基于上下文语义的社交网络用户人格预测

王江晴,陈思敏,刘晶,孙翀,毕建权

(中南民族大学 计算机科学学院,湖北省制造企业智能管理工程技术研究中心,武汉 430074)

摘 要 在利用文本信息预测用户大五人格的普遍方法中,对于文本特征的提取未充分考虑上下文语义信息,存在对语义特征提取不够精准的问题.针对该问题,提出了一种结合深度学习与上下文语义的方法:在 TF-IDF 中加入单词的上下文语义信息来计算单词权值,然后结合基于文本的卷积神经网络模型和由单词权值构成的上下文语义特征向量进行用户大五人格预测.实验数据使用 Facebook 中 myPersonality 应用的用户社交记录,实验结果表明:将文本上下文语义加入到深度学习预测模型后,人格预测的准确率有所提高.

关键词 人格预测;大五人格;上下文语义;深度学习;社交网络

中图分类号 TP391 **文献标志码** A **文章编号** 1672-4321(2020)03-0289-06

doi:10.12130/znmdzk.20200312

引用格式 王江晴,陈思敏,刘晶,等.基于上下文语义的社交网络用户人格预测[J].中南民族大学学报(自然科学版),2020,39(3):289-294.

WANG Jiangqing, CHEN Simin, LIU Jing, et al. Social network user's personality prediction based on context semantics [J].Journal of South-Central University for Nationalities(Natural Science Edition), 2020,39(3):289-294.

Social network user's personality prediction based on context semantics

WANG Jiangqing, CHEN Simin, LIU Jing, SUN Chong, BI Jianquan

(College of Computer Science,South-Central University for Nationalities,Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprises,Wuhan 430074,China)

Abstract In the general methods of using text information to predict user's Big Five personality, the extraction of text features does not fully consider context semantics information, and there is a problem that semantic features are not extracted accurately enough. Aiming at this problem, a method combining deep learning and context semantics is proposed. The context semantics information of words is added to TF-IDF to calculate word weight, and then combining text-based convolutional neural network model and context semantics feature vector composed of word weight to predict user's Big Five personality. The experimental data uses the user's social record of the myPersonality application in Facebook, and the experimental result shows that accuracy of user's Big Five personality prediction is improved after the text context semantics is added to the deep learning prediction model.

Keywords personality prediction; Big Five personality; context semantics; deep learning; social network

随着社交网络用户日益增多,网络用户行为已经成为社交网络领域重要的研究内容.人格是一种心理结构,旨在从一些稳定和可衡量的个体特征方面解释各种各样的人类行为[1].人格特质作为体现用户行为的重要因素,影响着人们的行为选择和习惯偏好,对社交网络用户的人格特质预测有许多重要的实际应用和研究价值.例如,在个性化推荐背景下,相似人格特质的人喜爱的产品也会高度相似[2];在心理问诊方面,心理疾病与人格特质存在一定的内在联系.在心理学领域,用来衡量一个人人

收稿日期 2019-09-30

作者简介 王江晴(1964-),女,教授,博士,研究方向:智能算法,E-mail:wjqing2000@ mail.scuec.edu.cn

基金项目 国家科技支撑计划项目(2015BAD29B01);湖北省技术创新专项重大项目(2019ABA101);湖北省自然科学基金资助项目(2017CFC886);中央高校基本科研业务费专项资金资助项目(YZY18002);中南民族大学研究生学术创新基金资助项目(2019sycxjj122)

格的最主流的模型是大五人格模型[3],大五人格模型从外向性(EXT)、神经质(NEU)、宜人性(AGR)、责任心(CON)以及开放性(OPN)等五个维度来分析和描述一个人的人格特质.

已有相关研究从社交网络文本中挖掘出一个人的人格特质与行为活动之间的潜在关系,验证了利用社交网络文本识别与预测用户大五人格的可行性[4-6].基于文本的用户大五人格预测主要工作有用户文本特征提取和分类模型构造.

