本项目将大数据和机器学习算法结合起来，针对项目涉及的流程，从题库题目的生成到题库题目的评价与维护，再到用户答题情况的分析，提出了基于机器学习模型的智能化自动出题算法、基于用户评价与机器学习的题目优化与增强、基于大数据的用户答题情况挖掘算法，实现了从文本到题目的智能化自动出题，从而减少人工出题的工作量，并在考虑用户对题目评价的基础上，运用机器学习方法对题目设置进行优化和调整，通过对用户答题情况的大数据分析，获取整体用户的学习情况以及每个用户的知识掌握情况、知识盲区等，从而为其制定个性化的题目练习和学习计划，以提升其学习效果。

一 基于机器学习模型的智能化自动出题算法

传统的题库建立、维护往往需要人工手动进行，不仅具有人力耗费大、时间消耗多、更新速度慢等弊端，而且无法满足微信答题小程序对于题目高时效、自动化的需求。为了解决这一矛盾，本项目提出了基于机器学习模型的智能化自动出题算法，实现了从给定素材文本到自动生成题目的。

智能化自动出题是指运用给定的文本段落素材，通过机器学习模型，自行生成包含题目题干、正确选项以及错误干扰选项在内的完整题目，其实现主要包含四个个过程：

（1）意群划分：将给定的文本段落划分为多个意群，各个意群包含着不同的含义，即不同的“知识点”，因此将这些意群分割开来是生成题目的第一步；

（2）自动文摘：从划分出的意群中抽取重要信息概括该意群的主要思想，作为题目的题干；

（3）主体识别：运用主体识别分析出文本主要思想中包含的主体内容，例如：时间、地点、事件、人物、组织等，作为题目的正确选项；

（4）干扰项生成：运用词向量及聚类分析，在语料库中选择与正确选项相近的词语，作为题目的错误选项。

1.1 意群划分

意群是指复句中由意义和形式关系相对密切的两个以上的分句所组成的结构中心，或者指的是我们的语言所表达的思想都是通过一群相互关联的意义单位体现出来的，而这些意义单元根据其所处语言片段的角色，有大有小，因此意群分割也就有一个多尺度问题。

句子是由词语和短语组成的，是具有一定语调并表达一个完整意思的语言运用单位。按照结构来分，句子通常划分为单句和复句，复句更具有普遍性。复句是由两个或以上意义相关，结构上互相不构成句子成分的分句组成的句子。复句相对于单句来说结构更加复杂，句子表达的含义也更多。通常，复句都包含多个分句，每一个分句都表达了独立的含义。为了处理的方便，可以将句子中的分句作为意群，运用逗号和分号作为句子意群的分隔符。

为了提高意群内容倾向分析的确性，我们将转折词也作为意群划分的依据。例如，若文中出现“然而”，“但是”等转折词语，可以将转折词前后划分为两个不同的意群。总而言之，意群通常是复句中的分句，或者被转折词隔开的短语。

