ISSN 1673-4785
CN 23-1538/TP

中国人工智能学会会刊

智能系统学报

CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

ISSN 1673-4785

9 771673 478120

4
2012
Vol.7 No.4

《智能系统学报》第二届编辑委员会委员名单

The Second Editorial Board of CAAI Transactions on Intelligent Systems

智 能 系 统 学 报
第 7 卷　第 4 期　　（总第 36 期）　　2012 年 8 月

目　　次

·信息交流·

期刊基本参数：CN 23-1538/TP＊2006＊b＊A4＊94＊zh＊P＊ ￥15.00＊1500＊14＊2012-08

CAAI Transactions on Intelligent Systems

Vol. 7 No. 4 (Sum No. 36) Aug. 2012

Contents

International

Serial parameters: CN 23-1538/TP * 2006 * b * A4 * 94 * zh * P * ¥15.00 * 1500 * 14 * 2012-08

DOI:10.3969/j.issn.1673-4785.201204017
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120712.1111.008.html

基于关联词的主题模型语义标注

周亦鹏[1], 杜军平[2]
(1.北京工商大学 计算机与信息工程学院,北京 100048; 2.北京邮电大学 智能通信软件与多媒体北京市重点实验室,北京 100876)

摘 要:互联网主题分析中经常采用概率主题模型对主题进行描述,但存在对于一般用户难以理解的问题,提出一种概率主题模型的自动语义标注方法.首先通过基于语义分类的关联规则挖掘关联主题词并建立候选标签集合,然后以关联词在数据集中的概率分布来设计相关性判别函数,计算候选标签和主题模型的相关度,最后根据最大边缘相关选择高语义覆盖度和区分度的标签.在食品安全和旅游领域主题模型标注的实验表明,与最大概率主题词标记方法相比,提出的方法能够明显提高标注的准确性,并且解决了多标签标记中语义类别单一的问题,能够以较少数量的标签表达更为丰富的语义,这有助于进一步实现更为准确的主题跟踪和主题信息检索.

关键词:主题分析;语义标注;生成模型;关联词;关联规则
中图分类号:TP391 **文献标志码**:A **文章编号**:1673-4785(2012)04-0327-06

Semantic tagging of a topic model based on associated words

ZHOU Yipeng[1], DU Junping[2]
(1. School of Computer Science and Information Engineering, Beijing Technology and Business University, Beijing 100048, China; 2. Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: In topic analysis field of Internet, the probabilistic topic model is often used to describe topic semanteme. But the semanteme of a topic model is difficult for users to understand. An automatic semantic tagging method of a probabilistic topic model is proposed. Firstly, an association rule mining algorithm based on semantic categories is presented to get associated topic words, which consist of a candidate tag set. Then, according to the probability of associated words, a semantic correlation function is used to calculate semantic correlation of candidate tags and topic model. At last, a maximal marginal relevance method is used to select tags with better semantic coverage and discrimination. The experimental results of food safety and tourism topic model proved that, compared with maximum probability topic words tagging method, the proposed method can improve accuracy of topic tagging obviously, and can express more abundant semantemes with a small number of tags, which solve the problem of single semantic category in the multi-tagging method. So it is helpful to achieve more accurate topic tracking and topic information retrieval.

Keywords: topic analysis; semantic tagging; generative model; associated words; association rule

主题分析通常采用概率生成模型,如LDA、PLSA等方法,以语义词概率分布的形式描述主题[1-2],这使得一般用户较难理解主题的内容.通常的方法是取概率较高的若干个语义词来表示主题含义[3],但这种方法也常常不能准确表示整个分布所覆盖的全部语义.因此,提出一种主题模型的自动标注方法,提取具有一定语义覆盖度和区分度的主题关联词来描述主题的内容.

1 主题模型的语义标注

主题模型的自动语义标注通常包括2个步骤:

收稿日期:2012-04-20. **网络出版日期**:2012-07-12.
基金项目:国家"973"计划资助项目(2012CB821206);国家自然科学基金资助项目(91024001,61070142);北京市自然科学基金资助项目(4111002).
通信作者:周亦鹏. Email:yipengzhou@163.com.

首先构造能够表达各种主题语义的候选标签集,标签可以是词、短语,也可以是句子;然后,为不同的主题模型选择与其语义相关的一个或多个标签进行标注.最常用的方法是最大概率主题词标注[4],这种方法的标签集由从文档中抽取的单个词语构成,标签的选择是根据词语在主题模型当中的分布概率来决定的.

