

Chatbot Topic Recommendation Based on Collaborative Filtering

Huan Li, Hui Xu

China University of Mining and Technology, Beijing
Email: 2578423277@qq.com

Received: May 6th, 2020; accepted: May 21st, 2020; published: May 28th, 2020

Abstract

With the rapid development of deep learning technology, there are more and more chatbots on the market, which are roughly divided into five categories according to their functions: customer service, small talk, education, personal assistant and question-and-answer chatbots. Their main functions are different. The question-and-answer type mainly meets the requirements of the user's information inquiry, while the small talk type is mainly responsible for the conversation with the user, bringing emotional comfort and spiritual company to the user. According to the different ways of reply, the chatbots are divided into two types: retrieval-based and generative-based. There is a common problem in both of them, that is, when chatting with users, it is easy to get bogged down, interrupt the chat and affect the user experience. The retrieval based chatbot constructs the topic and the corresponding answer in the knowledge base in advance. When the user raises the topic, it looks for the topic most similar to the user's topic in the knowledge base and gives the reply. There are two reasons why this type of robot chat is prone to deadlock. 1. The items related to the user's topic cannot be matched in the knowledge base, so the sentences with the highest probability in the library will be selected for reply, which is usually "mm-hmm", "I know" and other universal replies. 2. The knowledge base can retrieve the answers of users' topics, but the quality of the replies is poor or the users are not interested. Both of these situations can easily lead to a conversation impasse. When the neural network is trained to generate word-by-word responses based on generative conversational data, it is often found that the response statement is not smooth, that is, the response quality is poor, which leads to the deadlock of the chat. Replying for poor quality, the user that is not interested in reply makes chat deadlock problem, this paper puts forward: 1. based on collaborative filtering and keyword extraction way to collect user interest, use the keyword extraction collect user interest, let robot around the interest and the user to chat, to avoid the user stalled by is not interested in topic. The purpose of using collaborative filtering is to expand user interest. Other users' historical chat messages may also contain the interests of the current user, so the collaborative filtering method can be combined to expand the interests of users and optimize the recommendation effect. 2. Introduce external hot topics (Baidu hot topics, hot searches on Weibo) to generate topics that users may be interested in by combining with their interests; when the chat is deadlocked, recommend topics that users may be interested in to break the deadlock; enhance the user experience. Judging whether the chat is deadlocked is also the focus of this article, which involves whether recommendations can be made later. Therefore, this paper focuses on the similarity algorithm of short texts to detect whether the chat is deadlocked, that is, to judge the correlation between the user's speech and the robot's reply to judge whether you are deadlocked by the level of correlation. Fi-

nally, the method proposed in this paper is proved to be feasible by using the evaluation index of the number of rounds of continuous dialogue.

