文章编号:1000-5862(2013)02-0125-05

基于特征融合的社区问答问句相似度计算

杨海天,王 健,林鸿飞*
(大连理工大学信息检索研究室,辽宁 大连 116024)

摘要:提出了一种基于特征融合的问句匹配框架来解决问句相似度检测方法,利用答案特征、词序特征、统计特征和语义特征相结合来解决问句相似度计算问题. 在 Yahoo! Answers 上抽取的真实标注数据集上进行实验,实验结果表明:该方法在性能上得到了较好的结果.

关键词:问句相似度;社区问答;相似度计算;特征融合
中图分类号:TP 391 **文献标志码**:A

0 引言

近年来,社区问答系统已经成为在线寻求帮助信息的有效方法,人们可以在社区问答系统中提出自己的问题,并由其他用户回答. 由于任何人都可以在上面提问和回答,Yahoo! Answers 等成立几年来已经积累了大量的问答对. 如何有效地利用这些已有的问答对来回答用户提出的问题已成为研究的焦点. 由于自然语言中存在大量的同义词、语义特性和丰富的句法特征,所以从 CQA 系统中找到相似的问句并不是一件易事.

针对上述问题,本文提出了一种将答案特征、词序特征、统计特征和语义特征进行融合的方法,目的在于更加充分地挖掘句子之间的关系,从而有效地解决句子相似度计算问题.

1 相关工作

目前关于问句相似性计算的相关研究主要有 R. D. Burke 等[1] 和 V. Jijkoun 等[2] 利用向量空间模型计算查询问句向量和候选问句向量的夹角余弦;Cao Xin 等[3] 和 Duan Huizhong 等[4] 提出将语言模型应用到社区问答系统的问句检索中;Xue Xiaobin 等[5] 提出了基于翻译模型的问答系统检索模型. 以上这些方法以特征向量为处理对象,难以表示结构化的特征,存在数据稀疏问题. 文献[6] 提出一种统计信息和语义信息(基于 WordNet) 相结合的方法来计算 FAQ 中问句的相似度,该方法较好地解决了数据稀疏问题,并把语义信息引入到句子相似度计算,取得了较好的效果. 但文献[6] 提出的方法只利用了词表面的统计信息,并没有对不同的词表示相同的意思和不同的词在一个句子中所属的地位不同等问题进行研究. M. Collins 等[7] 提出了一种树核方法,通过计算 2 棵句法树之间的相同树片段的数量来比较句法树之间的相似度,但没有区别节点的深度特征和句法成分特征. Wang Kai 等[8] 通过使用树核对结构化特征进行建模,从而计算 2 个句子之间的相似度问题,但在相似度计算过程中没有考虑语义信息. J. Jeon 等[9] 通过问题答案之间的相似性来估计 2 个问句之间的相似度,并采用了基于翻译的检索模型,但在相似度计算过程中没有考虑句法特征和语义特征.

本文利用 WordNet 作为语义资源来充分挖掘句子的语义信息,有效地弥补上述模型中缺少语义信息的不足.

2 问句相似性度量方法

2.1 检索模型概述

首先对 Yahoo! Answers 网站抽取语料集中的问题与答案,建立问答对之间的索引. 对于任意的一个查询 Q,先用改进的统计模型和词序相似度线性结合,得到初步结果. 对于每个 Q 返回前 100 个结果,再用 QTFD(Question Topic and Focus Determination) 方法确定每个句子的关键词,对确定的关键词赋予较高

收稿日期:2012-11-15
基金项目:国家自然科学基金(61272370,60973068) 和辽宁省自然科学基金(201202031) 资助项目.
通信作者:林鸿飞(1962-),男,辽宁大连人,教授,博士生导师,主要从事搜索引擎,文本挖掘,情感计算和自然语言理解等方面的研究.

的权重,并重新用改进的统计模型进行检索,再次返回前 1 000 个结果,然后再用语义模型对初次检索的结果再次进行检索,最后融合答案信息重新检索,得到的最终结果用相应的评价指标进行评价. 框架的流程图如图 1 所示.

