基于向量空间模型的题库相似度检查算法①

汪忠国 吴 敏 (中国科学技术大学 现代教育技术中心 安徽 合肥 230026)

摘 要: 随着题库系统的广泛应用和题库中试题数量的日益增大，如何避免试题重复，成为研究的重要问题。利用向量空间模型,首先通过TF-IDF公式得到试题的文本权重向量,再通过余弦理论计算试题相似度，并与设定的相似度阈值比较，得到相似度检查结果。在现有题库的基础上进行的实验结果显示，算法计算出的试题相似度的准确率与专家人工判别相比达到94%。算法取得了较好的结果。

关键词: 向量空间模型；相似度检查；单文本词汇频率；逆文档频率；余弦理论

Similarity Checking Algorithm in Item Bank Based on Vector Space Model

WANG Zhong-Guo, WU Min

(Center of Modern Educational Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: With a wide use of item bank system and the increment of items in item bank system, how to avoid duplicate items becomes an important research topic. This paper first gets text with vectors with TF-IDF formula through the algorithm based on vector space mode(VSM) theory. Then, it gets the similarity of items by using cosine theory, which is used for the comparison with the threshold value initialized to get similarity checking resulting. Based on the existing item bank system, the experiment with this algorithm shows that the exact rate of 94% is gained, which is a good result compared with expert checking.

Keywords: vector space model; similarity checking; TF; IDF; cosine theory

随着计算机在教学领域的应用和发展，试题库的编制和应用也越来越显示出其重要性。题库系统的核心问题是优质的试题、合理的试题库结构[1] 。但是，随着题库中试题量的日益增大，依赖专家出题时人工判别试题是否相似或重复，难度较大，因此，如何建立一个合适的能够识别相似试题的算法成了建立优质题库的关键。

本文利用向量空间模型的理论，先使用 TF-IDF 公式将待检查试题的文本向量化，再通过余弦理论计算待检查试题的文本向量和题库中现有试题的文本向量的余弦值得到试题相似度，之后通过与设定的相似度区别阈值比较，得到试题是否重复的结果。在结果显示试题不相似的情况下，题库系统直接保存待检查试题入库，否则，题库系统对用户进行重新出题或更换试题等提示。

1 算法相关概念及模型介绍

1.1 试题相似度

试题相似度是指两道试题在元数据和内容上的相似程度．在[0，1]之间取一实数值，值越大表明两道试题越相似，当取值为 1 时，表明两道试题完全相同；值越小则表明两道试题相似度越低，当取值为 0 时，表明两道试题完全不同[2]。

1.2 向量空间模型

向量空间模型[3]是信息检索领域进行语句相似度比较的常用模型。在现代信息检索这本书中的定义为：

对于待检查文本 q 中的每一个单词，使用$(W_{i,q})$代表此文本中第 i 个单词的权重，同样使用$(W_{i,j})$代表已有文本 j 中第 i 个单词的权重，从而可以用

$\vec{q}=(W_{1,q},W_{2,q},\cdots,W_{t,q})$和$\vec{d_j}=(W_{1,j},W_{2,j},\cdots,W_{t,j})$

表示待检查文本 q 和已有文本 j 的单词权重向量。

① 收稿时间:2009-06-30

在得到试题的文本向量之后，通过余弦理论计算 $\vec{q}$ 和 $\vec{d_j}$ 这两个向量的余弦相似度从而可以得到待检查文本 q 和已有文本 j 之间的相似度。余弦相似度的计算公式如下：

$$similarity = \cos(\theta) = \frac{\vec{q}.\vec{dj}}{\|\vec{q}\| \times \|\vec{dj}\|}$$

$$= \frac{\sum_{i=1}^{t} Wi,j \times Wi,q}{\sqrt{\sum_{i=1}^{t} W_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} W_{i,q}^2}} \quad (1)$$

其中 $\|\vec{q}\|$，$\|\vec{d_j}\|$ 表示向量的模。由于 $W_{i,j}$，$W_{i,q}$ 均大于等于 0，所以(1)式的值是一个 0 到 1 的值，其中，0 表示两个文本的相似度为零，1 表示两个文本完全相似。

1.3 TF-IDF 公式

向量空间模型中单词的权重 W 使用 TF-IDF[4]公式计算。

TF(Term Frequency)为单词频率，表示一个单词与某个文档的相关性。某单词的 TF 可以通过这个单词在文档中出现的次数除以该文档中所有单词出现的总次数得到。简单考虑，也可以使用正则化处理，即 TF 等于单词在文档中出现的次数除以文档中出现频率最高的单词的次数。