相对于单个词语的标注方法,采用短语作为标签进行标注更容易表达主题模型的语义,因此需要生成短语标签集合.常用的短语生成方法是基于统计模型的短语抽取[5-6],即根据同现概率获得同现词,并通过互信息或 χ^2 测试从文本集中抽取可能的短语.但是,这些方法会受到同义词等问题的影响,因此抽取出的标签会出现语义重复问题,并且仅仅根据概率统计获得的标签也存在语义相关性低或语义覆盖性低的问题.此外,如果选择多个标签对主题模型进行标注,还存在如何比较多标签与主题模型的语义相关度、语义覆盖度以及标签间的语义区分度等问题.本文提出一种基于关联词的主题模型自动语义标注方法,其框架如图1所示.

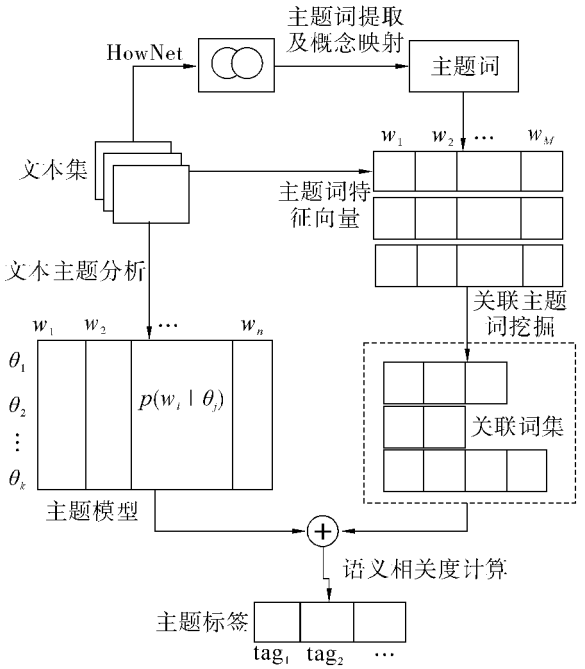


图1 基于关联词的主题模型语义标注框架

Fig. 1 Framework of topic model tagging based on associated words

首先从参考文本集中抽取词语,并根据语言本体将其映射为义原,实现词语在概念上的归并,从而获得描述语义概念的主题词,同时主题模型也从一般的词语分布转换为概念主题词分布;然后根据实体、环境、活动等不同语义类别对概念主题词进行分类,同时采用基于语义分类的关联规则挖掘获得具有语义关联的主题词,从而建立候选标签集;最后,将标签也以主题词概率分布的形式进行描述,并计算其与主题模型的语义相关度,选择具有高语义覆盖度和区分度的多个标签进行标注.

2 关联主题词的生成

2.1 主题词的语义概念

关联词是依据上下文关系经常搭配使用的词,在自然语言中,为了表达的需要,文本中常常会出现大量关联词.例如,在旅游信息中,“杭州”与“西湖”同时出现的几率非常大,在食品安全事件中,“婴儿”、“奶粉”与“三聚氰胺”同时出现的几率较大.如果将这些关联词作为多个语义单元,一方面会增加主题模型的维数,另一方面也降低了主题模型对文档的表达精度.虽然文本特征抽取可以通过预先设定的阈值来降低特征向量的维数,但它不是在保证语义精度的前提下,因此常常适得其反.

而在另一方面,为了解决主题的表达问题,也必须分析词与词之间的联系,不单是对文本中词的概率统计描述更应从语义上加以理解,此时就需要将具有语义关联性的词语抽取出来用于描述主题的内容.因此,利用关联规则挖掘构造关联词集是一个简单可行的方法,挖掘具有关联性的词语作为一个语义单元,既可以实现特征向量的降维,又可以增大主题表达的准确性.使用关联词集合可以有效地对文本特征空间的关联词进行归并,改进主题标注的效率和精度.

构造关联主题词集合需要解决2个问题:

1)同义词问题.由于中文文本存在语法修饰,不同的词汇表示相同的概念,因此,关联规则算法无法根据中文文本中的深层语义信息挖掘关联词,影响了关联词归并的质量.

2)语义相关性问题.虽然关联规则挖掘可以发现特征词的同现关系,但因为主要反映的是一种统计规律,所以存在某些规则不能很好反映特征词之间语义相关性的问题,即某些关联规则在语义上是无效的.

因此,本文采用“知网”作为概念空间,将特征词映射到概念空间,解决同义词问题,同时提出基于语义概念分类的关联规则挖掘方法来提高关联词的语义相关性.

“知网”[7](HowNet)是著名的采用汉语描述的本体论.它将汉语和英语的词语所代表的概念作为描述对象,同时描述了概念之间、概念所具有的属性之间的关系,并建立了反映这些概念和关系的知识库.“知网”中,单个或复杂的概念以及各个概念之

间、概念的属性和属性之间的关系是通过义原或义原的组合来进行标注的. 这样的好处是虽然新词不断出现,但义原的增加却极少. 因此,在“知网”中,词义就被定义为各种义原的组合.