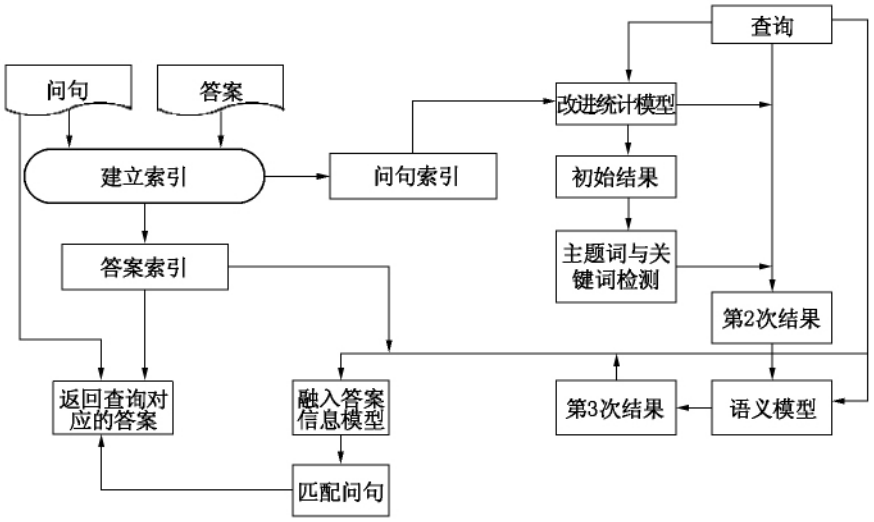

图 1 框架的流程图

2.2 词序相似度

词序相似度[10](*OrdSim*)是反映 2 个句子中所含相同词或同义词在位置关系上的相似程度,以 2 个句子中所含相同词或同义词的相邻顺序逆向的个数来衡量. 设 Q_1 和 Q_2 为 2 个句子,$OnceWord(Q_1,Q_2)$ 为 Q_1、Q_2 中都出现且只出现一次的单词集合,$\boldsymbol{P}_{\text{first}}(Q_1,Q_2)$ 表示 $OnceWord(Q_1,Q_2)$ 中单词在 Q_1 中的位置序列构成的向量,$\boldsymbol{P}_{\text{second}}(Q_1,Q_2)$ 表示 $\boldsymbol{P}_{\text{first}}(Q_1,Q_2)$ 中的分量按对应单词在 Q_2 中的词序排序生成的向量,$\text{Rev}Ord(Q_1,Q_2)$ 为 $\boldsymbol{P}_{\text{second}}(Q_1,Q_2)$ 各相邻分量的逆序数,则 Q_1、Q_2 的句序相似度为

$$OrdSim(Q_1,Q_2)=\begin{cases}1-\dfrac{\text{Rev}Ord(Q_1,Q_2)}{|OnceWord(Q_1,Q_2)|-1}, & |OnceWord(Q_1,Q_2)|>1,\\ 1, & |OnceWord(Q_1,Q_2)|=1,\\ 0, & |OnceWord(Q_1,Q_2)|=0.\end{cases}\tag{1}$$

2.3 改进的统计模型

传统的文档相似度计算使用向量空间模型(VSM: Vector Space Model),文档中含有大量的单词,把文档中的每个单词表示成 1 维向量,从而构成的高维向量可以较好地表示文档. 但对于句子相似度计算,由于一个句子由很少的单词构成,如果仍采用高维空间来表示句子,则会存在较大的稀疏问题,从而不能得到预期的结果. 统计模型[9](SM: Statistical Model)可以更好地表示句子,它的定义为: $QS=Q_1\cup Q_2$,其中 QS 表示问句 Q_1 和问句 Q_2 的并集(即 QS 表示 2 个句子中只出现一次的单词集合). $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$ 表示 Q_1 和 Q_2 所对应的向量,维度等于 QS 中单词的个数,分量遵循如下 2 个规则:

(i) 对于 QS 中的单词,如果在句子 $Q_i(i=1,2)$ 中不存在,则对应 $\boldsymbol{V}_i(i=1,2)$ 中的分量为 0.

(ii) 对于 QS 中的单词如果在句子 $Q_i(i=1,2)$ 中存在,则对应的 $\boldsymbol{V}_i(i=1,2)$ 中的分量为该单词在句子 $Q_i(i=1,2)$ 中出现的频率.