大部分大五人格研究者使用到的文本特征提取方法有 LIWC(Linguistic Inquiry and Word Count)、词袋模型[7]、TF-IDF[8]等.这些方法提取到的文本特征仅仅停留在词集的层面,很少对文本语义做研究.而文本的语义信息往往才是全面描述当前用户所要表达信息的载体,因此,我们认为分析文本潜在语义信息,从文本语义层面出发研究用户的大五人格,能更准确地挖掘出用户的人格信息.然而这些文本特征提取方法没有考虑社交短文本的上下文语义信息,使得对语义特征的提取不够精准,可能忽略掉很多文本关键信息,所以我们针对特征提取方法 TF-IDF,引入上下文词语的共现关系来提取更多的语义信息.

在自然语言处理(NLP)研究工作中,与传统的机器学习方法相比,近几年广泛利用分布式表示[9]和深度学习的方法来分析和挖掘文本信息,其效果突出.深度学习的模型在基于文本的大五人格分类和预测工作中也逐渐被应用.MAJUMDER 等[10]提出了一种使用 CNN 从意识流文章中提取人格特质的方法,提高了人格预测模型的精确度.WEI 等[7]使用了社交网络的异质信息包括文本、用户头像、表情符号、用户交互信息来预测大五人格,其中文本信息特征的提取,结合了词袋聚类、LIWC 和 CNN 等方法,对关键词只统计了词频,没有考虑上下文语义的有关信息,使得特征权重分配不佳.还有一些研究者使用了 RNN[11,12]及其变种等方法作为预测模型,其结果与 CNN 相差不大.由于 RNN 模型计算步骤之间有前后依赖关系,并行程度不高,而 CNN 的所有卷积都可以并行执行,相比 RNN 并行程度更高,效率更快,而且容易捕捉到一些全局的结构信息,关键性短语在句子编码过程中能保持含义不变性,因此本文采用基于文本的卷积神经网络模型(Text-CNN),结合上下文语义特征向量来对用户文本进行训练以预测用户的大五人格.实验结果证明引入上下文语义信息后的模型在预测准确率上有一定的提高.

1 模型描述

1.1 结合上下文语义信息的社交文本特征提取

主流的文本特征提取方法 TF-IDF 没有考虑特征词之间的语义联系,使得提取的特征词表示文本语义强度不佳,为解决该问题,本文在 TF-IDF 计算过程中加入了上下文语义信息.

用户文本集表示为 $D=\{d_j \mid j=1,2,\cdots,N\}$,$N$ 是用户文本集中的文本总数,词汇表表示为 $V=\{t_i \mid i=1,2,\cdots,M\}$,$M$ 是词汇表中的特征词总数,统计用户文本集中的所有单词得到词汇表.

首先计算文本中每个特征词的 TF-IDF[8]值,表示为:

$$tf\text{-}idf_{i,j}=tf_{i,j}\cdot idf_i,$$

其中,$tf\text{-}idf_{i,j}$表示单词 t_i 在文本 d_j 中的 TF-IDF 值,其中 $tf_{i,j}$表示为:

$$tf_{i,j}=\frac{n_{i,j}}{n_j}$$

其中,$tf_{i,j}$表示单词 t_i 在文本 d_j 中的词频.$n_{i,j}$是单词 t_i 在文本 d_j 中出现的次数,n_j是词汇表中所有单词在文本 d_j 中出现的次数之和.idf_i 表示为:

$$idf_i=\log\frac{\mid N\mid}{1+\mid\{j:t_i\in d_j\}\mid},$$

其中,idf_i 表示单词 t_i 的逆向文本频率,$j:t_i\in d_j$ 是包含单词 t_i 的文本个数.

然后统计词汇表中特征词 t_a 与特征词 $t_b(b\neq a)$ 同时出现在用户文本集的文本条数,如果文本条数不小于 2,则 t_a 与 t_b 是一对共现词对[13],记为 $t_{a,b}$,此时的文本条数代表 $t_{a,b}$ 的出现频率,记为 $ft_{a,b}$. 根据 $ft_{a,b}$ 计算单词的上下文语义值,公式为:

$$sw_{a,j}=\sum_{b=1}^{p} tf\text{-}idf_b\cdot\frac{ft_{a,b}}{N},$$

其中,$sw_{a,j}$表示文本 d_j 中单词 t_a 的上下文语义值,$tf\text{-}idf_b$ 是文本 d_j 中单词 t_b 的 TF-IDF 值.

