文章编号：2095-2163(2019)06-0041-04 中图分类号：TP18 文献标志码：A

相似问句判别研究

尹庆宇，张 宇，刘 挺
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：对于搜索引擎而言，如何能够正确理解用户提出的问题十分重要。而在识别问句的过程中，如何能够对形式不同而语义相似的问句进行相似性识别后，归一化处理，则会对整个搜索引擎的效果有一个明显的提升。对此，本文提出了一种基于机器学习的问句相似性判别模型，从数据集的构建到特征的提取，探究了相应的解决方案。本文创新性地从 5 个方面提取了不同类型的特征，并将其应用到整个分类器的建模过程中。实验结果表明，该方法能够在现有的语料上取得令人满意的结果，*F* 值达到了 83%。

关键词：相似度；问句；机器学习

Research on the identification of similar queries

YIN Qingyu, ZHANG Yu , LIU Ting
(Information Retrieval Lab, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 For a search engine, it is important for it to understand the queries from users correctly. Therefore, in the process of query understanding, to identify the similarity of different queries at the semantic level and then generate similar answers can bring much more better results. In this paper, we propose a machine learning method that identifies the similarity of queries. We propose 5 kinds of features and then apply these features for the classifying process. Experimental results show that our method achieve great performance in the dataset and the F-score is 83%.

【Key words】 similarity; query; machine learning

0 引 言

搜索引擎正确理解用户输入的查询是十分必要的。在实际应用中对于同一个问题，不同用户的提问形式往往不同。比如用户想得到一个 U 盘格式化的方法，那么有些人会问："如何对 U 盘格式化"，还有些人可能会问："怎么对 U 盘格式化"，或者"U 盘格式化的方法？"等等。如果一个搜索引擎能够将这些相似问题理解为同一个意思，就能够正确返回给用户结果。但是，有些问题虽然形式上比较接近，用户问的却是完全不同的意思。比如用户提问"姚明是谁的爸爸"和"姚明的爸爸是谁"，如果搜索引擎将这 2 个问题视为同一个，返回的结果之一就是错误的。因此，搜索引擎应该能够将这些问题很好地区分开。

本文将相似问句判别视为一个二元分类问题，即对于两个问句，或者二者可以归一化，或者不可以。现有的判别方法主要分为两种：基于规则的和基于统计机器学习的。基于规则的方法是根据数据的特点抽出一些模板，然后利用模板去匹配句子，如果句子匹配的模板为相似模板，那么二者为相似的句子，反之则不是。基于统计机器学习的方法是利用一些标注好的数据，抽取特征，选取一个适当的机器学习方法训练一个分类器，然后利用这个分类器对新数据进行二元分类。基于规则的方法受模板所限覆盖面不是很大，但是相对来说比较准确。模板的抽取方式可以采取人工方式或者从标注好的数据中自动抽取。基于统计机器学习方法的优点是适用面比较广，即便是对于数据集中没有出现过的形式，如果抽取的特征恰当，仍能够正确地对其进行分类。

1 研究方法

主要研究内容分以下几点。首先是数据问题，即如何获取数据，以及对于获取的数据应该做何处理。然后是具体的实现方法，这也是本课题的核心内容。最后是评价问题，即如何评价系统的判别结果的好坏。本文将对上述问题分别进行说明。

作者简介：尹庆宇(1990-)，男，博士研究生，主要研究方向：自然语言处理、深度学习；张 宇(1972-)，男，博士，教授，主要研究方向：自然语言处理、问答系统、机器学习等；刘 挺(1972-)，男，博士，教授，博士生导师，主要研究方向：自然语言处理、机器学习、社会计算等。
通信作者：尹庆宇 Email：qyyin@ ir.hit.edu.cn

收稿日期：2019-03-22

1.1 数据获取及处理

首先,需要获取到若干问题对,然后才能对这些问题对进行分类处理(可归一化,不可归一化)。在中文领域,没有公开的问题对语料,因此,选取了百度知道这个平台,从网上抓取需要的问题对。

爬虫算法的流程如图 1 所示。其基本流程为:从一系列种子(Seed)网页开始,使用这些种子网页中的 URL 链接去获取其它页面,把这些网页中的 URL 链接依次提取出来,访问 URL 链接对应的页面。在网络爬虫中,使用哈希表记录一个页面是否被访问过,未被访问的 URL 链接则放入队列。由调度算法,每次从队列中取出一个 URL,然后通过 HTTP 协议爬取对应页面,保存到网页库。整个过程不断重复,直到有足够的网页被访问过,或者已达到其它的既定目标。

