基于 VSM 和 LDA 模型的 FAQ 问答系统

郑 诚,刘娇丽,项 珑
(安徽大学 计算机科学与技术学院,安徽 合肥 230601)

摘 要:传统的搜索引擎返回的数据太过庞大,很多情况下用户不能快速地找到自己要的答案。在这种情况下,文中引入 FAQ 系统。FAQ 中如何找到最佳匹配答案,是文中的研究重点。改进了传统的 VSM 模型,使得它能更好地体现问题中词的权重。重点引入了 LDA 模型,并用计算机故障领域内的文档资料对它进行训练,得到主题-词的概率分布。通过主题-词中词的概率分布,计算词与词的相关度,提出通过词与词间相关度计算句子与句子间相似度的算法。对两个算法进行综合,得到最终的相似度算法。文中对 FAQ 进行整理,得到了 FAQ 问答系统的雏形。通过实验分析,说明相似度算法有很好的效果。

关键词:VSM;相似度计算;LDA (Latent Dirichlet Allocation);主题-词分布
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2014)01-0133-03
doi:10.3969/j.issn.1673-629X.2014.01.034

FAQ Answering System Based on VSM and LDA Model

ZHENG Cheng,LIU Jiao-li,XIANG Long
(School of Computer Science and Technology of Anhui University,Hefei 230601,China)

Abstract:The data returned by the traditional search engine is too large,users cannot quickly find the answer they want sometimes. In this case,introduce FAQ system. How to find the best match in the FAQ system is the focus. An improved VSM model is presented in this paper. This new model is used in order to reflect the weight of the terms in question better. LDA,which was trained with documentation within the domain of computer malfunction generates a probability distribution of topic-term by which the relevance between words is calculated. Then the algorithm of calculating similarity between sentences by calculating relevance between words was presented. Combined with the above two algorithm,get the final similarity algorithm. FAQ is collected and rudiment of FAQ answering system is implemented in this paper. The algorithm used is proved well by the experiments.

Key words:VSM;similarity calculation;LDA (Latent Dirichlet Allocation);topic-term distribution

0 引 言

随着科技发展,网络承载的信息量剧增,信息已处于爆炸的时代。而现在的搜索引擎,大部分还是按照关键词进行搜索。这在某种程度上阻碍了人们利用互联网进行有效的知识获取、共享和交换。

对于问答系统,用户可以把整个问题直接交给问答系统,而不需要把自己的问题分解成关键字。问答系统结合自然语言处理技术,通过对问题理解,能够直接返回给用户想要的答案[1]。

FAQ 是提问频率高的常见的问题,这些常见的问题和对应的答案存储在数据库中。用户提出问题以后,系统可以先对常见问题库进行检索,找出相似的问题,直接将问题对应的答案返回给用户。这样可以节省大量的时间。在上述过程中,相似度计算是找出相似度问题的核心。计算用户提出问题与 FAQ 问答系统中存储问题的相似度,如果相似度大于某一固定阈值,则把此问题对应的答案作为用户提出问题的答案,并返回给用户。

文中在总结以往句子与句子之间的相似度计算算法的基础上,提出了一种应用 LDA 模型的新的句子相似度计算方法,并成功应用在计算机领域的自动问答系统中。

1 相关工作

相似度算法是检索中必备的环节,也是问答系统的中心。目前研究的相似度算法主要有基于句子表层

收稿日期:2013-03-07 **修回日期**:2013-06-11 **网络出版时间**:2013-09-29
基金项目:安徽省自然科学基金资助项目(11040606M133)
作者简介:郑 诚(1964-),男,副教授,研究方向为数据挖掘与 Web 语义检索;刘娇丽(1987-),女,硕士研究生,研究方向为 Web 语义检索。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20130929.1544.041.html

关键字的相似度算法,它通过词的共现来体现句子的相似度,研究应用较多的是 VSM 模型[2-3];还有基于语义的相似度算法[4],这些文章中大部分应用了知网的接口[5-7]。另外,还出现了本体的相似度算法[8]以及基于粒度的相似度算法。