划分意群的算法如下所示：

输入：给定文本

输出：划分后的多个意群

步骤1： 根据逗号和分号将句子划分成一个个意群

步骤2： 将意群进行分词，并逐个读取划分后的词语，若词语属于转折词，那么截取该意群

步骤3： 若所有句子处理完毕，则转入步骤4，否则转入步骤1

步骤4： 算法结束，输出结果。

1.2 基于主题模型的自动文摘

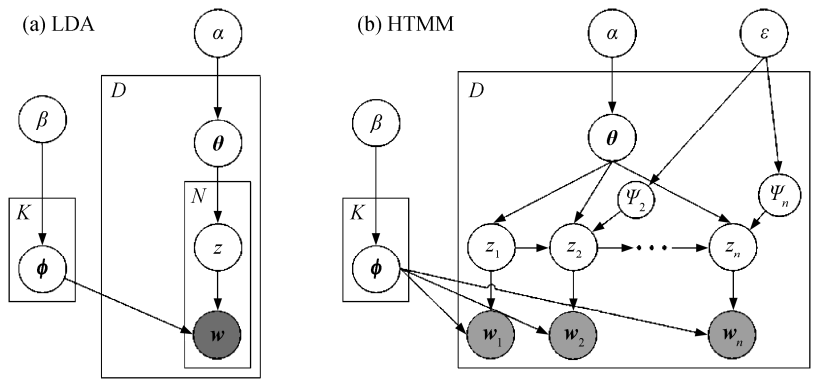
将输入的文本划分成多个意群后，利用自动文摘技术逐一处理每个意群，分析并输出各个意群的“中心思想”，将“中心思想”作为相对应的意群的概述，即可用于题目的题干。

自动文摘指自动从大量的文本信息中抽取重要的信息来代表文本的主要思想。自动文摘按照摘要的产生方式可以分为两类，即生成式和抽取式。生成式产生的文摘可以包括原文中没有出现的词和词组，一般基于实体信息、信息融合以及压缩技术等。生成式文摘的缺点在于领域适应能力弱, 因此生成式文摘的普遍适用性受到了约束由于生成式对自然语言处理技术要求非常高，目前还处于起步阶段，产生的文摘离实用化还有相当的距离。

目前，基于机器学习的方法是抽取式文摘的主流方法，通过选取原文中的句子形成文摘。其方法分为两类: 基于有监督的学习方法和基于无监督的学习方法。有监督的学习方法视自动文摘为分类问题, 文本中的句子被分成两类——文摘句和非文摘句，并按照归属类别的程度选择文摘句子。无监督的学习方法通过文本的语义信息对句子打分，得分高的输出为文摘句。我们使用基于隐主题马尔科夫模型的自动文摘技术进行自动文摘抽取。

主题模型是一种有效的文档浅层语义表示模型。主题可以看成是词项的概率分布, 起源于隐性语义索引主题模型*LSI*。在*LSI*基础上，发展出了三层贝叶斯概率模型*LDA*以及隐主题马尔科夫模型*HTMM*，其句子间的主题关系满足马尔科夫性质, 并且主题转移服从二项式分布.

*LDA* 是一个完整的文档生成模型, 通过主题和词项的概率分布模拟文档的产生过程。在文档的生成过程中, *LDA* 有两个假设: 主题独立假设和词袋假设。然而，*HTMM*在模型的构建上打破主题独立假设。*HTMM*生成文档过程中, 词由不同主题生成, 并且主题与主题之间服从二项式分布。不同于*LDA*模型, *HTMM* 的文档生成过程中需要考虑文档中词的顺序, 通过二项式分布描述句子间主题的转移。



*LDA*与*HTMM*的结构示意图

主题模型将主题表示为隐性变量, 仅文档内容可观测到, 最主要的参数为各主题下的词项概率分布()和各文档下的主题概率分布()。

基于主题模型的抽取式文摘(机械文摘)生成, 其文摘句来自文档中的句子。利用主题模型中的与概率分布, 计算句子的主题概率分布和文档的主题概率分布。文摘句具有与文档相似的主题概率分布, 因而选取与文档具有高相似主题概率分布的句子作为文摘句。生成文摘可分为三个过程：

1) 训练主题模型参数, 得到与概率分布

2) 计算句子主题分布及文档主题分布

3) 比较句子主题分布与文档主题分布, 从而抽取候选文摘句.

