二、确定矩阵中各元素的值。“词-上下文”共现矩阵根据其定义，里面各元 素的值应为词与对应的上下文的共现次数。然而直接使用原始共现次数作为矩 阵的值在大多数情况下效果并不好 [3]，因此研究人员提出了多种加权和平滑方 法，最常用的有 tf-idf、PMI 和直接取 log。

最常用的分解技术包括奇异值分解（SVD）、非负矩阵分解（NMF）、典型关联分析（Canonical Correlation Analysis，CCA）[22, 23]、Hellinger PCA（HPCA）[59]。

基于矩阵的分布表示在这些步骤的基础上，衍生出了若干不同方法，如经典的LSA [56] 就是使用“词-文档”矩阵，tf-idf 作为矩阵元素的值，并使用SVD分解，得到词的低维向量表示。在这类方法中，最新的为GloVe 模型[90]，下文简单介绍这一模型。

在由于需求的变更 或者 错误引起的变更中，可能存在的情况是：特定词（关键词）改变，引起了下级要素内容的改变，但依然存在未改变部分，此部分的相似度仍然是比较高的，例如：

因此，考虑变更中的此情况，考虑改进此处相似度算法（扩展查询没找到办法）

主要思路:相似排除法？1.对于改变了的部分，

（思考：可追溯性可能是没有降低的，主要是计算变化了的部分的相似度

对比前后两次的相似度的改变程度？ 交叉比较相关的两个构件的相似度，若区别不大，说明可能有问题，若 1.a 1.b相似，2.a 2.b 相似，1.a 2.b相似度要低，1.b 2.a相似度也低，说明正确的可能性高）

在技术上怎么处理？扩展查询可以吗？怎么确定是哪些词？

去除原关键词？计算df-idf,将原权重大的关键词（没法找）去除（通过原基线计算原关键词，通过新的） 有问题

提出一种 对变更后的数据进行相似度测算的方法

1. 测试数据去哪里找
2. 特点： 使用 向量减去的方法（n-0,0-n 这种情况下加大权重，这是一种拓展），

增，（直接）删，（直接） +改 面目全非，可以

方法确定

另外，设想由于Do178c所涉及的领域是有限的，具体可根据不同领域分别生成语料库。