

基于概率潜在语义分析模型的自动答案选择

张成¹, 曲明成², 倪宁³, 仇光², 卜佳俊²

(1. 中国残联信息中心, 北京 100034; 2. 浙江大学计算机科学与技术学院, 杭州 310027;

3. 浙江商业职业技术学院信息技术学院, 杭州 310053)

摘要: 问答社区中候选答案过多会增加提问用户选择最佳答案的负担。为此, 提出一种基于概率潜在语义分析(PLSA)模型的自动答案选择方法。在主题建模思想的基础上, 利用问答社区中的用户资料, 以 PLSA 模型表达问答社区中的用户兴趣分布, 依据答案和问题之间的主题匹配度对候选答案进行排序。实验结果表明, 该方法可有效挖掘用户兴趣, 提高答案选择的准确率。

关键词: 答案选择; 问答社区; 概率潜在语义分析; 主题建模

Automatic Answer Selection Based on Probabilistic Latent Semantic Analysis Model

ZHANG Cheng¹, QU Ming-cheng², NI Ning³, QIU Guang², BU Jia-jun²

(1. China Disabled Persons' Federation Information Center, Beijing 100034, China;

2. College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China;

3. College of Information Technology, Zhejiang Vocational College of Commerce, Hangzhou 310053, China)

【Abstract】 A novel answer selection method based on topic modeling techniques is proposed to mitigate the issue of question asker's burden of selecting the best answer stemming from too many candidate answers in question answering communities. Aiming at the problem, this paper presents an automatic answer selection based on Probabilistic Latent Semantic Analysis(PLSA) in question answering communities, and accordingly rank candidate answers based on similarity of interest between answers and questions. Experimental results show that the method can effectively excavation user interest and improve the accuracy of answer selection.

【Key words】 answer selection; question & answering community; Probabilistic Latent Semantic Analysis(PLSA); topic modeling

DOI: 10.3969/j.issn.1000-3428.2011.14.022

1 概述

随着互联网的迅猛发展, 交互式问答社区已成为信息获取和知识分享的一个重要平台。然而, 随着问答社区参与用户的日益增多, 候选答案数目增长迅速, 导致提问用户选择最佳答案的负担加大。因此, 对候选答案的自动排序成为问答社区的一个迫切需要。本文通过对问答社区用户交互特征的研究, 提出一种基于主题建模思想的答案排序选择方法。

2 相关工作

近年来, 问答社区逐渐成为互联网数据挖掘领域的研究热点。文献[1]采用分类框架整合问答社区中的各类文本信息, 研究问答社区中的文本质量预测和用户满意度预测问题。文献[2]采用翻译模型学习词语间的语义相似度, 并以此寻找相似问题。文献[3]研究了问答社区中的用户链接结构, 并提出用 HITS 算法预测用户权威度的方法。本文将研究问答社区中的答案自动排序选择方法。基于统计的主题建模广泛地应用于文档集的主题挖掘任务^[4-5]。文献[5]采用概率潜在语义分析(Probabilistic Latent Semantic Analysis, PLSA)方法, 构建一个中文信息检索框架。文献[6]采用 PLSA 方法实现问答系统中的问题推荐功能。本文将采用 PLSA 方法挖掘用户兴趣, 辅助问答系统中的答案选择。