使用 HTMM 模型获取与率分布，在参数估计的迭代公式中, HTMM 以句子为单位考虑主题转移概率。然后计算文档的主题概率分布和文档中所有句子的主题概率分布。一方面, 主题模型的参数估计可以直接获得文档的主题概率分布; 另一方面, 句子的主题概率分布来自两部分: 文档-主题的概率分布和主题-词项的概率分布。利用公式计算句子*S*的主题概率分布:

其中*D*表示句子*S*所在的文档, 为主题*Z*下词项*w*的概率, 为文档*D*下主题 *Z*的概率。式中假设词与词相互独立，同时假设文档出现概率均等, 句子出现概率也均等, 并且消除长句优势, 将上式简化为

其中*len(S)*表示以词为单位的句长。

使用 KL 散度值计算两种概率分布的相似程度:

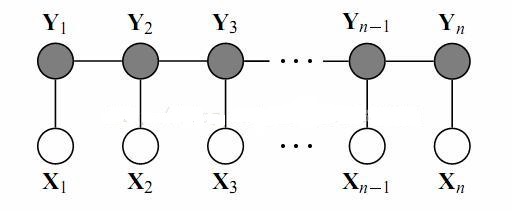
*P* 与 *Q* 为概率分布, *KL* 散度值越小, *P* 与 *Q* 概率分布之间的差异就越小。因此选取句子主题概率分布与文档主题概率分布 *KL* 散度值最小的句子集合作为候选题干, 即 *D*KL(*P*(*Z*|*S*)||*P*(*Z*|*D*))最小。

1.3 基于条件随机场的命名实体识别

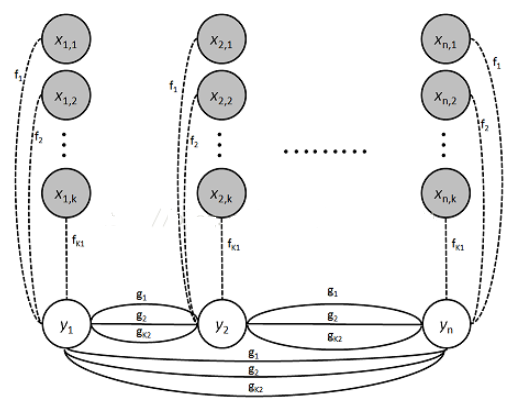
当题干产生后，需要在题干中抽取“问题”，即在题干中确定空白项位置。抽取出的“问题”即作为题目的正确选项。通常，“问题”取自于题干中在日常试题中常见的问题类型，例如时间、地点、人物、会议名称、组织名等等。在题干中抽取问题，可以利用命名实体识别技术。

命名实体识别的主要功能是从文本中抽取出特定的事实信息。当自动文摘生成了文本段落的摘要内容后，利用信息抽取技术，抽取特定信息，作为试题的选项。

命名实体识别（*NER*）是在自然语言处理中的一个经典问题，其应用也极为广泛。其任务是从文本中识别出诸如人名、组织名、日期、时间、地点、特定的数字形式等内容。传统的公认比较好的处理算法是条件随机场（*CRF*），它是一种判别式概率模型，是随机场的一种，常用于标注或分析序列资料，如自然语言文字或是生物序列。条件随机场在命名实体识别中的作用是，给定一系列的特征去预测每个词的标签。如下图：

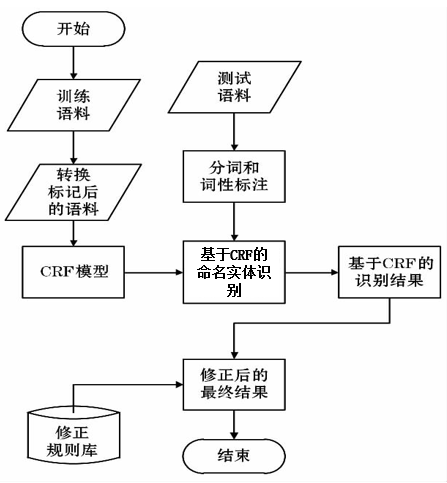


在上图中，*X*可以看做成一句话的每个单词对应的特征，*Y*可以看做成单词对应的标签。这里的标签就是对应场景下的人名、地名、事件名等等。特征*X*可以表示词性，如名词、动词等，然而单一的特征无法完成标签的学习，所以需要更多的特征。往往这些特征需要我们根据不同的场景去人工的抽取，比如抽取人名的特征我们往往可能看看单词的第一个字是不是百家姓等。所以，标签*Y*与特征*X*是一对多的关系。如下图：