最后由单词的上下文语义值和 TF-IDF 值计算出文本 d_j 中每个词的权值,公式为:

$$tw_{i,j}=\alpha\cdot tf\text{-}idf_i+(1-\alpha)\cdot sw_{i,j},$$

其中,$tw_{i,j}$表示文本 d_j 中单词 t_i 的权值,α 为权重.

结合上下文语义信息的词权值计算的时间复杂度分析如下:首先,计算特征词的 TF-IDF 值的时间复杂度为 $O(n)$;其次,计算特征词-特征词共现词对矩阵的时间复杂度为 $O(n^2)$;然后,计算单词的上下文语义值的时间复杂度为 $O(n)$;最后,计算文本中

每个词的最终权值的时间复杂度为 $O(1)$.综上,结合上下文语义信息的词权值计算的时间复杂度为$O(n^2)$.

1.2　基于 Text-CNN 的人格预测模型

上下文语义信息是人工提取的特征,与深度学习预测模型自动提取的特征相比,特征之间表达的含义不同,在预测模型中加入上下文语义特征,人格相关潜在特征得以丰富,从而达到优化预测效果的目的.为验证在预测模型中加入上下文语义信息是否能提高大五人格预测的准确率,我们选取 Text-CNN 作为人格预测模型,模型架构如图 1 所示,将卷积和池化操作得到的抽象特征向量与 1.1 节结合了上下文语义的特征向量连接后,送到全连接层以及输出层进行人格分类.

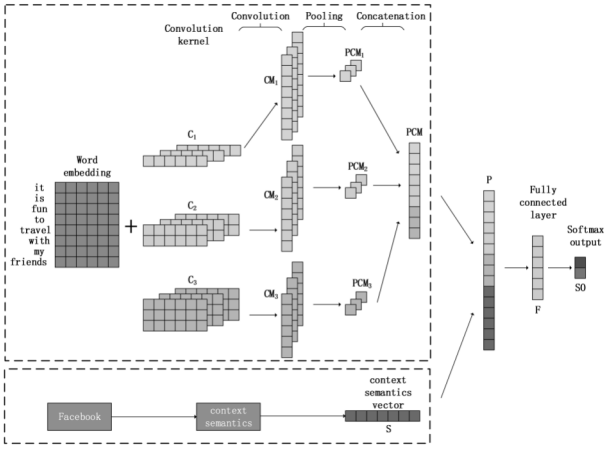

图 1　Text-CNN+context semantics 模型架构

Fig.1　Text-CNN+context semantics model architecture

模型使用到的变量定义如表 1 所示.

表 1　符号表

Tab.1　Symbol table

符号	描述
W	表示句子长度
E	表示词嵌入的长度
n	表示卷积核的核宽
K	表示卷积核的个数
V	表示词汇表中单词总数
A	表示输入的句子矩阵
C	表示卷积核
CM	表示卷积后的特征向量
PCM	表示池化后的特征向量
S	表示结合了上下文语义的特征向量

输入层:输入的句子通过对字典的 lookup 生成句嵌入,即二维矩阵,每一行表示为单词的词向量.因此,输入是一个数组 $A^{W\times E}$.

卷积层:卷积核定义为 $C_n \in A^{K\times n\times E}$,$n = 1,2,3$.卷积窗口从句子矩阵最上方开始向下滑动直到句子结尾,每次滑动考虑 n 个单词,通过卷积计算得到该句子的特征映射 $CM_n \in A^{K\times(W-n+1)\times 1}$,激活函数为 ReLU.

池化层:对 CM_n 做平均池化操作,得到输出特征向量 $PCM_n \in A^K$,将所有的 PCM_n 连接得到最终的池化结果 $PCM \in A^{(K\times n)}$.

结合了上下文语义的特征向量:对于输入的每个句子,其结合了上下文语义信息计算得到的特征向量为 $s_j=(tw_{1,j},tw_{2,j},\cdots,tw_{V,j})$,$s_j \in A^V$.将 s_j 与 PCM 连接作为下一阶段的输入向量 P,P 的计算公式为:

$$P = concat(PCM, s_j), P \in A^{(K\times n+V)}.$$

全连接层:将 P 与两个全连接层矩阵做运算,得到更深层的特征表示,如图 1 列向量 F.

输出层:使用 softmax 函数对最后的人格结果进行预测,得到二分类结果如图 1 列向量 SO.损失函数的计算公式如下:

$$loss = -\sum_i y'_i \log(y_i),$$

其中,y_i' 是该人格预测的概率值,y_i 是人格的实际值.