在主题的词语概率分布模型中引入“知网”作为背景知识,将主题词映射到义原,可以在一定程度上解决同义词替换的问题,使得相同概念、不同描述的词可以进行归并.

为了获得主题信息的概念集,首先对文本集 $D=\{d_1,d_2,\cdots,d_N\}$ 进行预处理,抽取每篇文本 d_i 中权重较高的特征词,构成基于特征词集的特征向量:

$$V(d_i)=[tfi(d_i,t_1)\ tfi(d_i,t_2)\ \cdots\ tfi(d_i,t_n)\ ].$$

式中:tfi 为特征词 t_i 在网页 d_i 中出现的频率,n 为特征词的数量.

然后引入知网,将特征词映射到义原. 在将文本 d_i 中的每个特征词 t 映射为义原时,首先对具有 2 个或 2 个以上语义解释的词 t 进行语义排歧,获取其对应每个语义解释的概率 p,然后以 p 作为权重为语义解释涉及到的每个义原 a 所对应的特征向量赋值. 由于目前知网收录的词条有限,有些特征词没有被知网收录,对于这些特征词予以保留,这样就形成了义原加特征词的特征向量:

$$V(d_i)=$$
$$[w(d_i,a_1)\cdots w(d_i,a_l)\ tfi(d_i,t_1)\cdots tfi(d_i,t_i)\cdots tfi(d_i,t_k)\ ].$$

式中:$t_i(1\leqslant i\leqslant k)$ 为没有被知网收录的特征词,$w(d_i,a_i)$ 为义原 a_i 在文本 d_i 中的权值:

$$t_w(d_i,a_i)=\sum_{j=1}^{n}t_w(d_i,t_j,a_i).$$

式中:$t_w(d_i,t_j,a_i)$ 为文档 d_i 中词条 t_j 对义原 a_i 的权重贡献:

$$t_w(d_i,t_j,a_i)=|ref(t_j)\cap\{a_i\}|\times p_j\times\lambda\times tfi(d_i,t_j).$$

式中:$ref(t_j)$ 为词条 t_j 对应的义原集合,λ 为该义原类别的权重系数.

为了进一步缩减向量维度并提高关联规则挖掘的支持度和置信度,通过计算义原间的相似度[8]可以进一步将相似义原进行归并. 义原相似度的计算方法如下:

$$\mathrm{Sim}(a_1,a_2)=\frac{\alpha}{\mathrm{dis}(a_1,a_2)+\alpha}.$$

式中:$\mathrm{dis}(a_1,a_2)$ 是义原 a_1 和 a_2 在知网层次结构中的语义距离,α 是一个可调节的参数.

2.2　基于语义分类的关联词集构造

经过分析,各类事件信息中的主题词根据语义可以分为 5 类,分别反映了信息中涉及的实体对象、环境、活动、事件和结果,它们从不同角度描述了事件信息的语义内容. 因此,在建立主题词的概念空间之后,文本特征向量中的义原和主题特征词分量被分为实体对象、环境、活动、事件和结果 5 类,并且根据其概率分布,对主题语义的贡献度赋予不同的权重系数. 通过挖掘这 5 类特征分量间的关联规则,在发现关联词的同时也有助于反映它们之间的语义联系. 为了避免同类特征项出现在关联规则之中,定义基于语义分类的关联规则如下.

设文档特征空间中包含的所有义原和特征词构成集合:$W=\{a_1,a_2,\cdots,a_l,t_1,t_2,\cdots,t_k\}$,其中每个元素属于一个语义类别 K. 则定义基于语义分类的关联规则 $A\rightarrow B$,其中 $A\subset W,B\subset W,A\cap B=\varnothing$,并且,对于规则左部 A 和右部 B 中包含的任意项 u、v,满足 $K_u\neq K_v$.

对于文本集 D,规则 $A\rightarrow B$ 的支持度为 $s=P(A\cdot B)$,置信度为 $c=P(B|A)$.

基于语义分类关联规则的关联词集构造算法如下:

1)利用关联规则算法[9]挖掘基于语义分类的关联规则,获得所有支持度和置信度分别大于 s 和 c 的关联规则;

重复 2)~5),对获得的每一条关联规则的左右部包含的关联词进行归并;

2)将关联规则右部包含的主题词从主题词集合中删除;

3)在归并后的主题词集合中查找含有关联规则任一边的主题词的归并主题词组合;

4)如果找到,则将另一主题词加入到该归并主题词组合中;

5)如找不到归并主题词组合,则以关联规则左右部的 2 个主题词构造一个新的主题词组合,并放入归并后的主题词集合中去;

6)在完成所有关联规则的归并后,得到新的主题词集合,集合内包含多个关联主题词组合,即得到关联词集合.