1.1 VSM 模型

空间向量模型以特征项作为文档表示的基本单位[9],每一个文档被看成由所有特征项组成的 n 维特征空间的一个向量:$D = (T_1, W_1; T_2, W_2; \cdots; T_n, W_n)$。其中 W_i 为第 i 个特征向量 T_i 在文档中的权重。在传统的向量空间模型中,权重 W_i 一般用 TF - IDF 来表示,它的主要思想为:如果某个词或短语在一篇文章中出现的频率 TF 高,并且在其他文章中很少出现,则认为此词或者短语具有很好的类别区分能力。

TF 为词频,表示词条在文档 d 中出现的频率,表示方式如式(1):

$$\mathrm{TF}(t,d) = \sum t \in d \tag{1}$$

IDF 文档频率,表示方式如式(2):

$$\mathrm{IDF}(t,D) = \log \frac{|D|}{|\{d \in D: t \in d\}|} \tag{2}$$

式中,$|D|$为总的文档数;$|\{d \in D: t \in d\}|$ 为包含词条的文档数。

两个句子 $Q_1(T_1, T_2, \cdots, T_n)$,$Q_2(t_1, t_2, \cdots, t_n)$,用两个向量的余弦表示它们的相似度,如公式(3):

$$\mathrm{Sim}(Q_1, Q_2) = \frac{\sum_{i=1}^{n} T_i \times t_i}{\sqrt{\left(\sum_{i=1}^{n} T_i^2\right)\left(\sum_{j=1}^{n} t_j^2\right)}} \tag{3}$$

1.2 LDA 模型

LDA 是一个三层的贝叶斯结构,由主题、文档、词组成[10-11]。文档集为 $D(W_1, W_2, \cdots, W_M)$,总共有 M 篇文档。M 篇文档中包含有 N 个词,文档 W 表示为(W_1, $W_2, \cdots, W_N$)。k 是主题的个数。图 1 中灰色的部分表示词典中的一个词,用 w 表示,它是唯一能观测到的变量。θ 表示文档在主题上的分布,对于文档 d,θ_d 服从 Dirichlet 分布 $\mathrm{Dir}(\theta_d \mid \alpha)$,$\alpha$ 为超参数。Z 为主题下词的概率分布。

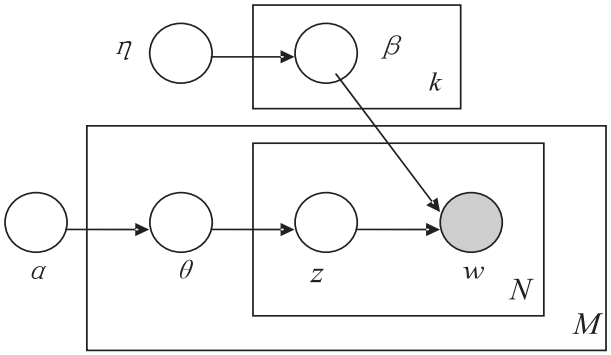

图 1 LDA 模型表示

2 基于 VSM 和 LDA 模型的相似度计算方法

2.1 问题处理

问答系统中由于用户提出的是一句问题,能确定这句话意思的核心尤为关键。任何句子都是由关键成分(主、谓、宾等)和修饰成分(定、状、补等)构成。通常情况下,句子的关键成分主语和宾语多为词,谓语多为动词。所以提取出名词和动词。如在问题“键盘/NN 是/VC 什么/PN ? /PU”中,提取的关键词为“键盘[名词]、是[动词]、什么[疑问词]”。首先,删除停用词、语气词。对问题进行分词,提取关键词,然后结合领域词汇表来提取核心词。