Keywords

Chatbots, Collaborative Filtering, Topic Recommendation

基于协同过滤的聊天机器人话题推荐

李欢, 徐慧

中国矿业大学(北京), 北京
Email: 2578423277@qq.com

收稿日期: 2020年5月6日; 录用日期: 2020年5月21日; 发布日期: 2020年5月28日

摘要

随着深度学习技术的迅速发展, 市场上出现了越来越多的聊天机器人, 按照功能大致分为5类, 客服、闲聊、教育、个人助理以及问答型聊天机器人。它们的主要功能不一样, 问答型主要满足用户信息查询的要求, 闲聊型主要负责和用户对话, 给用户带来情感慰藉和精神陪伴。按照回复方式的不同, 闲聊型聊天机器人又分为基于检索式和基于生成式。二者都普遍存在一个问题, 即和用户聊天时很容易陷入僵局, 中断聊天, 影响用户体验。基于检索式的聊天机器人预先在知识库中构建好话题及对应的答案, 当用户提出话题后, 寻找知识库中和用户话题最相似的并给出回复。这类机器人聊天容易陷入僵局原因有以下两点。1、知识库中匹配不到用户话题相关的条目, 此时会选择库中出现概率最大的句子进行回复, 往往是“嗯嗯”、“我知道了”等万能回复。2、知识库中能检索到用户话题的答案, 但回复的质量差或用户没有兴趣。这两类情况都容易使聊天陷入僵局。基于生成式的使用对话语料训练神经网络逐词的生成回复, 往往出现回复语句不通顺, 即回复质量差的问题, 使得聊天陷入僵局。针对回复质量差、用户对回复不感兴趣使得聊天陷入僵局的问题, 本文提出: 1、结合基于协同过滤和关键词提取的方式收集用户兴趣, 使用关键词提取收集用户兴趣, 让机器人围绕兴趣和用户聊天, 避免用户因对话题不感兴趣而陷入僵局。使用协同过滤的目的是扩展用户兴趣。其他用户的历史聊天信息中也可能包含当前用户的兴趣, 因此可以结合协同过滤方式来扩展用户的兴趣, 优化推荐效果。2、引入外部热点话题(百度热点、微博热搜)结合用户兴趣生成用户可能感兴趣的话题, 当聊天陷入僵局时, 给用户推荐可能感兴趣的话题, 以此打破僵局, 增强用户体验。另外判断聊天是否陷入僵局也是本文的重点, 这涉及到后续能否形成推荐。因此本文重点研究了短文本相似度算法, 以此来检测聊天是否陷入僵局, 即判断用户话语和机器人回复的相关性, 根据相关性高低来判断是否陷入僵局。最后通过对比实验, 使用持续对话轮数评价指标验证了本文提出的方法是可行的。

关键词

闲聊机器人, 协同过滤, 话题推荐

Copyright © 2020 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 研究背景

近年来,随着深度学习技术的迅速发展,人机对话系统的发展势头越来越猛,科研机构和许多大型互联网公司都纷纷加入了聊天机器人的研究,各大公司机构更是推出了自家的产品,2014年微软推出的智能机器人微软小冰、百度的小度、小黄鸡、Siri等。可以预想聊天机器人的发展前景是一片广阔的。

以上提到的机器人可以分为两类,一类是以完成用户任务为主的任务型聊天机器人,比如百度的小度、微软的Siri,另一类是以闲聊为主的闲聊机器人,比如小黄鸡和微软小冰。这是以它们的功能来进行分类的。本文主要介绍的是闲聊机器人。

闲聊机器人解决了人们情感等方面的需求。比如在用户感到孤独且无人陪伴,迫切需要对外交流的时候,如果有个闲聊机器人,滔滔不绝、不知疲倦地陪用户聊天,则能缓解用户的孤独感。当然这只是聊天机器人的一个应用场景,我们还可以设想,儿童在学习语言或者在获取知识的时候往往有源源不断的问题,家长可能很快就厌倦了回答重复的问题,但是机器人不会厌倦,这就为早期教育提供了帮助。而且人类语言学习能力是有限的,但是机器掌握一门语言则快得多,我们可以利用聊天机器人进行外语的学习或者承担翻译任务等等。

同时我们注意到在与人们聊天的过程中,机器人可以获取用户的语言习惯与生活习惯,从而能够分析出人们的行为特点。这也就为精准医疗提供了大量数据,世界上第一个聊天机器人就是为分析人们的心理问题而诞生的,可以设想未来医疗领域聊天机器人会有广阔的应用与美好的前景。

