：α,β,γ为加权系数，且α + β+ γ = 1。在具体应用场景下，加权系数可以根据文本不同级别权重值的取值范围及其重要程度进行设置。

再将中间权重wmid 与标题相似度权重Sweight 加权，得到句子Si 的初始权重w(Si)



式中，为了保证Sweight 和wmid 在同一数量级，而且为了能够动态调整wmid 的权重，引入了调节因子**δ**。

* + 1. 改进MMR算法

初始MMR 算法是用来根据查询对文档集合进行排序的，本文对MMR 算法进行了改进，使其适用于面向新闻文本的抽取式摘要。抽取式摘要的目的是对句子进行合理的排序，然后根据顺序选出一定比例的摘要集。根据这个目的，设计的面向摘要的MMR 算法如下：



式中：D 是通过MMR 算法得到的阶段性候选集，**λ**为控制摘要概括性与冗余性的一个系数。

该算法的公式共分为两部分，w(Si) 计算的是新闻本文中所有句子的初始权重，similarity(Si,D) 计算的是新闻文本中句子与已入选候选集句子之间的相似度，其中-(1-**λ**)是负值。如果当前句子与摘要集句子之间的相似度过大，那么该算法会对当前句子的权重进行惩罚，以尽可能地减少摘要集的冗余。

使用该算法对初始权重进行迭代的流程图如图6所示。

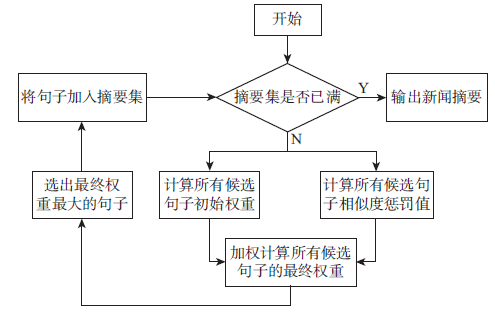


图 6 MMR算法流程图

1. **关键词提取模块**
   1. 改进的基于文档主题结构与语义的方法

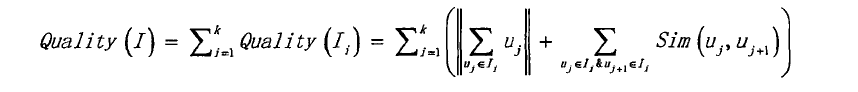
通过同义词林，计算中文词语之间的语义相似度，使得算法能够考虑到同义或者同类词语之间的语义关联，进一步提高关键词提取的效果。

加入语义因素后，将聚类中的簇对应的簇心与簇中其他所有成员的余弦相似度之和以及簇中所有相邻连续段的语义相似度之和，两者相加求得的值定义为目标函数Quality。对于第i个簇，其目标函数Quality(Ii)的值可以通过下面公式求得。



其中，Sim(uj,uj+1)表示分段uj和分段uj+1的语义相似度。

因此，聚类目标函数将通过下式求得。



在计算两个中文词语的语义相似度的算法中，将weightArr权重数组的值设置为[1.0,0.5,0.25,0.125,0.063,0.031]，前一项是其紧随后一项的两倍。把距离初始值initDis设为10。将β设置为5。

在计算簇的目标函数的时候，提及需要计算同一个簇中两个连续段的语义相似度Sim(uj,uj+1)。对于每个文本分段，可以将其看成词语的集合。只有当文本分段si包含词语tj，即tf(tj, si)>0的时候，词语tj才会出现在分段si的词语集合中。段与段之间的语义相似度问题可以转变成带权二分图Ｇ＝{V, E}的最佳匹配问题。其中，V指的是图G的顶点集，而E指的是图G的边集。对于分段ui和分段uj ，假设分段ui共有ｎ个词语，而分段uj共有ｍ个词语，那么对应的词语集合分别使用Xi＝{xi1,xi2,…,xin}和Yj＝{yj1,yj2,…,yjn}来表示。把词语集合Xi和Yj中的所有词语当作带权二分图中的顶点。集合Xi中的任意词语与集合Yj的词语都存在一个带权的边，但是集合Xi和集合Yj内部的顶点之间不存在边，构成边集Ｅ。带权二分图中的每条边的具体权重可以使用词语间的相似度Sim(Ta,Tb)求得，如图7所示。