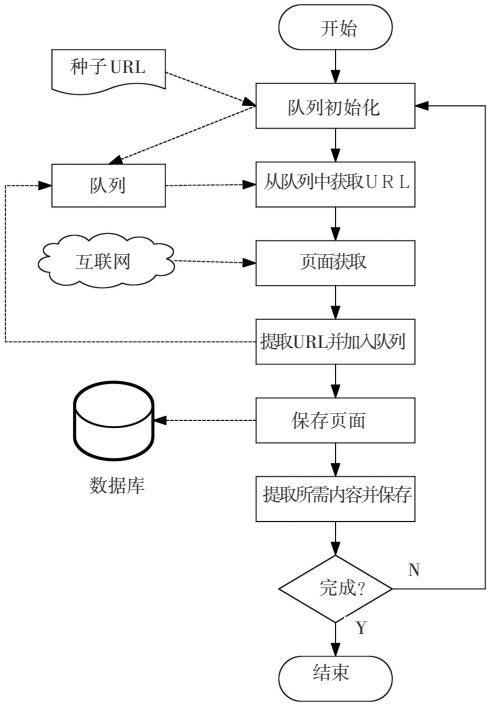

图 1 爬虫算法流程图

Fig. 1 Process of the crawler

由百度知道上爬取了若干网页原始数据后,需要从中抽取有用的信息,即问题对。由此可知在一个问题的页面中,存在有如下两部分内容—类似问题和相关知识,这两部分内容恰好可以构成所需要的问题对。如图 2 所示。问题是:iphone 好用么(http://zhidao.baidu.com/question/542432940.html)。人们抽取了其中的"类似问题"块同原始问题组成问题对,作为正例(可归一化的问题对),抽取其中"相关知识"块同原始问题组成负例(不可归一化的问题对)。这样,就获取了充足的问题对。

图 2 页面抽取块样例

Fig. 2 Example of web-page

1.2 一致性判别方法

研究中采用机器学习的方法来处理两个问句的一致性问题。采用逻辑斯蒂回归算法进行分类。为了能够更好地对问题进行判别,除一些基本特征外,人们还从 5 个方面抽取了问句的相似度信息。表 1 中列出了抽取的特征,下边将分别介绍在计算相似度上所使用的方法。

表 1 特征向量表

Tab. 1 Feature description

特征类别	说明
String kernel 特征	用 string kernel 方法从结构的角度获取两个句子的相似度
Hownet 特征	用 hownet 方法从语义的角度获取两个句子的相似度
Term Wieght 特征	利用搜索引擎从词在句子中的重要性角度来计算两个句子的相似度
Tf idf 特征	利用 tf idf 信息获取两个句子的相似度
Rank 特征	利用依存树,模仿 page rank 算法,从出入度上衡量词的权重,进而计算两个句子的相似度
句法结构特征	从依存关系中获取句子中的词信息,利用这些信息来判别句子是否相似
Normal 特征	包括了句子词数差、长度差等信息,利用这两个差来判断两个句子的相似度

HowNet(即知网)是一部详尽的中文语义知识词典,被广泛应用于计算词和句子的相似度任务上[1-4]。虽然和其它的语义词典一样,也有一个反映知识结构的树状层次体系,但实际上有着本质的不同。在 WordNet[5] 中,概念是描述词义的最小单位,所以,每一个概念都是这个层次体系中的一个结点。而在知网中,每一个概念由多个义原组成,概念本身不是这个层次体系中的结点,而义原才是。

对于 2 个词 W_1 和 W_2,如果 W_1 有 n 个概念[S_{11}, S_{12}, …, S_{1n}], W_2 有 m 个概念:[S_{21}, S_{22},…, S_{2m}], W_1 和 W_2 的相似度 $Sim(W_1, W_2)$ 为各个概念的相似度的最大值,如公式(1)。

$$Sim(W_1, W_2) = \max_{i=1..n,\ j=1..m} Sim)S_{1i},\ S_{2j}), \quad (1)$$

因为所有的概念都最终归结于用义原来表示，所以义原的相似度计算是概念相似度的基础。由于所有的义原根据上下位关系构成树状的义原层次体系，可以简单的通过语义距离计算相似度。义原的语义距离如公式(2)。

$$Sim(s_1,s_2)=\frac{\alpha}{d+\alpha}, \quad (2)$$

其中，s_1，s_2 表示2个义原，d 是 s_1，s_2 在义原层次体系中的路径长度。α 是一个可调节的参数，在本课题实现的基于 HowNet 的词汇语义相似度计算方法中 $\alpha = 0.5$。2 个词的相似度计算方法如公式(3)。

$$Sim(S_1,S_2)=\sum_{i=1}^{4}\beta_i\prod_{j=1}^{i}Sim_j(S_1,S_2). \quad (3)$$

树核(String kernel)算法是通过字符串结构上的特征来计算字符串之间的相似度[6-9]。String kernel 计算预处理后的问题对之间的相似度数值，主要是基于字符串核函数的方法。即首先将给定的字符串(问题对)拆分成子串集合(子串的长度可通过参数调节)，然后通过核函数计算子串集合之间的相似度，从而通过线性合并得出问题对之间的相似度。