但是人工智能发展到今天,闲聊类机器人有一定的缺陷。它离人脑有一定的差距,因此闲聊类的机器人缺少一定的人性化,很多情况下无法理解人类的话语,因此目前的闲聊类的聊天机器人不能长久地和用户进行聊天,基本上智能维持一轮或几轮对话,其中很大一部分原因就是聊天机器人有时无法理解用户的话语,这时候他会采取一种安全回复的做法。安全回复是所有机器人都会碰到的问题,也就是当用户说了一句话后,机器人无法理解这句话的含义或者说不知道怎么回复用户的问题,然后采取一种是而非的回答,比如“嗯嗯,我知道了”、“对不起”、“没事”或者随机生成其他话语等避开用户提问的回答,显而易见这是一种很影响用户体验的行为,因此如何让用户和聊天机器人能长时间地聊下去是目前面临的一个大问题。如果机器人了解了你的兴趣爱好,围绕你的兴趣爱好和你聊天,那么将可能改善对话持续轮数很少的情况。本文提出的基于协同过滤的话题推荐,其主要的目的就是根据用户以往的聊天记录发现用户的兴趣并根据用户的兴趣推荐其可能感兴趣的话题。当机器人无法回答用户提出的问题或回答不满意时,机器人可以根据用户的兴趣推荐可能感兴趣的话题,增强用户的聊天体验

2. 聊天机器人相关技术

聊天机器人中涉及到NLP(自然语言处理)中的许多技术,包括基础的文本向量化、中文分词、句法分析、关键词提取、文本相似度检测等,本节介绍主要的几种技术。

2.1. 中文分词

与英文不同,英文每一个单词都能表达完整的含义,因此英文文本并不需要分词处理,而中文文本不同,待处理的文本是由成句成段,因此需要先将其表示成一个个分开的中文词汇在中文自然语言处理的各项任务中,中文分词是最基础的一项任务。中文分词方法大概分为三种主流的流派,每一种中文分词方法下又包含各种不同的具体的分词方法。

2.1.1. 基于规则的分词

基于规则的分词方法是一种逻辑简单、机械方式的方法,需要事先建立好分词词典和规则库,核心思想是字符串匹配。它依赖于具体的词典,不断的从词语库中寻找词汇和待切分句子进行匹配,如果能

在词典中找到该词汇则进行切分, 否则不进行切分, 继续让下一个待切分的词汇与词典中的词汇进行匹配, 直到该句子被切分完毕。基于规则的分词方法又分为主要的三类, 分别是正向最大匹配、逆向最大匹配以及双向最大匹配。

2.1.2. 基于统计的分词

上文讲到基于词典的分词方法, 它的弊端是需要不断的维护庞大的词典, 随着词典的增大这必然会给分词带来更多的时间。基于统计的思想是“词是由字组成的”, 它基于这样一个假设, 如果连续的字多次出现在不同的文本那么这个连续的字越有可能是一个词。近年来基于统计的分词方法得到了快速的发展。随着近年来机器学习的迅速发展以及文本数据的爆发增长, 统计分词明显占据明显的优势。基于统计的分词是对中文文本中的最基本的单位“字”进行处理。连续出现的字在不同文本中出现的频率来衡量其是否属于一个词, 可以预先为其设定一个阈值, 当频率超过这个阈值那么则认定这些连续出现的字组成的是一个词语。否则就不是。

统计分词方法一般分为 2 步:

1) 建立统计语言模型;

2) 对待分词的句子中的每个字进行划分, 每个字都有一个标签代表其属于词首、词中、词尾, 然后用动态规划等方法计算出一条最大概率路径, 这条路径代表了这句话中每个字组成的一条最大可能的概率路径。也就是每个词最有可能的标记。从所有的路径中挑选出概率最大的那一条作为每个字的标记。目前主要的统计学习方法有 HMM (隐马尔可夫) 和 CRF (条件随机场)。

2.1.3. 基于混合的分词

尽管基于统计的分词方法在理论上优于基于规则的分词方法, 但是在具体的分词任务中, 分词效果差距并不明显, 事实上某些特定场景基于规则的分词方法分词效果稍微优于基于统计的分词方法, 因此都是针对不同的场景具体任务具体分析, 没有哪一种方法是占据绝对优势的, 大部分情况下, 都是几种分词方法混合使用, 具体的做法是先基于一种分词方法, 分词完毕后用其他的分词方法加一辅助来提升前一种分词方法的分词效果。目前常用的方法是先用基于规则的分词方法初步分词, 然后使用 CRF、HMM 隐马尔可夫模型加以辅助, 这样能很好的处理歧义词和未登录词。目前比较好的分词工具 jieba 分词正是基于这种思想。