对于 $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$ 采用 cosine 来计算其相似度

$$Sim_{\text{statistic}}=(\boldsymbol{V}_1\cdot\boldsymbol{V}_2)\Big/(\|\boldsymbol{V}_1\|\cdot\|\boldsymbol{V}_2\|).\tag{2}$$

采用改进的统计模型(ISM: Improved Statistical Model),利用语言学知识为句子中的动词和名词赋予较高的权重,并引入同义词对句子进行扩充. 同时使用与(2)式有所不同的相似度计算方法

$$Sim_{\text{statistic}}=(\boldsymbol{V}_1\cdot\boldsymbol{V}_2)\Big/(\|\boldsymbol{V}_1\|+\|\boldsymbol{V}_2\|).\tag{3}$$

2.4 问题主题和焦点确定

问题主题和焦点信息(QTFD: Question Topic and Focus Determination)采用文献[11]的方法确定. 对每个查询 Q 先用改进的统计模型检索,把检索出来的前 N 个结果和 Q 作为一个文档集,在此文档集中对 Q 中的每个单词 w 计算局部 $tf\cdot idf$,计算公式为

$$loc\_tf\cdot idf=tf_{wi}\times\log(loc\_docNum)/loc\_df_{wi}.\tag{4}$$

再把语料库中所有的问题和查询 Q 作为一个文档集,在此文档集中对 Q 的每个单词 w 计算全局 $tf\cdot idf$,计算公式为

$$glob\_tf\cdot idf=tf_{wi}\times\log(glob\_docNum)/glob\_df_{wi}.\tag{5}$$

对 q 中的每个单词 w 进行排序:

$$wRank = rank(glob\_tf \cdot idf) / rank(loc\_tf \cdot idf), \quad (6)$$

其中 $rank(glob\_tf \cdot idf)$ 和 $rank(loc\_tf \cdot idf)$ 的值为 Q 中的每个单词 w 分别根据 $loc\_tf \cdot idf$ 和 $glob\_tf \cdot idf$ 在 Q 中的排序值.

对于每个查询 Q，根据 $wRank$ 值排序结果，取前 50% 的词作为句子的主题信息词和焦点信息词，采用改进的统计模型重新进行检索，并在此次检索中对主题词和焦点词赋予较高的权重. 对于查询 Q 的主题词和焦点词的权重，将上述每个单词的 $wRank$ 值乘以 100 作为相应的权重. 因为，根据 $glob\_tf \cdot idf$ 和 $loc\_tf \cdot idf$ 比值得出来的数据比较小，直接用这个值作为权重不足以区分每个单词的重要性.

2.5　语义模型

在 WordNet 中 2 个单词之间距离越近认为它们之间语义相似性越大，反之则认为语义相似性越小. 对于给定 2 个特征词 w_1 和 w_2，利用 *WordNet* 来计算 2 个特征词之间的语义相似度公式为

$$Sim(w_1, w_2) = 1 \Big/ (dis(w_1, w_2) + 1), \quad (7)$$

其中 $Sim(w_1, w_2)$ 表示 w_1 和 w_2 之间的语义相似度，$dis(w_1, w_2)$ 表示 w_1 和 w_2 在 WordNet 中的语义距离. 公式(7) 较好地表现了 2 个单词的相似度随语义距离的增大而减小的特点. 另外，规定相同单词之间的语义距离为 0，此时的相似度为 1.

计算出词语的语义相似度后，再采用二分图[6] 的方法来进一步计算 2 个句子的相似度，计算公式为

$$Sim_{\text{semantic}} = \frac{1}{2}\Big(\sum_{a_i \in Q_1} \max\ ssim(a_i, Q_2) \Big/ \| Q_1 \| + \sum_{b_j \in Q_2} \max\ ssim(b_j, Q_1) \Big/ \| Q_2 \|\Big), \quad (8)$$