3　关联主题词的语义相关度计算

3.1　语义相似性计算

在抽取出主题词并得到关联主题词集合后,需要从其中选择与主题语义相关性高的词作为主题模型的标注词——标签,实现对主题模型的自动语义标注. 而语义相关性计算的难点在于标签和主题模型(主题词的概率分布)之间的匹配. 因此本文将标签也以概率分布的方式表示,这样就可以直接与主题模型相比较.

假设标签 l 以语义词分布 $\{p(w|l)\}$ 来表示,则可以使用 Kullback-Leibler(KL)距离算法计算 $\{p(w|l)\}$ 与主题 $\{p(w|\theta)\}$ 之间的相似度. 为了获

得标签 l 的语义词分布 $\{p(w|l)\}$，本文采用一种近似方法，通过数据集 D 来估计 $\{p(w|l, D)\}$，以代替 $\{p(w|l)\}$. 标签和主题之间的语义相似性通过式(1)进行计算.

$$\begin{aligned}S(l,\theta) &= -d(\theta\|l) = -\sum_w p(w|\theta)\ln\frac{p(w|\theta)}{p(w|l)} = \\ &-\sum_w p(w|\theta)\ln\frac{p(w|D)}{p(w|l,D)} - \\ &\sum_w p(w|\theta)\ln\frac{p(w|\theta)}{p(w|D)} - \\ &\sum_w p(w|\theta)\ln\frac{p(w|l,D)}{p(w|l)} = \\ &\sum_w p(w|\theta)\ln\frac{p(w,l|D)}{p(w|D)p(l|D)} - d(\theta\|D) - \\ &\sum_w p(w|\theta)\ln\frac{p(w|l,D)}{p(w|l)} = \\ &\sum_w p(w|\theta)\mathrm{PMI}(w,l|D) - d(\theta\|D) + \mathrm{Bias}(l,D).\end{aligned} \tag{1}$$

从式(1)可以看出，标签 l 和主题 θ 之间的语义相似性包括3个分量，其中分量 $d(\theta\|D)$ 表示主题和情境集之间的KL距离，它对于所有的标签都是相同的，因此在进行排序时可以忽略该值的影响. 分量 $\mathrm{Bias}(l,D)$ 可以看作为通过情境数据集 D 推导标签 l 和主题 θ 之间相关性的偏差，当主题模型和候选标签均来自于集合 D 时，可以假定没有偏差. 因此，标签和主题相近度的排序主要取决于第1个分量 $\sum_w p(w|\theta)\mathrm{PMI}(w,l|D)$，它可以是给定情境下标签 l 和主题模型主题词之间互信息的期望 $\mathrm{E}_\theta(\mathrm{PMI}(w,l|D))$.

3.2 语义覆盖度计算

好的标签应该对主题的语义内容有较高的覆盖度，语义相关性仅能保证所选择的标签与主题信息具有高相关性，但可能仅表达了该主题的部分语义. 因此，当选择多个标签对主题进行标记时，希望选择的新标签能够覆盖主题其他的语义部分，而不是已有标签已经涵盖的内容.

本文采用最大边缘相关(maximal marginal relevance, MMR)方法来选择高语义覆盖度标签. MMR方法常常用于多文档摘要问题，是一种十分有效的去冗余并且取得最大相关性和差异性的方法. 本文对MMR进行了一定简化以实现对标签的选择，通过最大化MMR来逐个选择标签，如式(2)：

$$\begin{cases}\hat{l} = \arg\max\limits_{l\in L-S}[\lambda S(l,\theta) - (1-\lambda)\max\limits_{i\in s}\mathrm{Sim}(l',l)], \\ \mathrm{Sim}(l',l) = -d(l'\|l) = -\sum\limits_w p(w|l')\ln\dfrac{p(w|l')}{p(w|l)}.\end{cases} \tag{2}$$

式中：S 是已经选择的标签，λ 是经验参数.