3 基于概率潜在语义分析的答案选择

由于用户根据自己的兴趣主题选择相应的问题进行回答, 因此问题的主题与回答用户的兴趣需要有较高的相关度。同时, 候选答案自身的主题与问题主题应保持一致。综合以

上考虑, 本文利用问题主题与候选答案主题及问题主题与用户兴趣的相似度对候选答案进行综合排序。

3.1 用户兴趣建模

在一个典型的问答周期中, 由于用户在回答问题时已隐式地选择了主题, 因此用户所回答的问题所属的主题能反映用户的兴趣。尽管目前的问答社区已对每个问题提供了显式的主题类别信息, 但这些显式的主题存在粒度过粗、类别局限等问题, 不适合对用户的细粒度进行兴趣建模。对用户进行聚类只能捕捉到用户最主要的兴趣主题, 不能有效描述用户的所有兴趣。本文采用的概率潜在语义分析模型^[4]可自动挖掘文档集中的潜在主题, 无需对主题类别进行预先定义, 同时能全面描述用户兴趣。在对用户 u 进行建模时, 以 PLSA 中的潜在变量 $z \in \{z_1, z_2, \dots, z_k\}$ 表示 u 所回答的问题集的潜在主题。用户 u 回答问题 q 的隐含过程是先选择主题 z , 再根据主题选择问题 q 。因此, u 和 q 的联合概率可以表示为:

$$\Pr(u, q) = \sum_z \Pr(q | z) \Pr(z | u) \Pr(u) \quad (1)$$

其中, $u \in u_1, u_2, \dots, u_m$ 表示问题集中的用户; $q \in \{q_1, q_2, \dots, q_n\}$ 表示问题集中的问题。

然而, 在真实的问答社区中, 每个用户所回答问题的数

基金项目: 国家科技支撑计划基金资助项目(2008BAH26B00)

作者简介: 张成(1970—), 男, 硕士, 主研方向: 数据挖掘, 语义分析; 曲明成, 硕士; 倪宁, 副教授; 仇光, 博士; 卜佳俊, 教授

收稿日期: 2011-02-14 **E-mail:** bjj@zju.edu.cn

目通常很少。为解决稀疏性问题, 本文采用<用户, 词语>对的共现信息进行主题建模, 其中共现数据为用户在回答某问题时输入的词语。类似于式(1), 用户和词语的联合概率表示为:

$$\Pr(u, w) = \sum_z \Pr(w|z) \Pr(z|u) \Pr(u) \quad (2)$$

其中, $w \in \{w_1, w_2, \dots, w_l\}$ 表示用户 u 在回答问题集问题时所输的词语。

在 PLSA 模型中, 参数包括 $\Pr(w|z)$ 、 $\Pr(z|u)$ 、 $\Pr(u)$ 。本文使用期望最大化(Expectation Maximization, EM)方法进行参数估计:

(1)期望步骤

$$\Pr(z|u, w) = \frac{\Pr(w|z) \Pr(z|u) \Pr(u)}{\sum_{z'} \Pr(w|z') \Pr(z'|u) \Pr(u)} \quad (3)$$

(2)最大化步骤

$$\Pr(z|u) \propto \sum_w c(u, w) \Pr(z|u, w) \quad (4)$$

$$\Pr(w|z) \propto \sum_u c(u, w) \Pr(z|u, w) \quad (5)$$

其中, $\Pr(u)$ 的值为用户 u 所回答的词频总和在所有用户的词频总和中所占的比例, 计算如下:

$$\Pr(u) = \frac{\sum_w c(u, w)}{\sum_{u', w} c(u', w)} \quad (6)$$

$\Pr(w|z)$ 和 $\Pr(z|u)$ 的初始值为随机数, 通过迭代期望步骤和最大化步骤计算问题集合对数似然度。

3.2 答案选择

本文在计算候选答案的排序得分时需要考虑用户兴趣和答案主题与问题主题的相似性, 计算如下:

$$Score_{q,a} = \alpha \Pr(a|q) + (1-\alpha) \Pr(u|q) \quad (7)$$

其中, $Score_{q,a}$ 表示问题 q 中候选答案 a 的最终得分, 本文将选取得分最高的候选答案作为最佳答案; $\Pr(a|q)$ 表示给定问题 q , 候选答案 a 的后验概率, 即 a 作为 q 最佳答案的概率; $\Pr(u|q)$ 表示给定问题 q , 提供该候选答案的用户 u 的后验概率, 即 u 为 q 提供最佳答案的概率。最终得分是 2 个概率值的线性加权之和。加权系数 α 的取值在 $[0, 1]$ 区间, 其中, 当 $\alpha=0$ 时, 得分即为用户 u 提供最佳答案的概率; 当 $\alpha=1$ 时, 得分为候选答案 a 作为最佳答案的概率。根据贝叶斯定律, 式(7)可转化为 $\Pr(a, q)$ 和 $\Pr(u, q)$ 经归一化后的线性加权和:

$$Score_{q,a} = \alpha \Pr(a, q) + (1-\alpha) \Pr(u, q) \quad (8)$$

最终, 将 q 的候选答案按 $Score_{q,a}$ 降序排列, 取首个答案为最佳答案。下文将阐述 $\Pr(u, q)$ 和 $\Pr(a, q)$ 的计算方法。

3.2.1 $\Pr(u, q)$ 计算

$\Pr(u, q)$ 的值为问题 q 中所包含的词语概率的乘积, 再将问题 q 的长度作归一化:

$$\Pr(u, q) = \left(\prod_i \Pr(u, w_i) \right)^{\frac{1}{|q|}} \quad (9)$$

其中, w_i 为问题 q 中的词语; $|q|$ 是 q 的词语总数。

3.2.2 $\Pr(a, q)$ 计算

问答社区中的候选答案在不断动态更新, 但系统不可能在收到新的答案时重新计算 PLSA 模型。本文采用 fold-in 方法把候选答案的文本内容调入到已有的 PLSA 模型:

$$\Pr(z|a, w) = \frac{\Pr(w|z) \Pr(z|a) \Pr(a)}{\sum_{z'} \Pr(w|z') \Pr(z'|a) \Pr(a)} \quad (10)$$

$$\Pr(z|a) \propto \sum_w c(a, w) \Pr(z|a, w) \quad (11)$$

其中, $\Pr(z|a)$ 表示给定答案 a 中主题 z 的概率, 其初始值为随机数, 并通过以式(10)、式(11)反复迭代得出最后结果; $c(a, w)$ 表示答案 a 中词语 w 的出现频率; $\Pr(a)$ 表示答案 a 的先验概率, 计算如下:

$$\Pr(a) = \frac{\sum_w c(a, w)}{\sum_{a', w} c(a', w)} \quad (12)$$

答案和词语的联合概率可表示为:

$$\Pr(a, w) = \sum_z \Pr(w|z) \Pr(z|a) \Pr(a) \quad (13)$$

将问题 q 中包含的词语概率累乘并进行归一化得到:

$$\Pr(a, q) = \left(\prod_i \Pr(a, w_i) \right)^{\frac{1}{|q|}} \quad (14)$$

4 实验与结果

4.1 实验准备

本文从 Yahoo! Answers 问答社区中抽取 3 个类别的问题作为实验数据集, 包括天文学(Astronomy)、温室效应(Global Warming)、哲学(Philosophy), 并剔除仅包含一个答案的问题。在数据集中每个问题的最佳答案都已标注。表 1 为数据集的统计信息。在每个类别中, 选取 85% 的问题作为 PLSA 模型的训练数据, 剩余 15% 作为测试数据。

表 1 Yahoo! Answers 数据集

类别	问题数	答案数	用户数
Astronomy	8 920	49 297	16 391
Global Warming	8 330	82 788	22 015
Philosophy	9 477	84 953	22 822

本文对数据集进行以下预处理:

- (1)剔除停用词;
- (2)使用 Porter Stemmer(<http://tartarus.org/~martin/PorterStemmer/>)进行词根还原;
- (3)剔除文档频率(document frequency)小于 3 的词语;
- (4)剔除提出及回答问题数目小于 3 的用户。

实验结果以最佳答案再根据 $Score_{q,a}$ 排序列表中的位置进行衡量:

$$accuracy = \frac{|R| - R_B}{|R| - 1} \quad (15)$$

其中, $|R|$ 是答案排序列表中包含的答案个数; R_B 是最佳答案在答案排序列表中的位置。

作为对比, 本文实现了余弦相似度(Cosine Similarity)算法。余弦相似度方法通过比较问题向量和答案向量之间的相似度选择答案, 向量的值采用 tf.idf 权重:

$$s(q, a) = \frac{\sum_w tf.idf_{q,w} tf.idf_{a,w}}{\sqrt{\sum_w tf.idf_{q,w}^2} \sqrt{\sum_w tf.idf_{a,w}^2}} \quad (16)$$