对于大五人格的五维人格特质,我们训练 5 个独立的 Text-CNN 模型,它们的网络结构一致.

2　实验及分析

2.1　数据集

实验采用 Facebook 中 myPersonality 应用的公共数据集.myPersonality 中包括 essay 和 Facebook 用户文本,发表这些文本的用户已经填写了大五人格量表问卷并得到大五人格的评测结果,这些文本已标注用户大五人格类别.我们通过人格识别计算研讨会的共享任务[14]获得 Facebook 的用户文本数据.其中 80%的数据集用于训练,剩下 20%用于测试.

2.2　文本预处理

在自然语言处理中,文本分类结果的好坏,一方面取决于分类器的好坏,另一方面与文本前期的预处理工作有很大关系.文本的处理步骤如下:

1)去掉文本中的邮箱地址和网址.这些信息与人格特征关系不大;

2)拼写检查更正.使用 pyenchant 类库检查单词拼写,找出错误后,根据需要来更正;

3)缩写词还原.如“I'm”还原成“I am”;

4)将单词转化为小写,并引入停用词表删除一些无效字符,以降低词汇表的维度;

5)去除数字和一些标点符号,并保留如“!!!”、“!!!!!!”等标点符号,因为这些重复的符号是用户用来强调情绪的直接表现.同理我们还保留了如“yayayaya”、“freeeeee”、“ahhhhhh”等含重复字母的单词;

6)词形还原.一个单词会有单数、复数和时态等多种不同的形式.我们使用自然语言处理工具(nltk)将文本中的单词还原成原形形式,从而生成最终的词汇表.

2.3 实验参数设置

通过不断调整超参数来降低随机梯度,以使训练的模型最佳.对输入的句嵌入的向量维度、词嵌入的向量维度、卷积核的核宽以及每种卷积核的个数等进行设置.对于训练,每迭代 100 次进行一次验证,并保存结果.表 2 展示了实验设置的超参数.

特别地,对于实验参数 Batch_size,表示一批训练数据的文本条数,取值范围为{20,30,40,50},选 20 至 50 之前,用更大范围的数值训练过模型,发现在 20 至 50 区间效果最好,所以在这个区间更细粒度地训练了模型.其中每一种取值测试 20 组数据,共测试 80 组,训练五个人格维度模型则为 400 组.结果如表 3 所示,展示了每种取值下 Text-CNN+context semantics 各人格维度模型预测准确率的平均值和最高值.表 4 展示了未加入上下文语义时 Text-CNN 模型预测准确率的平均值和最高值.我们将预测准确率最高时的 Batch_size 取值作为最终生成的模型的 Batch_size 值,即得到的 Text-CNN+context semantics 五个人格维度模型的 Batch_size 取值分别为{20,50,50,20,20},Text-CNN 五个人格维度模型的 Batch_size 取值分别为{40,50,50,50,30}.

表 2 实验参数设置

Tab.2 Experimental parameters setting

参数名	参数值
Word embedding	16
Sentences dimension	50
Convolution filter	1,2,3
Num of filter	200
Learning rate	0.001
Dropout probability	0.8

表 3 Batch_size 取不同值时 Text-CNN+context semantics 模型预测的准确率

Tab.3 Accuracy of Text-CNN+context semantics model prediction when Batch_size takes different values

Batch_size	EXT		NEU		AGR		CON		OPN	
	平均值	最大值	平均值	最大值	平均值	最大值	平均值	最大值	平均值	最大值
20	0.596	0.624	0.557	0.568	0.591	0.608	0.587	0.607	0.667	0.702
30	0.565	0.611	0.552	0.574	0.593	0.605	0.585	0.605	0.623	0.688
40	0.597	0.615	0.557	0.575	0.590	0.613	0.582	0.600	0.654	0.668
50	0.565	0.600	0.547	0.580	0.596	0.618	0.583	0.600	0.646	0.656

表 4 Batch_size 取不同值时 Text-CNN 模型预测的准确率

Tab.4 Accuracy of Text-CNN model prediction when Batch_size takes different values