3.3 语义区分度计算

以上标签选择方法仅考虑了对单一主题的标注，当对多个主题进行标注时，则需要考虑不同主题间的区分，因为如果一个标签在多个主题内都具有较高的相关度，则该标签对于人们区分不同的主题是缺乏帮助的，因此为多个主题选择标签既需要考虑相关度，也需要考虑区分度，在这种情况下，对式(1)进行修正，提出了考虑区分度的语义相似性计算方法：

$$\begin{cases}S'(l,\theta_i) = S(l,\theta_i) - \alpha S(l,\theta_{-i}), \\ S(l,\theta_{-i}) = -d(\theta_{-i}\|l) \approx \mathrm{E}_{\theta_{-i}}(\mathrm{PMI}(w,l|D)) \approx \\ \dfrac{1}{k-1}\sum\limits_{j=1,2,\cdots,i-1,i+1,\cdots,k}\sum\limits_w p(w|\theta_j)(\mathrm{PMI}(w,l|D)) = \\ \dfrac{1}{k-1}(\sum\limits_{j=1,2,\cdots,k}\mathrm{E}_{\theta_j}(\mathrm{PMI}(w,l|D)) - \mathrm{E}_{\theta_i}(\mathrm{PMI}(w,l|D)).\end{cases} \tag{3}$$

式中：θ_{-i} 表示除主题 θ_i 之外的其他 $k-1$ 个主题，即 $\theta_{1,2,\cdots,i-1,i+1,\cdots,k}$，$k$ 为主题数.

式(3)通过 $S'(l, \theta_i)$ 计算跨主题的标签语义相似度并进行排序，可以为多个主题生成语义相关且具有一定覆盖度和区分度的标签.

4 实验结果及分析

4.1 实验方案

实验选择旅游信息和食品安全事件信息中的4 200条文本数据构成训练文档集，采用LDA主题分析方法[10]在文本集上建立主题模型，利用快速Gibbs采样进行参数估计，设定主题数 $K=30$，超参数 $\alpha=50/K$，$\beta=0.1$，迭代次数为1 000.

采用本文提出的主题词生成方法进行主题词的提取和关联词集构造，将其作为主题标注的候选标签集，然后在候选标签集合上，采用本文提出的语义相关度计算方法选取能够描述主题语义的标签进行自动标注.

实验抽取主题词数 $N = 1\,000$，为了控制程序运行时间，设定概念空间维数为20，关联归并的支持度 $s=1\%$，置信度 $c=1.5\%$. 最后选择286个关联主题词，每个关联主题词对应1~3个主题词，构成主题的候选标签，该标签集记为TagSet-1. 同时，为了与本文的候选标签生成方法进行对比，采用N-gram方法($n=1,2$)抽取关键词，并通过 χ^2 测试选择前300个主题词建立另一个候选标签集，记为TagSet-2，从而利用这2个标签集分别进行主题标注，以评价标签集的有效性. 在食品安全和旅游信息领域采用以上2种方法分别建立的部分候选标签如表1所示.

表 1　部分候选标签

Table 1　Some candidate labels

食品安全候选标签		旅游信息候选标签	
TagSet-1	TagSet-2	TagSet-1	TagSet-2
婴儿奶粉	奶粉	游览路线	景区
记者采访	婴幼儿	进入景区	游客
医院治疗	肾结石	游客	交通
政府监管	乳制品	酒店住宿	游览路线
食物检测	新闻报道	食物制作	酒店价格
企业生产	食品安全	味道	住宿

4.2　主题标注结果

表 2 和表 3 分别列出了食品安全和旅游领域的部分主题的标注结果.

表 2　部分食品安全主题及相应标签

Table 2　Some food safety topics and corresponding labels

主题模型	人工标注	本文自动标注
医院,奶粉,患,出现,婴儿,结石,治疗,发现,食用,专家,记者,原因,孩子,时,肾结石	问题奶粉	婴儿奶粉 患 肾结石
鸡蛋,含,发现,苏丹红,中,含有,进行,鸭蛋,蛋,问题,超市,食品,健康,时,报道	红心鸭蛋	蛋 含有 苏丹红
销售,生产,还有,造成,我们,事件,标准,作为,有毒,发生,需要,部门,危害,应该,管理	政府监管	有关部门 标准 管理

表 3　部分旅游主题及相应标签

Table 3　Some tourism topics and corresponding labels

主题模型	人工标注	本文自动标注
门票,气温,参观,园,人,气候,直接,路线,之间,票,购买,进入,值得,地区,不宜	景区游览	游客 游览路线 购买门票
菜,味,小,肉,风味,制作,味道,茶,中,时候,食品,食,吃,品尝,香	地方美食	食物制作 品尝 风味
景区,酒店,游客,旅游,价格,宾馆,市区,当地,方便,他们,很多,最好,附近,选择,景点	景区住宿	酒店住宿 游客 价格

表 2、3 中列出了每个主题模型中概率最大的前 15 个词,以及根据本文方法自动标注的标签. 为了便于比较,表中也给出了每个主题模型的人工标注标签. 人工标注的具体方法是将每一主题的主题模型(主题词概率分布)、代表性文档及候选标签集展示给志愿者,由他们选择合适的标签进行人工标注.