其中, $tf.idf_{q,w}$ 和 $tf.idf_{a,w}$ 分别代表问题 q 和答案 a 中词语 w 的 $tf.idf$ 权重。

4.2 实验结果

在本文实验中, 首先对 $\Pr(a, q)$ 和 $\Pr(u, q)$ 权重系数的取值进行分析。本文设定潜在变量的个数 $k=100$, 从 0~1 调整权重系数。图 1 为在 3 个类别的数据集中的实验结果。可以看出, 当权重系数分别为 0.95、0.25 和 0.20 时, 准确率分别达到该类别的最优结果 0.766、0.746 和 0.765; 同时, 将 $\Pr(a, q)$ 和 $\Pr(u, q)$ 线性加权所得的结果好于单独采用 $\Pr(a, q)$ 或 $\Pr(u, q)$ 的结果, 验证了本文同时考虑问题与答案及用户相似度的有效性。

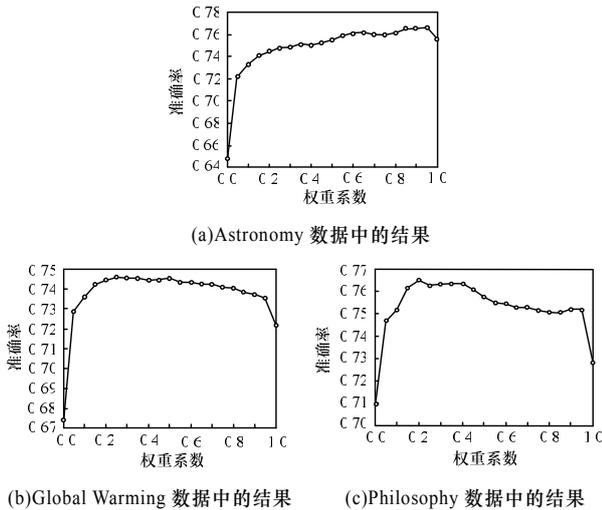


图1 本文方法实验结果

表2为本文方法与余弦相似度方法的实验结果。本文方法设定潜在变量数 $k=100$ 。可以看出，本文方法的结果在3个测试数据集中均好于余弦相似度。实验结果表明，PLSA模型可有效地挖掘用户兴趣，提高答案选择的准确率。

表2 2种方法的准确率比较

类别	余弦相似度方法	本文方法
Astronomy	0.693	0.766(权重系数=0.95)
Global Warming	0.679	0.746(权重系数=0.25)
Philosophy	0.713	0.765(权重系数=0.20)

5 结束语

本文提出一种基于概率潜在语义分析模型的答案选择方

编辑 陆燕菲

(上接第69页)

```

FSE执行结果显示
打开文件  预编译
0.  []
1.  [(A, 1)]
2.  []
3.  [(A, 3)]
4.  [(A, Y3=2->3◆1)]
5.  []
6.  [(B, 5)]
7.  []
8.  [(C, 7)]
9.  [(B, 5), (C, W3=6->7◆C1)]
10. [(A, Y3=2->3◆1), (C, Y3=4-> (C, W3=6->7◆C1)◆C1)]
11. [(A, Y3=2->3◆1), (C, Y3=4-> (C, W3=6->7◆C1)◆C1), (B, Y3=4->6◆B1)]

```

图3 例1的符号执行结果

4.2 符号执行算法的时间复杂度

传统的符号执行算法会遍历程序流程中的每一条路径。如图1所示，以代码行#define C 1;的可执行路径为例，在此代码行前包含了2个#if预处理指令合计4条路径，路径(a,d)是一条可能路径。在传统的符号执行方法中，路径条件的最终大小是与#if预处理指令的个数呈指数关系的。这种情况下的计算复杂度将不仅仅是 $O(2^n)$ 甚至可能是 $\Theta(2^n)$ ， n 指的是连续的#if预处理指令的个数。