IDF(inverse document frequency)为逆向文档频率，是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到。

对于系统中现有的文档，令 N 表示文档总数，n_i 表示包含单词 k_i 的文档数目，$freq_{i,j}$ 表示文档 d_j 中单词 k_i 出现的次数，文档 d_j 中单词 k_i 的正则化单词频率(TF)用 $f_{i,j}$ 表示，其中 $f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}}$；相应的，单词 k_i 的逆文档频率(IDF)用 idf_i 表示，其中 $idf_i = \log\frac{N}{n_i}$，则单词 k_i 的权重 W_i 可以用如下公式计算：

$$W_i = f_{i,j} \times \log\frac{N}{n_i} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \times \log\frac{N}{n_i} \quad (2)$$

对于待检查文本，现代信息检索一书中定义其中的单词的权重为：

$$W_{i,q} = (0.5 + \frac{0.5 \times freq_{i,q}}{\max_l freq_{l,q}}) \times \log\frac{N}{n_i} \quad (3)$$

上两式中，Salton and Buckley 指出，IDF 因子的使用原因是出现在许多文档中的单词对于辨别相关性文档不十分有用[5]。本文算法中在剔除低频词之后，文档中的单词均对相似度检查有着重要作用，从而，计算已有文本中单词的权重时，(2)式可以简化为

$$W_i = \frac{freq_{i,j}}{\max_l freq_{l,j}} \quad (4)$$

计算待检查文本中的单词的权重时，(3)式可以变更为

$$W_{i,q} = (0.5 + \frac{0.5 \times freq_{i,q}}{\max_l freq_{l,q}}) \quad (5)$$

2 相似度检查算法设计

2.1 算法中的相关定义

相似度检查算法中涉及到的数据库中的表主要有以下两个：

a) ItemFreWord 表：用于记录试题的 ID，此试题中的高频词汇及其出现次数。

表结构如图 1 所示：

名称	数据类型	大小
Item_id	varchar	32
FrequenceWord	varchar	50
WordCount	int	4

图 1　ItemFreWord 表的结构

b) FreWord 表：用于记录类似于 a，ok，this，that 等英语中频率最高的一些词汇，用以剔除这些词汇对相似度检查的影响。

FreWord 表中只有一项，结构如图 2 下：

名称	数据类型	大小
FrequenceWord	varchar	50

图 2　FreWord 表的结构

相似度检查算法中，主要用到的自定义的类 TwordItem 定义如图 3 所示：

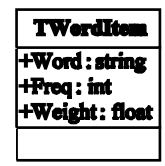

图 3　TwordItem 类的结构示意

其中，Word 表示单词内容，Freq 表示在单词某

个试题文本中出现的次数，Weight 表示单词在 TF-IDF 公式中计算出的权重。

2.2 算法实现步骤

经过中科大外语教师和外语专家的讨论研究确定，进行相似度检查，在相似度阈值设为 0.9 的情况下，如果算法计算的相似度小于阈值，则可以认为没有找到相似的试题，题库将执行提交操作，保存新出试题到题库；如果计算的相似度大于阈值 0.9，则可以认为新出试题与某个已有试题相似，题库将提示出题教师进行[重新编辑]、[放弃此题]、[替换此题]等进行选择。

在上述总体原则下，相似度检查算法定义 Item 为待检查试题，Q 为 Item 的文本，QueryFreqWordList，QueryFreqWordList*，CompareItemIdList，CompareItemFreqWordList，similarList 等的数据类型均为集合类型，用于存放高频词汇或者相似试题。

相似度检查的算法流程如下所示：

a 通过正则表达式，获得 Q 中所有的单词及其出现频率，存入 TwordItem 结构中，并将每个 TwordItem 存入 QueryFreqWordList；

b 从 QueryFreqWordList 中剔除数据库中常见单词表 FreqWord 表中的单词，得到 QueryFreqWordList*；

c 以 QueryFreqWordList*中的单词逐个检查数据库的 ItemFreWord 表，将包含相同单词的试题 ID 取出，存入 CompareItemIdList，得到将要与 Item 进行比较的试题 ID 列表；