可以看出,自动标注的标签基本涵盖了主题的语义,尤其在食品安全领域,例如“婴儿奶粉”、“患”、“肾结石”、“蛋”、“含有”、“苏丹红”等标签已经很好地表达了主题语义,与“问题奶粉”、“红心鸭蛋”等人工标注结果较为吻合. 某些情况下比人工标注还要准确,例如志愿者因受媒体报道等的影响,将禽蛋类食品中发现苏丹红的主题标注为“红心鸭蛋”,这是因为最早发现苏丹红是在鸭蛋中,所以媒体将此类事件报道为“红心鸭蛋事件”;而实际上,主题模型中包括鸡蛋和鸭蛋,本文标注方法将它们映射为义原“蛋”并据此生成标签,因此语义上更为准确.

4.3　主题标注的有效性

为了能够准确评价主题标注的有效性,采用评分法将本文标注方法与人工标注和最大概率主题词标注方法进行比较. 其中,最大概率主题词标注根据主题模型中词语的概率分布选择概率最高的前 3 个词作为主题标签.

标注结果的具体评分方法是:通过 5 名志愿者对 3 种方法的标签进行打分,即将随机排序的主题及其主题词分布、标签和该主题的最相关文档提供给志愿者,由志愿者对 3 种方法产生的标签分别打分,然后统计平均得分. 打分规则是总分为 5 分,由志愿者将这 5 分按照其对标签准确性的评估分别分配给 3 种方法生成的标签. 并且,要求志愿者对仅使用 1 个标签和使用 3 个标签进行标注的情况分别打分,结果如表 4 所示.

表 4　主题标注的有效性对比

Table 4　Comparison of topic labeling methods

数据集	标签数	人工标注	最大概率主题词标注	本文自动标注
食品安全	1	2.16	0.89	1.94
	3	1.92	1.41	1.67
旅游	1	3.04	0.76	1.22
	3	2.28	0.97	1.76

从表 4 中可以看出,虽然在所有情形下人工标注的得分都是最高的,但本文标注方法的得分明显高于最大概率主题词标注方法. 在食品安全领域,本文方法已经接近于人工标注的得分,这主要是因为在食品安全领域中,不同主题的主题词之间具有更高的区分度,尤其是一些专有名词和术语主要在特定主题中出现.

此外,在食品安全领域仅采用 1 个标签的情况下,本文方法相比最大概率标注方法优势明显,但若采用 3 个标签,则优势不大. 然而在旅游领域使用 3

个标签的情况下,本文方法仍具有较大优势,这主要是因为旅游领域除特定地点或景点主题外,主题词多是一些通用词,且某些高概率词的语义类别单一,并不能充分表达主题语义.而通过概念映射和建立关联词,则可以将属于不同语义类别且具有语义相关性的主题词组织起来,从而提供更为丰富的语义.例如,"菜"、"肉"、"茶"等主题词被映射为概念"食物",与关联词"制作"共同构成标签"食物制作",这样可以表达更明确的语义.

为了比较不同标签集生成方法的有效性,采用本文提出的语义相似性计算方法,分别利用 TagSet-1 和 TagSet-2 2 个候选标签集对主题进行标注,并对标注结果打分,总分 2 分,评分结果如表 5 所示.

表 5 标签集的有效性对比

Table 5 Comparison of tag sets

数据集	标签数	TagSet-1	TagSet-2
食品安全	1	1.08	0.92
	3	1.12	0.88
旅游	1	1.30	0.70
	3	1.22	0.78

从表 5 可以看出,本文方法建立的关联词集 TagSet-1 在总体得分上均高于 N-gram 关键词集 TagSet-2,这主要是因为 TagSet-2 中存在的多个同义或同语义类别词分散了语义相似度的计算结果,如"鸭蛋"、"鸡蛋"和"禽蛋"被作为 3 个标签分别计算语义相似度,导致计算结果偏低,影响了标签的选择.而且,TagSet-2 中的标签也存在语义类别单一的问题,降低了每个标签的语义表达能力.

5 结束语

提出了一种概率主题模型的自动标注方法,通过主题词提取和语义概念空间上的关联词挖掘方法来生成候选主题词,并且给出了主题词语义相关性计算以及高语义覆盖度和区分度标签的选择方法,实现了主题模型的自动语义标注,解决了对主题词模型进行语义理解的问题.该方法被用于食品安全主题和旅游信息主题的自动标注,实验证明该方法的标注效果优于最大概率主题词标注方法.尤其在食品安全等专业领域,由于充分考虑了专业术语与一般词汇的语义区分度和语义覆盖度,使得本文方法能够取得更好的效果

参考文献:

[1] BLEI D M, NG A Y, JORDAN M I, et al. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(7): 993-1022.