2.2. 文本向量化

在自然语言处理的各项任务中, 词向量化是不可避免的, 用非常简单的话来说明词向量化的作用就是提供了一种方法把人类能够理解自然语言表示成机器能够理解的数字表示[1]。自然语言处理中比较常见的词向量模型是分布式模型(distributed representation)和独热编码(one-hot Representation), 本文主要讲解 word2vec, 它是一种分布式的词向量表示模型, 把词语映射从高维的空间映射到低维的语义空间, 将词语表示成稠密的连续向量。它分为 CBOW 和 Skip-gram。

2.2.1. CBOW

CBOW 的思想是根据上下文去预测中心词, 网络结构只有三层, 输入, 映射和输出层, 它是一种根据给定窗口中上下文词语来预测中心词概率的语言模型[2], 输入是窗口中的上下文词汇, 输出是中间词, 输入的词汇用独热编码进行表示, 然后通过和一个映射矩阵相乘, 然后把所有向量相加或者相加后取均值, 然后又经过一个矩阵相乘, 最后用一个 softmax 激活函数得到每个词语的概率, 神经网络训练目的是调整网络中的参数使得中心词 y 的生成概率最大化。当网络训练完毕, 输入层到映射层的映射矩阵即为最后所求的词向量。如图 1 所示。

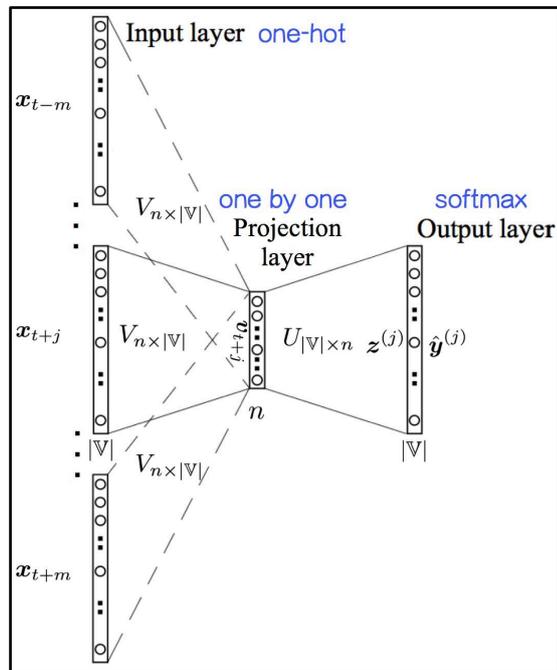


Figure 1. CBOW model
图 1. CBOW 模型

2.2.2. Skip-Gram

Skip-gram 的思想和 CBOW 正好相反，前者利用上下文去预测中心词[3]，后者利用中心词去预测上下文，网络结构和 CBOW 一样，只是前者输入是窗口中的上下文，后者的输入是一个词，CBOW 输出是一个 $1 \times V$ 的概率分布，skip-gram 输出是多个 $1 \times V$ 的概率分布。如图 2 所示。

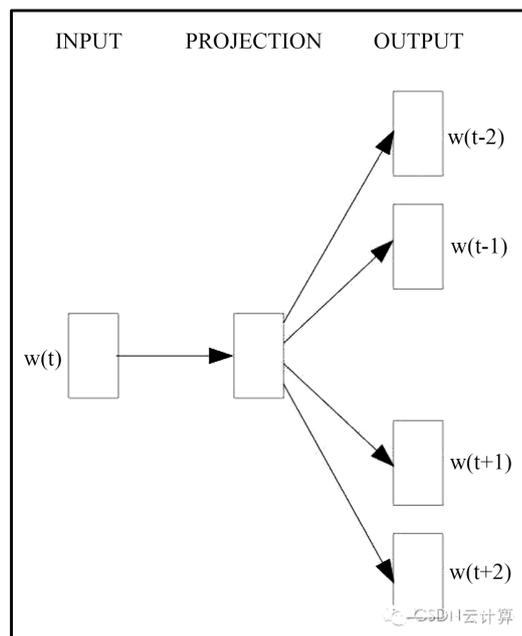


Figure 2. Skip-gram
图 2. Skip-gram

现在给定一句话“the cat sat on the mat”，假设窗口 c 为 2，中心词是 sat，位置是 3，按照 skip-gram 的思想要去预测它上下文，因为窗口是 2，即预测前两个和后两个词汇，一般不预测常用词和停用词，预测位置 2 单词是“cat”，在位置 6 上预测单词是“mat”。