本文所用到的符号预测算法没有涉及到回溯或者多路径的情况。如例1所示，在代码行7之前已有2个#if预处理指令，但没有引起预测时间或者预处理变量的条件值大小以指数方式增加，预处理变量的条件值大小的增长都是近似线性的。代码行2处的#if预处理指令并没有影响预处理变量C的c-value，因此，这类独立的预处理指令是不会增加代码行8的预测时间的。本符号执行算法的时间复杂度近似为 $O(n)$ 。相比传统的符号执行算法，其时间复杂度是大大降低了的，这同样适用于具有大量复杂的条件编译代码的预处理分析。

法。在问答社区中的用户兴趣通过PLSA模型的主题建模进行挖掘针对动态实时更新的答案数据，本文将答案数据调入已训练的PLSA模型中，以满足问答社区动态多变的应用环境。实验结果表明，本文方法可以有效挖掘用户兴趣，准确选择最佳答案。

参考文献

- [1] Agichtein E, Castillo C, Donato D, et al. Finding High-quality Content in Social Media[C]//Proc. of WSDM'08. [S. l.]: ACM Press, 2008: 183-194.
- [2] Jeon J, Croft W B, Lee J H. Finding Similar Questions in Large Question and Answer Archives[C]//Proc. of the 14th ACM International Conference on Information and Knowledge Management. [S. l.]: ACM Press, 2005: 84-90.
- [3] Jurczyk P, Agichtein E. Discovering Authorities in Question Answer Communities by Using Link Analysis[C]//Proc. of the 16th ACM Conference on Information and Knowledge Management. [S. l.]: ACM Press, 2007: 919-922.
- [4] Hofmann T. Probabilistic Latent Semantic Indexing[C]//Proc. of the 15th Conference on Uncertainty in Artificial Intelligence. [S. l.]: ACM Press, 1999: 50-57.
- [5] 罗景, 涂新辉. 基于概率潜在语义分析的中文信息检索[J]. 计算机工程, 2008, 34(2): 199-201.
- [6] Qu Mingcheng, Qiu Guang. Probabilistic Question Recommendation for Question Answering Communities[C]//Proc. of the 18th International Conference on World Wide Web. [S. l.]: ACM Press, 2009: 1229-1230.

5 结束语

通过本文研究，利用条件值可以有效地对C/C++预处理过程进行符号预测。

符号预测算法尚有一些不足：(1)在预处理过程中由于存在头文件的相互包含可能会出现迭代的情况，如头文件A.h中包含代码#include "B.h"，B.h中包含代码#include "A.h"，那么在进行符号预测时可能会出现#if预处理指令重复包含的情况，这在现阶段的算法设计中仍然没有解决。(2)本文的符号执行算法仍无法对于复杂宏命令进行处理，而宏命令的使用在头文件中是非常普遍的，从而导致本系统应用具有一定局限性。需要就以上2点对算法做进一步的改进与优化。

参考文献

- [1] 郑人杰. 计算机软件测试技术[M]. 北京: 清华大学出版社, 1992.
- [2] 杨宇, 张健. 程序静态分析技术与工具[J]. 计算科学, 2004, 31(2): 171-174.
- [3] 付剑平, 陆明燕. 软件测试性度量框架研究[J]. 计算机工程, 2009, 35(14): 60-62.
- [4] Hu Ying, Merlo E. C/C++ Conditional Compilation Analysis Using Symbolic Execution[C]//Proc. of the International Conference on Software Maintenance. Washington D. C., USA: IEEE Computer Society, 2000.
- [5] 尚卫东. C++静态预处理技术及其支持工具的研究与实现[D]. 北京: 北京航空航天大学, 2004.

编辑 顾逸斐