[2] COHN D, HOFMANN T. The missing link—a probabilistic model of document content and hypertext connectivity [EB/OL]. [2010-05-10]. http://books.nips.cc/nips13.html.

[3] GILDEA D, JURAFSKY D. Automatic labeling of semantic roles[J]. Computer Linguist, 2002, 28(3): 245-288.

[4] 石晶,李万龙. 基于 LDA 模型的主题词抽取方法[J]. 计算机工程, 2010, 36(19): 81-83.
SHI Jing, LI Wanlong. Topic words extraction method based on LDA model [J]. Computer Engineering, 2010, 36(19): 81-83.

[5] BANERJEE S, PEDERSEN T. The design, implementation, and use of the ngram statistics package [C]//Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. Mexico City, Mexico, 2003: 370-381.

[6] 刘铭,王晓龙,刘远超. 基于词汇链的关键短语抽取方法的研究[J]. 计算机学报, 2010, 33(7): 1246-1255.
LIU Ming, WANG Xiaolong, LIU Yuanchao. Research of key-phrase extraction based on lexical chain [J]. Chinese Journal of Computers, 2010, 33(7): 1246-1255.

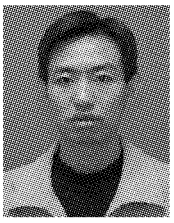
[7] 孙景广,蔡东风,吕德新,等. 基于知网的中文问题自动分类[J]. 中文信息学报, 2007, 21(1): 90-95.
SUN Jingguang, CAI Dongfeng, LÜ Dexin, et al. HowNet based Chinese question automatic classification [J]. Journal of Chinese Information Processing, 2007, 21(1): 90-95.

[8] 夏天. 汉语词语语义相似度计算研究[J]. 计算机工程, 2007, 33(6): 191-194.
XIA Tian. Study on Chinese words semantic similarity computation[J]. Computer Engineering, 2007, 33(6): 191-194.

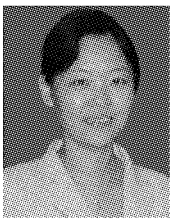
[9] 黄名选,严小卫,张师超. 基于矩阵加权关联规则挖掘的伪相关反馈查询扩展[J]. 软件学报, 2009, 20(7): 1854-1865.
HUANG Mingxuan, YAN Xiaowei, ZHANG Shichao. Query expansion of pseudo relevance feedback based on matrix-weighted association rules mining[J]. Journal of Software, 2009, 20(7): 1854-1865.

[10] 石晶,范猛,李万龙. 基于 LDA 模型的主题分析[J]. 自动化学报, 2009, 35(12): 1586-1592.
SHI Jing, FAN Meng, LI Wanlong. Topic analysis based on LDA model [J]. Acta Automatica Sinica, 2009, 35(12): 1586-1592.

作者简介:



周亦鹏,男,1976 年生,讲师. 主要研究方向为人工智能和 Web 挖掘.



杜军平,女,1963 年生,教授,博士生导师. 主要研究方向为人工智能和数据挖掘,承担国家"863"、"973"计划、国家自然科学基金、北京市自然科学基金项目等多项,发表学术论文 150 余篇.

中国人工智能学会粗糙集与软计算专业委员会

中国人工智能学会粗糙集与软计算专业委员会(简称"专委会")以学术活动为中心开展工作,每年组织各种学术活动,推动中国粗糙集与软计算理论研究及其应用发展。

专委会现任主任委员是王国胤教授(重庆邮电大学),副主任委员是苗夺谦教授(同济大学)、吴伟志教授(浙江海洋学院)和梁吉业教授(山西大学),秘书长是张清华教授(重庆邮电大学)。现在已有委员 128 人,会员从最初的 100 多人逐渐增加到目前的 400 余人。

举办全国学术会议

◆中国 Rough 集与软计算学术会议。专委会 2001 年在重庆邮电学院成功召开"第一届中国 Rough 集与软计算学术研讨会(CRSSC2001)",此后,分别在重庆、苏州、舟山、鞍山、金华、太原、新乡、石家庄、重庆、南京和合肥等地成功举办了 11 届粗糙集与软计算全国学术会议。

◆中国 Web 智能学术研讨会和中国粒计算学术研讨会联合学术会议(CRSSC-CWI-CGrC)。2007 年在山西大学成功召开了第一届中国 Web 智能学术研讨会(CWI2007)及第一届中国粒计算学术研讨会(CGrC2007)(与 CRSSC2007 会议)联合召开),从 2008 年至今,分别在新乡、石家庄、重庆、南京和合肥等地召开了 6 届中国粒计算学术研讨会和中国 Web 智能学术研讨会(与 CRSSC 会议联合召开,即 CRSSC-CWI-CGrC)。