基于条件随机场的命名实体识别分三步进行：

1. 将试题库文本分为训练语料与不重叠测试语料，将训练语料进行标注转换后，利用*CRF*模型对转换后的语料进行训练，最终生成模型参数。
2. 利用分词技术对测试语料进行分词和词性标注，并利用上一步得到的*CRF*模型进行命名实体的识别。最终将词形和词性标注序列转换为本文定义的标注集序列。
3. 通过挖掘未识别实体样本的内部特点和上下文特征，设计了大量的人工规则，在上一步识别的基础上，进行二次识别。

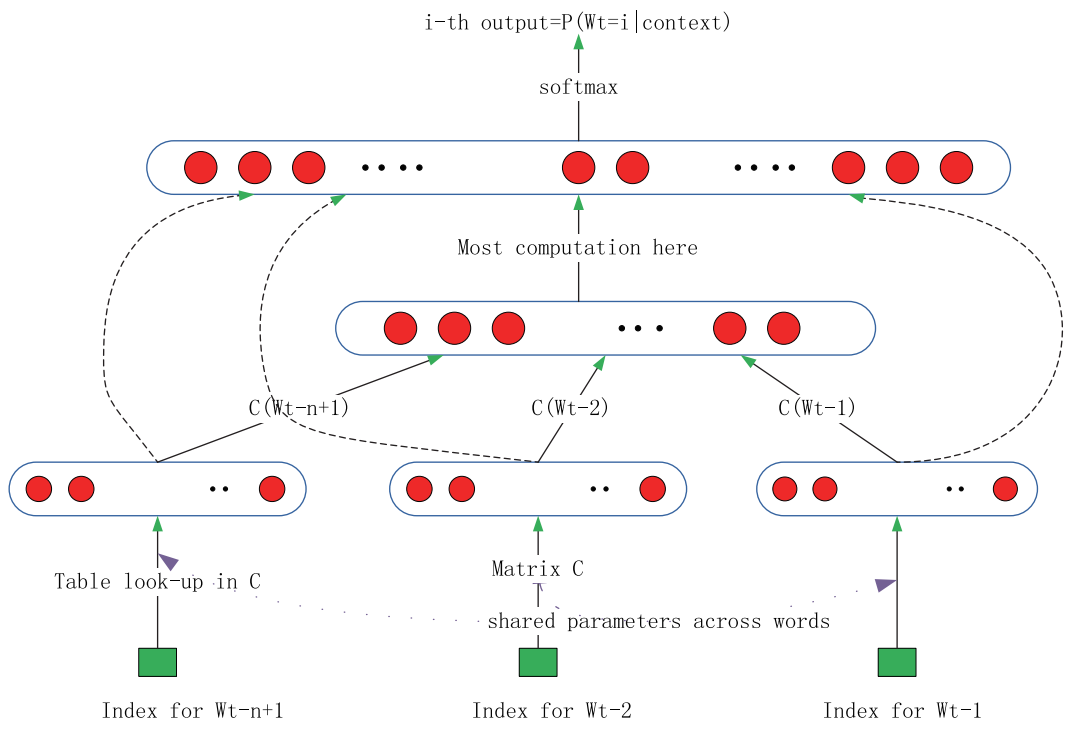


1.4 基于神经网络算法和聚类的干扰选项生成

利用命名实体识别选择出了题目的选项后，需要智能生成与正确选项相近的错误干扰选项。由一些属性相近的词组成的词可以看成是单个词语到语意一般概念的映射。我们将中文词语看成一系列独立词的“词袋模型”，这种模型将语言中词语之间的关系做了简化，仅仅考虑词语的统计特性；之后使用深度神经网络算法将词转化为*n* 维向量，它在传统三层神经网络算法的基础上做了延伸，将网络从三层扩展到多层；最后用*k-means* 算法计算对这些向量进行聚类，同一类中的词语具有类似的特征，选择与正确选项词语具有类似特征的词即可作为干扰选项。