利用 Term Weight 来计算相似度的方法也是基于向量空间模型(VSM)文本相似度量的一种方法[10-11]。与用 tfidf 计算相似度的方法不同之处在于给词项赋权的方法。本文没有直接用词频等统计信息来给词项赋权，而是利用了搜索引擎，通过搜索结果的重合率来为句子中的词项赋权。为词项赋权所用具体方法如下：

(1)将整个句子放进 baidu 中检索，记录前 20 个检索结果。

(2)去掉一个词，再将句子放入 baidu 中检索，记录前 20 个检索结果。

(3)计算第二次的检索结果占第一次的检索结果的百分比，然后用 1 减，得到的数值即可认为是这个词在句子中的重要性分数。词的分数越大，说明越重要，其权重就越大。

通过这个方法可以得到一个句子中每个词项的权重。但是，考虑到如果对每个句子都要放入搜索引擎中检索多次，时间消耗比较大，所以采用机器学习中的 SVM-RANK 算法[12]，通过学习来达到自动对句子中的词项赋权的目的。

对于句子，首先要做预处理，预处理包括分词，词性标注，句法依存分析等，以获得词语本身的词性特点以及词语之间的句法上的关系。对于句子中的每个词项选取以下特征：

(1)NOUN：该词是否是名词。

(2)S&C：该词是否是主语或者宾语。

(3)TermFreq(词频)：词语在整个文档中出现的次数。

(4)DocFreq(文档频率)：整个文档中出现该词的文档的个数。

通过这种方式，可以得到一个句子中每个词项的权重，同 tfidf 方法[13]一样，为每个文本(问句)建立向量空间模型，通过余弦计算得到 2 个句对之间的相似度。

在网页排序算法中[14]，一般认为，如果一个网页被很多其它网页链接，那么这个网页相对来说是比较重要的网页。模仿这种思想，从一个句子的依存树中，通过各个词项的依存关系，也对各个词的权重做出了衡量。

利用这个方法得到的权重，也能够从一定方面反应词项在句子中的重要性。利用这样的方法，通过入度给一个句子中的词赋权。对于词 w，其权重公式为 W = In/Norm。这里的 In 表示词 w 的依存链入度。Norm 为这个句子中所有词的入度和。赋权后，利用权重为问句建立向量空间模型，然后通过余弦计算得到 2 个句对之间的相似度。

其它特征的提取包括了一些比较常规的特征，如 2 个句对的词数差、句对的长度差、句对的包含关系等。上文所述的种种特征都能够从某些方面来得到 2 个问句的相似信息，但是并没有对句子中的词序做出区分。比如对于这样两个问句："谢霆锋爸爸是谁"，"谢霆锋是谁爸爸"。已经提取的特征没有办法区分这种关系，因此引入了另外一类特征—句法结构特征。

在这类体征中，人们借助了三元组的思想，对于每个问句构建了一个"三元组"。构建方式如下：通过依存句法树，然后在这颗树上获取 HED、SBV、VOB，3 个节点的信息作为句子的三元组，然后通过比较 2 个句子的三元组成分是否一致作为特征加入分类器。对于句子"谢霆锋是谁儿子"，其三元组抽取为"谢霆锋"，"是"，"谁"。而句子"谢霆锋儿子是谁"的三元组则会被抽取为"儿子"，"是"，"谁"。通过这类特征的提取，能够很好地从词序的关系上获取问句的相似信息。

1.3　评价规则

本系统中采用在自然语言处理领域常用的 3 个

评价指标,对实验结果进行评价。即准确率(precision)、召回率(recall)和 F 值(F1 Score)。

2 实验结果

实验中共计标注了 4 000 个问题对。采用测试和训练的语料比例为 1 ∶ 4,即 80%的数据用来训练,余下 20%的数据用来测试。在测试的过程中,采用 5 轮迭代取平均的测试方法,得到最终的准确率 P,召回率 R 和 F 值见表 2。