完成带权二分图的最佳匹配的计算后，将会产生min(|Xi|,|Yj|)对匹配，任意的两条匹配边不存在公共顶点，并且使得全部匹配边的权重和取得最大值。使用KM（Kuhn-Munkres )算法来解决最佳匹配的问题，而后，将最终得到的权重和除于min(|Xi|,|Yj|)，避免分段的不同长度大小造成的噪音影响。最终，得到两个分段ui，和uj的语义相似度Sim(uj,uj+1)的值。

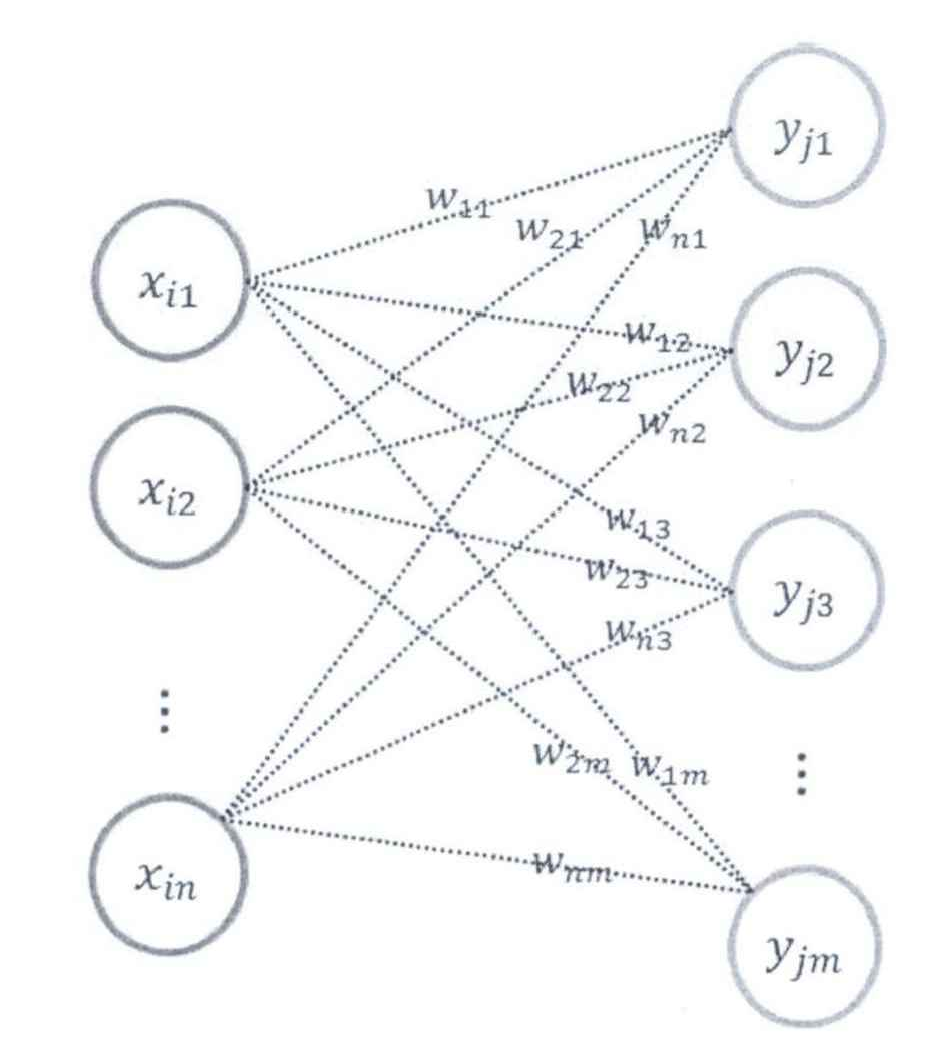


图 7 分段ui和分段uj对应的带权二分图

* 1. 算法流程

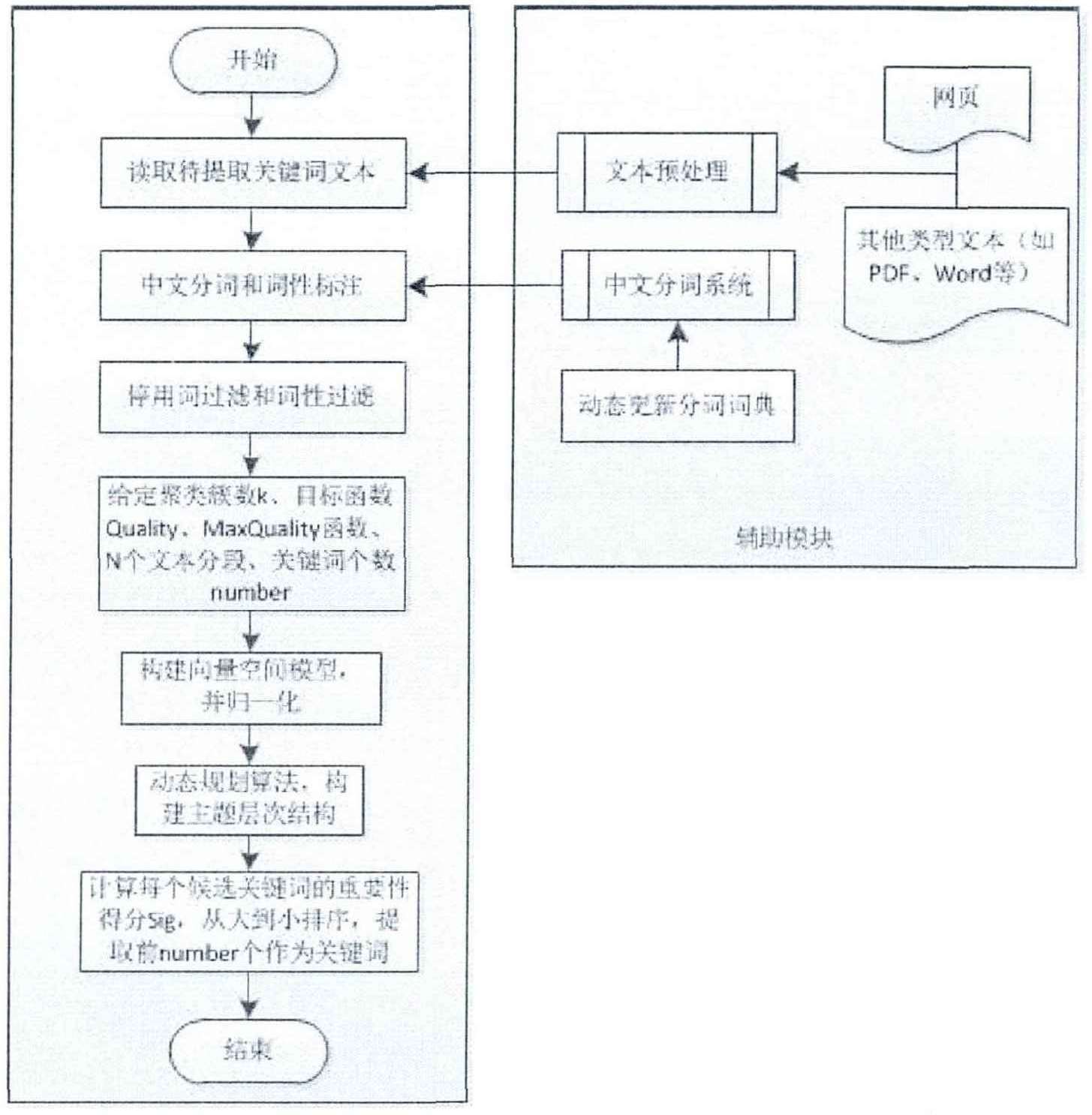


图 8 算法的流程图

1. **新闻热度模块**

算法流程：