◆ 国内学术专题研讨会

专委会不定期组织专题学术研讨会和暑期讨论班,研讨一些当前的研究热点问题。

2010 年在合肥召开了"商空间与粒计算"研讨会;

2011 年在上海召开了"不确定性与粒计算"研讨会;

2012 年 8 月在北京召开了"云模型与粒计算" 研讨会。

举办国际学术会议

◆RSKT 国际学术会议

为了加强与国际学术界的交流与融合,促进本领域研究工作的国际化,专委会发起召开了粗糙集与知识技术(RSKT)系列国际学术会议。2006 年在重庆组织召开了 The 1st Int. Conf. on Rough Sets and Knowledge Technology (RSKT2006);此后,分别在加拿大多伦多、成都、澳大利亚黄金海岸、北京和加拿大班夫等地召开,形成了在中国和其他国家轮流召开的机制。

◆IFTGrCRSP2006 和 IFKT2008 等国际论坛

为了及时研讨一些新兴热点研究问题,专委会适时组织相关学术论坛。2006 年在南昌召开了 Int. Forum on Theory of GrC from Rough Set Perspective (IFTGrCRSP2006); 2008 年在重庆召开了 2008 Int. Forum on Knowledge Technology (IFKT2008)。

出国参加学术会议

专委会自成立以来,定期组织会员赴国外参加学术会议,促进国际学术交流与合作。

2004 年王国胤教授、刘清教授、黄厚宽教授赴瑞典出席 RSCTC2004 国际学术会议,同年王国胤教授还应邀赴波兰出席 MSRAS2004 国际学术会议;

2005 年组织 14 名中国学者赴加拿大出席了 RSFDGrC2005 国际学术会议;

2007 年组织中国学者参加了在加拿大多伦多召开的国际粗糙集联合学术会议(JRS2007),同年组织国内学者参加了在美国硅谷召开的国际粒计算学术会议(GrC2007);

2008 年组织中国学者出席在美国召开的粗糙集与当前计算趋势国际学术会议;

2009 年组织国内学者参加了在澳大利亚黄金海岸召开的第 4 届粗糙集与知识技术国际学术会议(RSKT2009);

2010 年组织国内学者参加了在波兰华沙召开的 The Seventh International Conference on Rough Sets and Current Trends in Computing (RSCTC 2010);

2011 年组织 10 多名中国学者赴加拿大班夫出席了 RSKT2011 国际学术会议和在俄罗斯莫斯科召开的 RSFDGrC2011 国际学术会议。

本期执行编辑：马兰兰

★ 国家自然科学基金资助期刊
★ 工业和信息化部优秀期刊
★ 中国高校特色科技期刊
★ 中文核心期刊
★ 中国科技核心期刊
★ 美国《剑桥科学文摘》收录期刊
★ 波兰《哥白尼索引》收录期刊
★ 英国《科学文摘》收录期刊

智 能 系 统 学 报

Zhineng Xitong Xuebao

(双月刊,2006 年创刊)

第 7 卷第 4 期(总第 36 期)2012 年 8 月

CAAI Transactions on Intelligent Systems

(Bimonthly,started in 2006)

Vol.7 No.4(Sum. 36) Aug. 2012

主管单位:中华人民共和国工业和信息化部
主办单位:中国人工智能学会
哈尔滨工程大学
编辑出版:《智能系统学报》编辑部
哈尔滨市南岗区南通大街 145-1 号楼(150001)
联系电话:0451-82518134
主　　编:钟义信
印刷单位:黑龙江龙江传媒有限责任公司
订　　阅:全国各地邮局
国内发行:哈尔滨市邮政局
国外发行:中国国际图书贸易总公司(北京 399 信箱)

Supervised by Ministry of Industry and Information Technology of the People's Republic of China
Sponsored by Chinese Association for Artificial Intelligence and Harbin Engineering University
Edited and Published by Editorial Department of CAAI Transactions on Intelligent Systems
Add: 1st Building 145 Nantong Street of Nangang District,Harbin 150001,China
Tel: 86-451-82518134
Chief Editor: ZHONG Yixin
E-mail: tis@vip.sina.com
Distributed by China International Book Trading Corporation (P.O.Box 399,Beijing,China)

连续出版物号:ISSN 1673-4785 / CN 23-1538/TP　　邮发代号:14-190　　国外发行代号:BM4940　　国内定价:15.00 元