词向量具有良好的语义特性，是表示词语特征的常用方式。词向量每一维的值代表一个具有一定的语义和语法上解释的特征。所以，可以将词向量的每一维称为一个词语特征。词向量具有多种形式，一个词向量是一个稠密、低维的实值向量。词向量的每一维表示词语的一个潜在特征，该特征捕获了有用的句法和语义特性。可见，词向量将词语的不同句法和语义特征分布到它的每一个维度去表示。

可以采用一个三层的神经网络来构建语言模型，如下图所示。最下方的就是前*n-1*个词，模型会根据这*n-1*个词预测下一个词*wt*。*C(w)*表示词*w*所对应的词向量，整个模型中使用的词向量，存储于矩阵*C*，*C*的维度为。其中表示词表的大小，*M*表示词向量的维度。*w*到*C(w)*的转化就是从矩阵中取出一行。



神经网络模型

网络的输入层（第一层）是将这*n-1*个向量首尾相拼接起来，形成一个维的向量，下面记为x。网络的隐藏层（第二层）直接使用计算得到，其中d为一个偏置项，初始化值随机。使用*tanh*作为激活函数。网络的第三层（输出层）节点用表示，一共有个节点。下一个词为*i*的未归一化*log*函数。最后使用*softmax*激活函数将输出值*y*归一化成概率。最终*y*的计算公式为：

式中*U*是隐藏层到输出层的参数，他是一个维的矩阵，整个模型的多数计算集中在*U*和隐藏层的矩阵乘法中。式中的矩阵*W*，包含了从输入层到输出层的直连边。最后使用随机梯度下降法优化模型。优化结束后，即可从输出*y*中获取目标词汇对应的词向量。

获取到词向量表之后，我们就可以计算不同词向量直接的余弦距离，作为不同词之间的“距离”。在这之后，使用 *K-means* 聚类算法将近似的词聚集到一起。

*K-means*算法接受输入量*k*，然后将*n* 种数据对象划分为 *k* 组聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”来进行计算的。算法基本步骤：

1) 从*n* 个数据对象任意选择*k* 个对象作为初始聚类中心;

2) 根据每个聚类对象的均值，计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分;

3) 重新计算每个（有变化）聚类的均值;

4) 计算标准测度函数，当满足一定条件，如函数收敛时， 则算法终止；如果条件不满足则回到步骤 2。

通过聚类算法，分析出与命名实体识别中识别出的题目正确选项属于同一类的词汇，作为题目正确选项的“同义词”，即题目干扰错误选项。

二 基于用户评价与机器学习的题目优化与增强

传统的学习系统中，题库的知识和题目一旦设定便几乎很少进行调整和更新，但是实际应用中，经常会遇到题目设置不合理的问题，例如有些题目超出范围、内容过时、题目选项不合理，甚至有些题目答案出错。采用这样的题目让用户学习的话，不仅起不到学习效果，还会打击用户的积极性。因此，用户对于题目的评价是极为重要的。如果能够将用户对于题目的评价收集起来，通过分析评价的具体内容，从而针对性地对题目进行调整和增强，则会提高系统的用户满意度，提升用户学习水平。

基于用户评价与机器学习的题目优化与增强算法将运用机器学习算法从题目属性挖掘与题目分类、题目维度拆分、题目难度调整等方面结合用户评价对题目进行优化和增强。

2.1 题目属性挖掘与题目分类

题库中的每一道题目由两部分组成，即题目属性指标和题目本身。题目属性指标有多项，反映有关题目的信息，这些信息将某一题目与其他题目区别开，同时标记着题目使用的历史和使用情况(如使用次数，被高分员工答对的次数和被低分员工答对的次数等，以及难度系数等)。