其中 Q_1 和 Q_2 表示给定的 2 个问句，a_i 和 b_j 分别表示 Q_1 和 Q_2 的特征词，$\| Q_1 \|$ 和 $\| Q_2 \|$ 分别表示 Q_1 和 Q_2 中特征词的个数目

$$\max\ ssim(a_i, Q_2) = \max(sim(a_i, b_1), sim(a_i, b_2), \cdots, sim(a_i, b_{|Q_2|}));$$

$$\max\ ssim(b_j, Q_1) = \max(sim(b_j, a_1), sim(b_j, a_2), \cdots, sim(b_j, a_{|Q_1|})). \quad (9)$$

2.6　基于答案信息模型

基于答案信息模型[12]，利用问题的答案对句子进行扩展，使句子具有更丰富的信息. 传统的查询扩展通常采用基于相关反馈[13] 和基础查询日志[14] 方式，主要以提高召回率为目标，而本文采用的查询扩展是为了更好地表示句子信息，从而更有利于计算句子的相似度.

令 x 表示一个新的问题，则 x 的上下文向量 $QCV(x)$ 的计算过程为：

(i) x 作为 Q/A 问答对集合 C 中的问题；

(ii) $R(x)$ 表示由 x 在 Q/A 问答对集合 C 中检索出来的前 N 个问答对的集合 $(p_1, p_2, \cdots, p_n)$；

(iii) 对每个 p_i 计算向量 $\boldsymbol{v}_i$，其中向量 $\boldsymbol{v}_i$ 的每一维为 p_i 中单词的权重；

(iv) 对每一个 $\boldsymbol{v}_i$ 取其前 m 个最高的权重的单词，构成新的向量 $\boldsymbol{v}_i$.

$QCV(x)$ 的计算公式为

$$QCV(x) = \frac{1}{n}\sum_{i}^{n} \frac{v_i}{\| \boldsymbol{V}_i \|}. \quad (10)$$

查询扩展方法步骤(iii) 的权重计算使用信息检索领域最常用的 $tf \cdot idf$ 方法，每个 Q/A 问答对 p_j 中每个单词 t_i 的权重 w_{ij} 为

$$w_{ij} = tf_{ij} \times \log N/df_i, \quad (11)$$

其中 tf_{ij} 为 t_i 在 p_j 中出现的频率，N 为 Q/A 问答对集合中问答对的总数，df_i 为包含 t_i 问答对的个数.

最后使用 cosine 值来计算问题 x 和问题 y 的相似度

$$\cos(QCV(x), QCV(y)) = \sum_{i=1}^{n}(cx_i \times cy_i) \Big/ \Big(\sqrt{\sum_{i}^{n} cx_i^2}\sqrt{\sum_{i}^{n} cy_i^2}\Big), \quad (12)$$

其中 cx_i 和 cy_i 分别是向量 $QCV(x)$ 和向量 $QCV(y)$ 中每一维的权重.

3　实验结果与分析

3.1　实验数据

实验语料是取自 Yahoo! Answers 的一个数据集. 根据每个类别中问句数量对类别进行排序，选取问句数量最多的 60 个类别，每个类别的问句数量均在 1 000 个以上. 对选出的 60 个类，选取问句数最多的前 30 个类别作为相关类别，剩下的 30 个类别作为非相关类别. 以前 30 个类别为相关类别是因为本文的查询由前 30 个类别提供，共计 143 个查询，这些查询由不同的长度、句式构成.

由于选取的语料都是已解决的问题答案对，即每个问题都有最佳答案. 根据有共同最佳答案的问句是最相似的来判定 2 个句子是否相似，因此分别选出这 143 个查询所对应有共同最佳答案的问句作为标准答案集，每个查询都对应 5 ~ 6 个这样的对应问句.

从前 30 个相关类别中去除这 143 个查询问句以及标准答案问句，重新对每个类别随机选择 500 个问句，共选取 15 000 个句子作为噪音句，同时把这 143 个查询对应的标准答案集和这 15 000 个噪音句融合在一起，并在此数据集上进行实验；采用相同的

方法从另外的非相关类中选取15 000个问句作为噪音句,按上面的方法组成数据集,并在此数据集上进行同样的实验. 实验的语料信息如表1和表2所示.