2.2 VSM 和 LDA 模型计算相似度

在下面的讨论中,令用户提问的句子为 q_1,FAQ 问答系统候选问题为 q_2。

Step1:分别对 q_1、q_2 进行问题预处理得到词汇集 Q_1、Q_2。$Q_1 = \{w_1, w_2, \cdots, w_n\}$ ($w_i \neq w_j, i, j \in \{1, \cdots, n\}, i \neq j$),$Q_2 = \{u_1, u_2, \cdots, u_m\}$ ($u_i \neq u_j, i, j \in \{1, \cdots, m\}, i \neq j$)。

其中,$w_i (i = 1, \cdots, n)$ 为 Q_1 中的关键词;$u_j (j = 1, \cdots, m)$ 为 Q_2 中的关键词。Q_1 与 Q_2 里共同词汇个数记作 φ。这样得到两个问题中互不相同的总的词汇个数 N,$N = n + m - \varphi$。得到总的词汇集 $Q_1 \cup Q_2 = \{v_1, v_2, \cdots, v_N\}$。

Step2:按照下面的方式构造向量 $\overrightarrow{Q_1}$ 和向量 $\overrightarrow{Q_2}$。令 $\overrightarrow{Q_1} = (t_1, t_2, \cdots, t_N)$,$\overrightarrow{Q_2} = (t'_1, t'_2, \cdots, t'_N)$。对于 Q_1,Q_2 中的每个词汇 V_i 它的权值如公式(4):

$$\begin{cases} t_i = 1 & V_i \in \text{核心词} \\ t_i = 0.8 & V_i \in \text{名词} \\ t_i = 0.7 & V_i \in \text{动词} \\ t_i = 0.5 & V_i \in \text{疑问词} \\ t_i = 0 & \text{其他} \end{cases} \tag{4}$$

用公式(3)的方法计算相似度。得到向量空间模型的相似度算法如公式(5)所示。

$$\mathrm{Sim}_{\mathrm{VSM}}(Q_1, Q_2) = \frac{\sum_{i=1}^{N} t_i \times t'_i}{\sqrt{\left(\sum_{i=1}^{N} t_i^2\right)\left(\sum_{j=1}^{N} t_j^2\right)}} \tag{5}$$

Step3:LDA 模型通过领域语料库的训练,得到一个 model. twords 文件。它是一个词 - 主题分布。每个主题下有一系列词的概率分布。在某个主题 D_j 下,对于 Q_1,Q_2 中的每个词汇 V_i 它的权值取值为:如果 V_i 不在主题下,$t_{i,Dj}$ 为 0,否则 $t_{i,Dj}$ 的值为词汇在当下主题中

的概率。同样取向量夹角的余弦表示距离。最终取相似度最大的值表示它们的距离。整理后如公式(6)。

$$\mathrm{Sim}_{\mathrm{LDA}}(Q_1,Q_2)=\mathrm{Max}\frac{\sum_{i=1}^{N}t_{i,Dj}\times t'_{i,Dj}}{\sqrt{\left(\sum_{i=1}^{N}t_{i,Dj}^2\right)\left(\sum_{j=1}^{N}t'^2_{j,Dj}\right)}}\quad(0<j<k)\tag{6}$$

Step4:句子的相似度算法如式(7):

$$\mathrm{Sim}(Q_1,Q_2)=\mathrm{Sim}_{\mathrm{SVM}}(Q_1,Q_2)\times\mathrm{Sim}_{\mathrm{LDA}}(Q_1,Q_2)\tag{7}$$

3　实　验

3.1　实验数据

通过对网络资源和图书资源的整理分析,得到一个计算机故障相关 FAQ 常用问题库。用 FAQ 问题库和爬取的领域相关的网页文件对 LDA 模型进行训练,得到领域相关主题。

文中用到的分词工具为复旦大学分词系统。实验中,$\alpha=0.2$,$\beta=0.01$,$k=30$,$\lambda=0.4$,迭代次数为2 000次。