Word2vec 中用到了哈夫曼树和层次化 softmax (复杂度 $O(N)$ 到 $O(\log N)$)、层次化采样(复杂度从 $|V|$ 降低到了树的高度)，大大节省了计算成本和存储空间。

2.3. 关键词提取之 TextRank

顾名思义，关键词是一篇文本中最能表达文本语义，最能代表文本含义和文本高度相关的一些词汇 [4]。由于目前的文本大部分内容冗长，短时间难以从中获取有用信息，因此关键词抽取此时就能发挥其作用。关键词抽取是指把一些能代表文本语义、和文本高度相关的词语从自然语言文本中抽取出来。关键词提取在自然语言处理的许多任务中发挥着举足轻重的作用，尤其是在文本检索、文档比较、摘要生成、文档分类和聚类等方面发挥着重要的作用。

TextRank

直接翻译过来的意思是文本排名，具体来说是文本中每个词汇的重要度排名。其思想来源是上面的 PageRank 算法，只不过做了细微的改动。PageRank 是根据网页之间的链接关系构建图模型，而 TextRank 算法是根据文本中词语的相邻关系来构建图模型，主要思想还是如下。

1 如果文本中的某个词汇出现在很多其他的词汇后面，那么这个词汇越重要，相应的有很高的 TextRank 值。

2 如果某个单词出现在一个权重很高的词汇后面，那么这个单词也很重要，相应的权重就会越高。

每个词汇的 PageRank 值计算公式如下：

$$S(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j) \quad \text{公式 1}$$

$S(v_i)$: 词语 i 的 TextRank 值，一般每个词语的 TextRank 值初始化为 1。

d : 阻尼系数，一般取 0.85(这个 d 取值 0.85 是通过无数的实践得出的结果，实际上的作用相当于做一个平滑处理)。

$Out(v_j)$: 出现在词汇 j 后面的词汇集合。

W_{ji} : 连接顶点 i 和 j 的边的权重。

有了上面的计算公式后，通过不断的迭代计算每个顶点的权值直到收敛，从上面的公式可以看出某个词汇 PageRank 值和出现在它前面的词汇的权重以及连接的边的权重有关。

3. 实验过程和结果展示

3.1. 实验环境

本实验使用的是谷歌开源深度学习框架 Tensorflow。TensorFlow 是一个基于数据流编程(dataflow programming)的符号数学系统，被广泛应用于各类机器学习(machine learning)算法的编程实现，其前身是谷歌的神经网络算法库 DistBelief。

Tensorflow 拥有多层次结构，可部署于各类服务器、PC 终端和网页并支持 GPU 和 TPU 高性能数值计算，被广泛应用于谷歌内部的产品开发和各领域的科学研究。

种平台多种设备，甚至是移动设备。Tensorflow 是基于数据流图的处理框架，图中的节点表示数值计算，边代表运算节点之间的数据交互数值计算，边代表运算节点之间的数据交互。

3.2. 实验数据

3.2.1. 对话语料

本文用到的是聊天机器人公开语料, 包含各种主题(文学、体育、科技、娱乐等), 以下是部分对话语料截图, 如图 3 所示。

```
告诉我关于美国内战
你觉得南方是正确的?
你知道美国内战
我在各州之间的战争非常感兴趣。
什么是历史
历史: 政治, 经济, 军事事件随着时间的推移, 从人的黎明AI时代的进程。
什么样的历史
我喜欢谈论机器人和电脑的历史。
你对历史感兴趣
我对历史很感兴趣了。你喜欢什么时期?
解释史
历史上有两大解释, 这取决于您是否接受个人作为重要与否的作用。
```