Batch_size	EXT		NEU		AGR		CON		OPN	
	平均值	最大值	平均值	最大值	平均值	最大值	平均值	最大值	平均值	最大值
20	0.588	0.610	0.553	0.572	0.573	0.593	0.561	0.591	0.635	0.669
30	0.583	0.622	0.554	0.573	0.572	0.597	0.572	0.592	0.652	0.700
40	0.584	0.623	0.542	0.561	0.577	0.598	0.569	0.592	0.650	0.663
50	0.590	0.616	0.554	0.578	0.582	0.614	0.572	0.599	0.632	0.678

2.4 评估指标

本文以准确率(Accuracy)来评估实验结果的好坏,其公式为:

$$Accuracy = \frac{\text{预测对的样本数}}{\text{样本总数}}.$$

2.5 实验结果分析比较

本节将讨论模型训练中的收敛情况,以及 5 个人格维度上的卷积神经网络模型在引入上下文语义后,预测准确率上的差别.

图 2 给出了引入上下文语义后,开放型人格(OPN)维度上的 Text-CNN+context semantics 模型在训练过程中损失率和准确率的变化折线图.以 OPN 维度上的 Text-CNN+context semantics 模型为例,可以看出模型随着训练步数的增长,准确率逐渐增加,

损失函数逐渐减小,在 3000 步左右的时候模型趋于收敛.

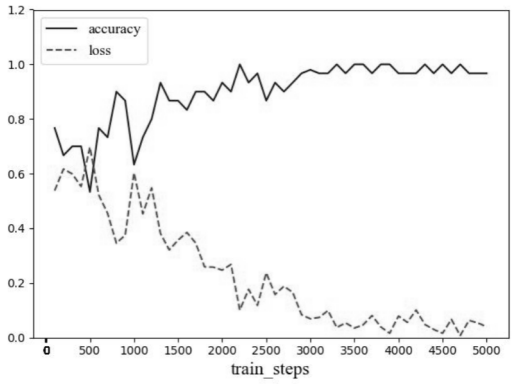

图 2　Text-CNN+context semantics/OPN 上的损失率和准确率

Fig.2　Text-CNN+context semantics/loss and accuracy of OPN

图 3 给出了引入上下文语义之前,开放型人格维度上的 Text-CNN 模型训练过程中损失率和准确率的变化折线图.可以看出 Text-CNN 模型在训练步数的增长时准确率的增加以及损失函数的减小,在 3800 步左右的时候趋于收敛.其他 4 个人格维度上的两种模型对比也有类似结果.经过比较可以看出,Text-CNN+context semantics 模型,在参数相同的情况下,模型收敛的速度要快于 Text-CNN 模型,因为加入上下文语义后,模型学习到有关人格特质的特征速度更快.

WEI[7] 和 MAJUMDER[10] 在预测用户大五人格时均使用了 Text-CNN 模型,为了验证实验中加入了上下文语义信息的效果,我们与 Text-CNN 模型进行比较.表 5 展示了本文方法与 Text-CNN 模型、文献[8]的 SMO 算法以及文献[15]的全连接架构在用户大五人格 5 个维度上的预测准确率.

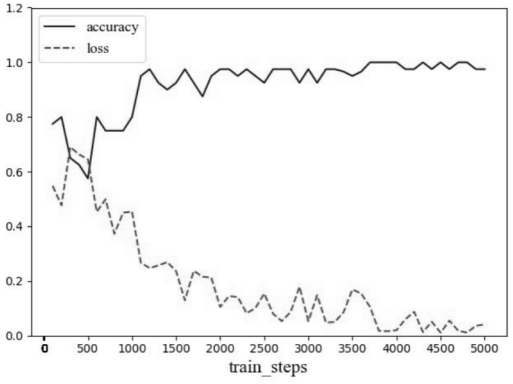

图 3　Text-CNN/OPN 上的损失率和准确率

Fig.3　Text-CNN/loss and accuracy of OPN

可以看到,在五个人格维度上的准确率,Text-CNN+context semantics 模型均比 Text-CNN 模型要高,Text-CNN + context semantics 模型在外向型(OPN)人格维度上的准确率最高达到 70.2%,模型预测准确率相对较高的原因在于加入上下文语义后,提取的文本语义特征更加丰富,模型学习到的有关人格特质的特征更多,模型更精准;同时,本文方法预测大五人格准确率仅在神经质型人格(NEU)上的准确率比 SMO 低 1.33%,但整体上的准确率比 SMO 以及只使用全连接层的神经网络要高.