3.2　实验结果

以用户问题"为什么屏幕变黑了"为例,对 FAQ 中的相关句子进行相似度计算,见表 1。

表 1　用户问题与 FAQ 中问题相似度计算

LDA 中问题	VSM	基于语义依存的汉语句子相似度	改进的 VSM+LDA
显示器画面不亮了怎么办	0.21	0.78	0.86
为什么屏幕有亮点	0.325	0.46	0.19
电脑关机画面黑了怎么办	0.28	0.67	0.71

从表中可以看出,虽然句子结构差异较大,但是语义有相似的情况下,改进的算法表现良好。

对整个 FAQ 问题集做一次测试,首先整理出 n 个问题,这 n 个问题能在FAQ系统中找到答案;然后整理 m 个问题,这 m 个问题在FAQ中不能找到答案,再提出 t 个与此领域不相关的问题。这样组成总问题 $N=m+n+t$。

在 FAQ 问题集中进行句子相似度计算,如果相似度大于 0.7,则返回结果,说明找到答案。如果相似度小于0.7,则说明找不到结果。这里,$m=350$,$n=100$,$t=50$,$N=500$。

用 C 表示查找结果正确的问题个数,用准确率 P 来评价算法的性能。则 $P=\frac{C}{N}\times 100\%$。得到的结果如表 2。

表 2　对 N 个问题检索答案的结果

相似度方法	问题总数/个	准确数/个	准确率/%
VSM	500	187	36.2
基于语义依存的汉语句子相似度	500	351	70.2
改进的 VSM+LDA	500	424	84.8

4　结束语

文中对传统的 VSM 算法进行改进,并提出了一个基于 LDA 的相似度算法,从而得出一个适合领域 FAQ 问答系统的相似度算法。

该算法考虑了句子的词语结构,又考虑到了词之间的语义关系,并且它使用了领域内语料进行训练,得到的语义关系也更为突出。笔者进行了大量的实验,实验表明文中的相似度算法在计算机故障检查这一领域内得到了很好的效果。

参考文献:

[1] 毛先领,李晓明. 问答系统研究综述[J]. 计算机科学与探索,2012,6(3):193-207.

[2] 郭庆琳,李艳梅,唐　琦. 基于 VSM 的文本相似度计算的研究[J]. 计算机应用研究,2008,25(11):3256-3258.

[3] 韩如冰,叶得学. 基于 VSM 的权重改进文档相似度算法研究[J]. 软件,2012,33(10):103-105.

[4] Moreda P,Llorens H,Saquete E,et al. Combining semantic information in question answering systems[J]. Information processing and management,2011,47(6):870-885.

[5] 张玉娟. 基于《知网》的句子相似度计算的研究[D]. 北京:中国地质大学,2006.

[6] 江　敏,肖诗斌. 一种改进的基于《知网》的词语主义相似度计算[J]. 中文信息学报,2008,22(5):84-89.

[7] 李　彬,刘　挺,秦　兵,等. 基于语义依存的汉语句子相似度计算[J]. 计算机应用研究,2003(12):15-17.

[8] 谷志锋,刘　勇,郭跟成. 本体映射过程中概念相似度计算方法的改进[J]. 计算机工程与应用,2008,44(8):67-70.

[9] Wang S K M,Ziarko W,Wong P C N. Generalized vector space model in information retrieval[C]//Proceedings of the 8th annual international ACM SIGIR conference on research and development in information retrieval. [s. l.]:[s. n.],1985:18-25.

[10] 张小平,周雪忠,黄厚宽,等. 一种改进的 LDA 主题模型[J]. 北京交通大学学报,2010,34(2):111-114.

[11] Porteous I,Newman D,Lhler A,et al. Fast collapsed gibbs sampling for latent Dirichlet allocation[C]//Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. [s. l.]:[s. n.],2008:569-577.

基于VSM和LDA模型的FAQ问答系统

作者： 郑诚， 刘娇丽， 项珑， ZHENG Cheng， LIU Jiao-li， XIANG Long
作者单位： 安徽大学 计算机科学与技术学院,安徽 合肥,230601
刊名： 计算机技术与发展
ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2013(1)