d 对 CompareItemIdList 中的 itemID 从数据库 ItemFreWord 的表中提取此试题包含的高频词；

e 将提取出的每个高频词及其出现次数存入 TwordItem 结构中，其中 TwordItem 的初始 weight 均为 0，将此 TwordItem 加入 CompareItemFreqWordList；

f 对 QueryFreqWordList*中和 CompareItemFreqWordList 的单词使用向量空间模型计算相似度，步骤如下：

f1 获得 QueryFreqWordList*中最高频单词的频率；

f2 获得每个单词的频率，并根据公式(5)计算待检查试题的 QueryFreqWordList*中每个单词的 TF-IDF 权重(weight)；

f3 获得 CompareItemFreqWordList 中最高频单词的频率；

f4 获得 CompareItemFreqWordList 中每个单词的频率，并根据公式(4)计算每个单词的 TF-IDF 权重(weight)；

f5 将 f2，f3 步骤中得到的文本向量，根据公式(1)计算这两道试题的余弦相似度；

f6 如果 f5 中计算的相似度大于设定的阈值 0.9，则将当前的 itemID 计入 similarList 中；

g 重复步骤 d，直到 CompareItemIdList 中的所有 ItemId 均提取并进行了上述 f 下的各个分步骤；

h 如果 similarList 中的 item 数目大于 0，表示题库中存在相似试题，则提示出题教师进行[重新编辑]、[放弃此题]、[替换此题]的选择；否则直接将所出的试题存入数据库。

3 算法实验结果

表 1 50 道试题的相似度检查结果

余弦相似值	0.996	0.925	0.977	0.951	0.961
相似度判定	相似	相似	相似	相似	相似
余弦相似值	0.908	0.826	0.962	0.902	0.928
相似度判定	相似	不相似	相似	相似	相似
余弦相似值	0.946	0.963	0.96	0.994	0.955
相似度判定	相似	相似	相似	相似	相似
余弦相似值	0.938	0.903	0.929	0.944	0.963
相似度判定	相似	相似	相似	相似	相似
余弦相似值	0.962	0.939	0.819	0.993	0.905
相似度判定	相似	相似	不相似	相似	相似
余弦相似值	0.934	0.969	0.962	0.905	0.918
相似度判定	相似	相似	相似	相似	相似
余弦相似值	0.985	0.906	0.939	0.885	0.92
相似度判定	相似	相似	相似	不相似	相似
余弦相似值	0.959	0.932	0.909	0.938	0.93
相似度判定	相似	相似	相似	相似	相似
余弦相似值	0.912	0.98	0.912	0.905	0.975
相似度判定	相似	相似	相似	相似	相似
余弦相似值	0.963	0.927	0.924	0.915	0.971
相似度判定	相似	相似	相似	相似	相似

算法的实验题库是实验室和上海外语出版社联合开发的分级题库中，目前共有 7313 题。实验首先由英语专家随机取出 50 题，并由将此 50 题的内容进行修改，并确保修改后的试题与原试题比较英语专家是认为相似的，即通过英语专家的工作，得到专家认为的与题库中试题相似的 50 道试题。之后，将此 50 道试题输入题库，并运行上述算法进行相似度检查，得到的检查结果如表 1 所示。

从上表中可以看出，50 道专家判定相似的试题中，算法检查出 47 道题是相似的，即本文设计的算法的检查结果与专家判定相比，准确率高达 94%，效果十分良好。

4 结论

在出题时，对试题进行相似度检测，可以使新出试题和已有试题避免重复，从而保持题库的高质量。使用基于向量空间模型的相似度检查算法，通过 TF-IDF 公式得到试题的文本权重向量，之后通过余弦理论计算出试题相似度，并与专家认为的相似度阈值进行比较，得到相似度检查的结果。算法的实验结果表明，使用基于向量空间模型的相似度检查算法对试题进行相似度检查的准确率高，算法效果良好，同时算法流程简单，在实际应用中易于实现。

参考文献

1 王宇颖,陈振,苏小红.自动组卷中试题去重技术研究.哈尔滨工业大学学报, 2009,41(1):85－88.

2 世平,樊孝忠.基于多示例学习的题库重复性检测研究.北京理工大学学报, 2005,25(12):1071－1074.

3 Salton G, Lesk ME. Computer evaluation of indexing and text processing. Journal of the ACM,1968,15(1):8－36.

4 Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. Addison Wesley, 1999.38－42.

5 Salton G, Buckley C. Term-weighting approaches in automatic retrieval. Information Processing & Management, 1988,24(5):513－523.