在设定好题目属性之后，可以根据题目的属性对题目的属性利用聚类算法进行聚类，得到不同的题目类别。在获得题目类别之后，可以通过对答题情况的分析获得用户不同题目类型的掌握情况，从而针对用户擅长的题目、类型、答题水平、知识盲区来推荐对应的题目。

2.2 题目维度拆分

题库并非是单一线索下的线性知识归集，相反却是极具多维性的，而且实际中不同维度之间可能是相互关联、相互作用的，例如对于党建类的题目而言，题目可能包含历史、时政、政经、哲学原理、实践教育等维度，而用户对于题目的答题情况往往暗含了其对于某个或者某些维度的掌握情况，因此有必要将题目维度进行拆分。

本项目提炼出知识的结构维度、群聚维度、关联维度，基于对以上三个维度的分析，从多维度融合的视角，构建以结构维度、群聚维度和关联维度为主干的题目维度框架。

（1）知识的结构维度。知识的结构是知识序化的基础，是知识系统化、学科化的保障，组织结构维度主要聚焦于领域知识的内部组织结构。换句话说在特定领域内，知识概念间的种属关系（父子、兄弟等）构成了具有一定中心性和层级性（偏序性）的组织结构。目前在揭示知识组织结构方面，最具有代表性的当属主题词表和领域本体。由于主题词表和领域本体内部结构严谨、逻辑严密，因此无论是用于资源聚合还是进行知识推理都取得了较为丰富的成果，并且基于概念的语义映射正逐步向多本体协同的方向发展。

（2）知识的群聚维度。知识的群聚性是客观知识世界中知识概念社会性的具体表现。知识世界与人类世界类似，知识概念间总是有远近亲疏之分，相似的知识、相近的知识之间总是具有更紧密的社会关系，从而形成不同程度的知识群簇。如果说结构维度中的“中心”或“层级”特征是学科分类体系的基础线索，那么这种知识的群聚维度则体现的是学科无关的知识聚类。这种知识聚类形成知识客观世界中的知识群落，而知识群落则通过各个知识节点之间的互惠性共同支撑某一项科学技术的发展。

（3）知识的关联维度。关联维度主要是从知识间关联关系的角度出发，重点观测知识间相互影响、相互扰动的状态、规律与模式。所谓“关联”关系既可以是知识单元间的关联关系，也可以是知识单元与知识群落（按照某一标准的知识单元集合）的关联关系，还可以是知识群落与知识群落的关联关系，甚至是资源间的关联关系、语义关联关系、数据关联关系等。目前这一维度下的主要研究方法，包括文献计量链接分析和关联数据等。

这三种维度事实上也是目前最能代表题目的三个发展方向和路径。通过这三种主要维度的相互融合交叉，能够更综合全面地而不是孤立地看待知识维度拆分的问题。并且三种维度绝不是相互隔绝相互独立的，而是相互影响相辅相成的，共同构造了知识的维度网络。

2.3 题目难度调整和知识增强

与传统根据主观经验判断题目难度这一常规做法不同，本项目将对用户原始答题数据进行挖掘，通过题目的答题情况进行题目的难度调整。

影响题目难度的因素有很多，比如考查的知识点的数量：题目中考察知识点的数目可以当做衡量题目难易度的因素，考察的知识点数量越大，题目就越难；反之，考察的知识点数量越少，题目就越容易；内容模块的涉及量：内容模块的涉及量也就是指知识点的分散度，所考察的知识点是否在一个内容模块上，或者说是否同属于一个学科。一般来说，内容模块的涉及量越大，就需要被试者的知识储备越多，那么题目越难；反之内容模块的涉及量越小，题目难易度就越小；创新性、内容情境的新颖度、内容的表达方式等等。除了题目本身属性的难易判别外，还有一些可以通过用户答题情况和答题习惯来获得，比如用户对于某道题目的答题时间长短、答错某道题的次数、某道题目用户整体的答题错误率，用户掌握情况不好的题目难度系数应当提升。