表 2 实验结果

Tab. 2 Experimental Results

P	R	F
0.89	0.78	0.83

从结果中不难看出,在提问类句子归一化问题上,基本达到了实用的水平,能够从一定程度上对问句是否能归一化做出判断。

3 结束语

随着大数据时代的到来,人们被海量的信息淹没。如何从海量的信息中找到所需要的信息是目前的一大挑战。对于同一个问题,不同用户的提问形式往往不同,因此如何判断用户输入查询的语义是否一致对改善搜索性能具有重要意义。本研究将这一问题定义成了一个二元分类问题,即查询的语义是否一致。然后,在百度知道上面爬取了大量的语义查询对,并对其进行了人工标注。为了能够覆盖查询的语义信息,人们对问句从不同方面提取了几十个特征,如 HowNet 相似度、String Kernel 相似度、tf-idf 相似度、Term Weight、依存句法分析等特征。选择了二项逻辑斯蒂回归方法构建分类器,该方法在标注的数据集上 F 值达到了 0.83。本文在问句一致性的研究上提出了相对有效的语义一致性判断算法。为提问类句子归一化研究做出了一些探索。虽然,本文取得了不错的实验结果,但是还存在很多问题:例如训练数据稀疏问题;自然语言处理工具的分析错误等问题,这些问题将有待进一步研究解决。

参考文献

[1] 刘群, 李素建. 基于《知网》的词汇语义相似度计算. 中文计算语言学. 2002 Aug;7(2):59-76.

[2] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算. 中文信息学报. 2006;20(1):16-22.

[3] Zhu, Yan-Lan, et al. "Semantic orientation computing based on HowNet." Journal of Chinese information processing 20. 1 (2006): 14-20.

[4] Zhendong, Dong, Dong Qiang. Hownet and the computation of meaning (with Cd-rom). World Scientific, 2006.

[5] Miller, George A. "WordNet: a lexical database for English." Communications of the ACM 38.11 (1995): 39-41.

[6] Zhou, Guodong, et al. "Tree kernel-based relation extraction with context-sensitive structured parse tree information." Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007.

[7] Vert, Jean-Philippe. "A tree kernel to analyse phylogenetic profiles." Bioinformatics 18.suppl_1 (2002): S276-S284.

[8] 孔芳, 周国栋. 基于树核函数的中英文代词消解. 软件学报. 2012 May.

[9] 车万翔. 基于核方法的语义角色标注研究 (Doctoral dissertation, 哈尔滨: 哈尔滨工业大学).

[10] 庞剑锋, 卜东波, 白硕. 基于向量空间模型的文本自动分类系统的研究与实现 (Doctoral dissertation).

[11] 刘少辉, 董明楷, 张海俊, 李蓉, 史忠植. 一种基于向量空间模型的多层次文本分类方法. 中文信息学报. 2002;16(3):9-15.

[12] Elisseeff, André, and Jason Weston. "A kernel method for multi-labelled classification." Advances in neural information processing systems. 2002.

[13] 黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 (Doctoral dissertation).

[14] Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab; 1999 Nov 11.

(上接第 40 页)

[13] SALMAN E H, NOORDIN N K, HASHIM S J, et al. An analysis of periodogram based on a discrete cosine transform for spectrum sensing[J]. Wireless Personal Communications, 2018, 101(4):1261-1279.

[14] CANDES E J, TAO T. Decoding by linear programming[J]. IEEE Transactions on Information Theory, 2005, 51(3): 4203-4215.

[15] 何国林, 丁康, 林慧斌. 基于匹配追踪的齿轮箱耦合调制振动信号分离方法研究[J]. 机械工程学报, 2016, 52(1): 102-108.

[17] SCHNASS K. Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation [J]. IEEE Signal Processing Letters, 2018, 25(12): 1865-1869.

[18] 陈鹏清, 黄尉. 基于 l_1-l_2 范数的块稀疏信号重构[J]. 应用数学和力学, 2017, 38(8): 932-942.

[19] 许志强. 压缩感知[J]. 中国科学: 数学, 2012, 42(9): 865-877.

[20] 阮越雄. 基于多变换域特征提取和机器学习的滚动轴承故障诊断方法[D]. 长沙: 湖南大学, 2018.

[21] 何正嘉, 陈进, 王太勇. 机械故障诊断理论及应用[M]. 北京: 高等教育出版社, 2010.

[22] ESTER M, KRIEGEL H P, SANOER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [J] International Conference on Knowledge Discovery in Databases and Data Mining (KDD-96). Portland: dblp, 1996. 226-231.

[23] MOHAMMAD K B, RAHIM A A, MASOUD M. Spatio-temporal modeling of seismic provinces of iran using DBSCAN algorithm [J]. Pure and Applied Geophysics. 2017, 174(5): 1937-1952.