表 5　不同模型准确率对比

Tab.5　Comparison of accuracy of different models

方法	EXT	NEU	AGR	CON	OPN
Text-CNN+context semantics	0.624	0.580	0.618	0.607	0.702
Text-CNN	0.623	0.578	0.614	0.599	0.700
SMO 算法[8]	0.582	0.593	0.585	0.582	0.658
全连接架构[15]	0.518	0.530	0.497	0.498	0.645

3　总结与展望

传统的利用文本信息来分析和预测大五人格的方法中,对于文本特征的提取阶段,未充分考虑上下文语义,语义特征的提取不够精准,会忽略掉很多文本关键信息,本文针对此问题引入短文本上下文的共现词对,结合上下文语义权重向量与 Text-CNN 模型,得到 Text-CNN+context semantics 模型来预测用户大五人格,实验结果表明本文的方法在准确率上有所提高.对于加入上下文语义前后,模型最佳时的参数 Batch_size 在不同人格维度上的取值不同,后续工作会继续增加 Batch_size 各个取值训练的次数,以探究其原因.未来我们会考虑将提取的上下文语义加入到其他深度学习模型如 RNN、长短期记忆网络(LSTM)中,验证上下文语义结合到预测模型中的通用性.

参 考 文 献

[1] VINCIARELLI A, MOHAMMADI G. A survey of personality computing[J]. IEEE Transactions on Affective Computing, 2014, 5(3): 273-291.

[2] 张磊, 陈贞翔, 杨波. 社交网络用户的人格分析与预测[J]. 计算机学报, 2014,37(8): 1877-1894.

[3] COSTA P T, MCCRAE R R. The revised neo personality inventory[J]. The SAGE Handbook of Personality Theory and Assessment, 2008, 2(2): 179-198.

[4] SCHWARTZ H A, EICHSTAEDT J C, KERN M L, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach[J]. PloS One, 2013, 8(9): e73791.

[5] QIU L, LU J, RAMSAY J, et al. Personality expression in Chinese language use[J]. International Journal of Psychology, 2017, 52(6): 463-472.

[6] LI Y, ZHU T, LI A, et al. Web behavior and personality: a review[C]//IEEE.2011 3rd Symposium on Web Society. Port Piscataway: IEEE, 2011: 81-87.

[7] WEI H, ZHANG F, YUAN N J, et al. Beyond the words: Predicting user personality from heterogeneous information[C]//ACM. Proceedings of the 10th ACM International Conference on Web Search and Data Mining. New York: ACM, 2017: 305-314.

[8] ALAM F, STEPANOV E A, RICCARDI G. Personality traits recognition on social network-facebook[C]//AAAI. 7th International AAAI Conference on Weblogs and Social Media. Massachusetts: AAAI, 2013.

[9] SRIDHAR V K R. Unsupervised text normalization using distributed representations of words and phrases[C]// NAACL HLT.Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. Denver: NAACL HLT, 2015: 8-16.

[10] MAJUMDER N, PORIA S, GELBUKH A, et al. Deep learning-based document modeling for personality detection from text[J]. IEEE Intelligent Systems, 2017, 32(2): 74-79.

[11] SUN X, LIU B, CAO J, et al. Who am I? Personality detection based on deep learning for texts[C]// IEEE. 2018 IEEE International Conference on Communications. Port Piscataway: IEEE, 2018:1-6.

[12] XUE D, WU L, HONG Z, et al. Deep learning-based personality recognition from text posts of online social networks[J]. Applied Intelligence, 2018, 48(11): 4232-4246.

[13] 张群, 王红军, 王伦文. 一种结合上下文语义的短文本聚类算法[J]. 计算机科学, 2016, 43(S2): 443-446.

[14] CELLI F, LEPRI B, BIEL J I, et al. The workshop on computational personality recognition 2014[C]//ACM. Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM,2014:1245-1246.

[15] YU J, MARKOV K. Deep learning based personality recognition from facebook status updates[C]//IEEE. 2017 IEEE 8th International Conference on Awareness Science and Technology. Port Piscataway: IEEE, 2017: 383-387.

(责任编辑 曹东)