2.4 基于用户评价的题目优化

当用户对题目有疑义时，可以向系统提交对该题目的针对用户提交的评价内容。在获取到用户的评价之后，先对其进行关键词抽取，并将抽取的结果运用机器学习方法进行分类，从而参照分类结果对题目进行调整和优化。

在对评价内容进行分析时，主要对评价内容中所用词语进行分析，这里主要包括对文本内容的分析。对于文本内容的分析目前较为典型的做法是通过使用普适性的词语来界定评价内容，并对其及进行量化处理。

三 基于大数据的用户学习情况挖掘算法

数据挖掘（Data mining）一般是指从大量的数据中通过算法搜索隐藏于其中信息的过程。数据挖掘通常与计算机科学有关，并通过统计、在线分析处理、情报检索、机器学习、专家系统（依靠过去的经验法则）和模式识别等诸多方法来实现上述目标。

通过利用大数据技术对用户的答题情况进行数据挖掘和分析，可以得到用户擅长的题目类型、答题水平、知识盲区、知识雷达图等，以及用户在整体中的掌握水平，为用户制定符合自身情况和学习梯度的练习题目和知识路线。

具体实现方法有两种：一是通过分析用户曾经答题的题目属性、难度等信息，提取题目特征然后构建用户画像，随后将用户画像同题目特征进行匹配，得出的结果就是用户和题目之间的关联度，关联度越好，说明该题目用户掌握得越好，对其推荐地越低；另一种方式是若用户若用户u、v在相同的题目的掌握情况相近，那么用户u对题目i的掌握情况也非常有可能与v的掌握情况相似。

采取第二种实现方式中的k最近邻模型，简称kNN，是被最早应用且效果很好的模型。kNN的基本思想是通过与目标样本最相邻的k个样本的值加权平均估算目标样本的值。常见的两种推荐方法是基于用户的推荐和基于题目的推荐。

其中，基于用户的推荐和基于题目的推荐，相似性的计算就是个关键问题。通常的相似度计算有以下几种：

a)余弦相似度



b)皮尔森相似度



c)欧几里得距离



d)Tanimoto系数



类似的相似度计算还有对数似然度、曼哈顿距离等，不同的相似度计算有其优势也有其缺点，比如在处理稀疏性的能力上，在计算效率上，在不同数据的可靠性上。

获得相似度后我们就可以利用最近邻的方法进行打分预测，也就是通过个最相近的用户进行协同过滤运算，这里预测方式也有多种比如最常用的求平均：



通常经过试验比较有效的还有一种方法，利用用户本身的偏好和相似度作为权值相结合的方法：



上式中，引入了用户的平均这是假设不同用户对不同题目的答题情况通常性的评价，引入这个假设后会让预测结果更合理和平滑。利用上式我们就可以根据相似用户的信息对目标用户进行运算和推荐了。

评价协同过滤推荐算法的常用测度使用平均均方误差其定义为：



平均均方根误差：



平均绝对误差：



需要指出的是，在对用户某道题目的答题情况进行评价时，应当建立梯度式评价规则，对于不同用户的答题情况设置不同的评分值，比如答题时间长短、答错选项与题目的关联度。以客观题型为例：一个选择题有四个答案，其中一个是正确答案，但从出题者的角度看，另外三个选项并非毫无根据地随便加上去的，有一些错误选项离正确答案“距离”很近，用户选择错误选项的原因并非对此知识点一窍不通，而仅仅是因为考虑不周或大意，被具有迷惑性的选项所欺骗，在传统考试评价过程中，这一特点应当得到准确反映。相反，其中有个别选项确实与该知识点没有任何关联，而选择此选项基本可以判断被试用户对此知识点一窍不通。显然这两类被试者是不同的，在选取其对题目的分析评价内容时应该对其有区别对待。