表1 相关语料集的统计信息情况 单位:个

问题数	答案数	最佳答案数量	类别数	问题平均答案数
165 847	1 506 186	165 911	30	9.08

表2 非相关语料集的统计信息情况 单位:个

问题数	答案数	最佳答案数量	类别数	问题平均答案数
35 151	342 255	35 164	30	9.73

3.2 实验结果与分析

为了评价本文检索方法的性能,采用 $P@N$ 和 MAP(Mean Average of Precision)2个指标对试验进行评价,同时采用了7种不同的检索系统进行对比实验,表3中是7中不同方法的描述,其中(1)和(2)是文献[6]提出的方法(Baseline),

表3 实验方法和表述

方法名称	方法描述
(1) SM	统计模型
(2) SM + SEM	在SM基础上上引入语义信息(基于WordNet)
(3) ISM	改进的统计模型
(4) ISM + SEM	在模型(3)基础上引入语义信息(基于WordNet)
(5) ISM + QTFD	在模型(3)的基础上引入主题和焦点信息
(6) ISM + QTFD + SEM	在模型(5)的基础上引入语义信息
(7) ISM + QTFD + SEM + BAIM(our)	在模型(6)的基础上引入答案信息(本文方法)

采用相关类中的实验结果如表4所示.

表4 在相关语料集上7种模型的 MAP 和 P@3 值

模型	MAP	$P@3$
SM	0.552 6	0.564 2
ISM	0.596 3	0.603 9
SM + SEM	0.689 4	0.723 6
ISM + SEM	0.739 6	0.764 7
ISM + QTFD	0.715 6	0.740 3
ISM + QTFD + SEM	0.742 7	0.786 4
ISM + QTFD + SEM + BAIM(our)	0.757 1	0.822 8

从表4结果可以看出:(ⅰ)改进后的统计模型(ISM)比原来的统计模型(SM)在 MAP 和 $P@3$ 上都有了小幅提高,说明遵循语言学知识,名词和动词在一个句子中的重要性相应地高于其他词性的词,同时进行同义词扩展也有利于2个句子之间相似度的比较. 弥补了原来模型的不足.(ⅱ)在ISM中引入语义信息(ISM + SEM)比ISM引入QTFD信息(ISM +QTFD)在MAP上高出0.0240,而在 $P@3$ 上也高出0.024 4. 说明在相关类中语义特征起到主要作用,而词特征起到次要作用.(ⅲ)在ISM中同时引入语义信息和QTFD信息(ISM + QTFD + SEM)相对于两者单独引入(ISM + SEM和ISM + QTFD),不管是在 MAP 还是在 $P@3$ 上都有相应的提高,说明综合运用词信息,语义信息和统计信息可以得到较好的结果.(ⅳ)在ISM + QTFD + SEM中引入BAIM信息即引入答案信息相对于ISM + QTFD + SEM在MAP上提高1.9%,在 $P@3$ 上提高4.6%,说明综合利用词信息、语义信息、统计信息和答案信息可以更加充分的挖掘句子之间的信息,从而更有利于句子之间的相似度计算. 同时也证明提出的方法是可行的.

采用非相关类中的实验结果如表5所示.

表5 在非相关语料集上7种模型的 MAP 和 $P@3$ 值

模型	MAP	$P@3$
SM	0.643 5	0.663 2
ISM	0.709 1	0.687 4
SM + SEM	0.732 5	0.753 5
ISM + SEM	0.761 2	0.776 3
ISM + QTFD	0.816 2	0.824 3
ISM + QTFD + SEM	0.826 8	0.836 2
ISM + QTFD + SEM + BAIM(our)	0.845 0	0.878 7

从表5和表4的对比可以看出:(ⅰ)不管是 MAP 值还是 $P@3$ 值,整体上非相关领域的效果比相关领域好. 这是因为在相关领域中,噪音句和标准答案句之间相似性较高,存在的干扰较大;而在非相关领域,噪音句和标准答案句之间的相似度较低,存在的干扰度较小,所以整体上非相关领域的结果要好些.(ⅱ)在ISM中引入语义模型(ISM + SEM)比ISM引入QTFD信息(ISM + QTFD)在MAP上降低了0.0550,而在 $P@3$ 上也降低了0.0480. 说明在非相关类中语义特征起到次要作用,而词特征起到主要作用.(ⅲ)从表4和表5中ISM + SEM与ISM + QTFD结果比较可以看出,在相关领域中,语义特征起到了主要作用,因为相关领域在同一个类别下同义词较多,基于语义特征的方法正好解决此问题;而在非相关领域中,词特征起到主要作用,因为非相关领域同义词出现的概率较少,而且非相关领域歧义情况也出现较少.(ⅳ)本文提出的多种特征融合的方法(ISM + QTFD +SEM + BAIM)在相关和非相关领域都得到了较高的评价结果,由此表明该方法是一个可行的通用方法.

4 结束语

本文探讨了社区问答系统中的问句相似度计算

方法，建模时较全面地运用了语言学知识特点，充分挖掘了句子的结构信息；同时根据社区问答系统存在大量问答对的特点，把答案信息引入到句子相似度计算中. 多角度的对比实验结果表明，将答案特征、词序特征、统计特征和语义特征等进行适当融合，用于问句的相似度计算，能够有效解决社区问答系统中句子匹配的问题. 当然，本文描述的方法尚存在着速度等方面的问题，这需要在今后的研究中加以改进.

5　参考文献

[1] Burke R D，Hammond K J，Kulyukin V A，et al. Question answering from frequently asked question files：experiments with the faq finder system [J]. AI Magazine，1997，18(2)：57-66.

[2] Jijkoun V，Rijke M D. Retrieving answers from frequently asked questions pages on the web [C]. Bremen：CIKM，2005：84-90.

[3] Cao Xin，Cong Gao，Cui Bin，et al. The use of categorization information in language models for question retrieval [C]. Hong Kong：CIKM，2009：256-274.

[4] Duan Huizhong，Cao Yunbo，Lin C Y，et al. Searching questions by identifying questions topic and question focus [C]. Columbus：HLT，2008：156-164.

[5] Xue Xiaobing，Jeon J，Croft W B. Retrieval models for question and answer archives [C]. New York：NY，2008：475-482.

[6] Song Wanpeng，Feng Maijie，Gu Naijie，et al. Question similarity calculation for FAQ answering [C]. New York：SKG，2007：298-301.

[7] Collins M，Duffy N. Convolution kernels for natural language [M]. massachusetts：MIT Press，2001：625-632.

[8] Wang Kai，Ming Zhaoyan，Chua T S. A syntactic tree matching approach to finding similar questions in community-based QA services [C]. Boston：MA，2009：187-194.

[9] Jeon J，Croft W，Lee J. Finding semantic similar questions based on their answers [C]. New York：ACM Press，2005：617-618.

[10] 吕学强，任飞亮，黄志丹，等. 句子相似模型和最相似句子查找算法 [J]. 东北大学学报：自然科学版，2003，24(6)：531-534.

[11] Zhang Yaoyun，Wang Xiaolong，Wang Xuan，et al. Diversifying question recommendations in community-based question answering [C]. Berlin：Springer，2011：177-186.

[12] Wang Jun，Li Zhoujun，Hu Biyun. A context approach to measuring similarity between questions in the community-based QA services [J]. Fuzzy Systems and Knowledge Discovery，2010，5：2408-2411.

[13] Xu Jinxi，Croft W B. Query expansion using local and global document analysis [C]. Zurich：ACM，1996：4-11.

[14] Wang Xuanhui，Zhai Chengxiang. Mining term association patterns from search logs for effective query reformulation [C]. California：CIKM，2008：479-488.

Question Similarity Calculation Based on Feature Fusion in Community Question Answering

YANG Hai-tian，WANG Jian，LIN Hong-fei*

(Information Retrieval Laboratory，Dalian University of Technology，Dalian Liaoning 116024，China)

Abstract: A testing method based on the framework combining different features with the characteristic of the answer，the word order characteristic，the statistical characteristic and semantic characteristic is proposed to caculate question simularity of question answering. The experiments on the true labeling dataset extracted from the Yahoo! Answers shows that the proposed method achieves a better performance.

Key words: questions similarity; community question answering; similarity calculation; feature fusion

(责任编辑：冉小晓)