Figure 3. Dialogue corpus

图 3. 对话语料

有了聊天语料就可以使用第二章介绍的关键词提取算法从对话中提取关键词, 然后经过简单分析得到用户的初步兴趣, 然后使用协同过滤方法进一步扩展用户的兴趣, 以便后续话题推荐的效果更好。

3.2.2. 热点新闻语料

当机器人的回复是无意义的话语或者是安全回复时, 本文提出了基于用户兴趣的热点新闻话题推荐来代替先前的回复, 用到的语料从网络爬取, 是当前热点话题, 包括微博热搜、百度热搜等新闻或者是视频等, 涵盖体育、娱乐、科技等主题, 如图 4 所示。

```
郎平向武汉捐物资
美国累计病例破百
伊朗翻译诊疗方案
多所高校网络答辩
大家都宅怕了吧? 推荐几个旅游景点
天呐! 一只会给主人暖手手的猫! 好可爱啊
女孩被批评后“语重心长”教育家长: 要用古诗来骂人, 全程高能
这个视频看得真舒服, 拍得很用心——植物的栽培与嫁接技术, 通俗易懂, 简单好操作。
西藏纳木措, 太美了!
c罗重回伯纳乌, 在自己的观席上观看德比。虎扑体育
为普通的风景照增加一些趣味吧
```

Figure 4. Corpus of hot topics

图 4. 热点话题语料

3.3. 整体实验流程

3.3.1. 关键词提取并生成兴趣

本文用到的关键词提取算法是 TextRank 算法, 该算法已经在第二章介绍过, 它不需要语料库作为依赖, 仅需要提供当前文档就能建立图模型生成给定个数的关键词, 本文提出的用户兴趣发现是基于协同过滤的方式, 具体的做法是抽取十段对话语料, 当作是十个用户的对话语料, 然后分别提取关键词, 然后使用协同过滤方式生成用户兴趣。

3.3.2. bert 中文问答对检测

bert 中文问答对检测的作用是针对给定一个问答对, 检测问题和回答的相关程度。可以用 bert 中文

问答对检测用户和机器人聊天的过程中的问答的相关程度, 或者说用户对机器人回答的满意程度, 以此来决定是否用用户感兴趣的热点话题来代替当前不满意的回答。以下是统计 100 个问答对的情况。其中程序判断相关并且人为判断相关的有 68 个, 程序判断不相关并且人为也判断不相关的 13 个, 正确率为 81%, 如表 1 所示。

Table 1. Sample comparison

表 1. 算法示例

	人为判断	程序判断	重叠
相关	78	83	68
不相关	22	17	13
总计	100	100	81

3.3.3. 热点话题推荐

当使用 bert 中文问答对检测程序对机器人的回复进行检测后, 如果程序检测出机器人回答和用户话语不相关或相关性很弱, 那么就会根据用户的兴趣和当前话题给用户推荐可能感兴趣的话题、视频来代替不相关的回答, 即达到了话题或者新闻推荐的目的, 又可以增强用户的聊天体验。以下是原始的聊天机器人程序和加入了本文提出的基于协同过滤的话题推荐方法后的聊天机器人程序对比。

部分对比示例如表 2 所示。

Table 2. Chat comparison

表 2. 聊天对比

原始聊天机器人对话	Q: 你觉得 c 罗和梅西谁更厉害 A: c 罗: 葡萄牙足球运动员, 梅西: 阿根廷足球运动员	Q: 因为这次疫情我在家待的快发霉了 A: 嗯嗯	Q: 你觉得马云是外星人吗 A: 你说的对	Q: 好无聊啊, 不能让我不那么无聊啊 A: 那我陪你聊天吧!	Q: 周星驰是最好的喜剧演员吗 A: 嗯嗯, 我不太了解, 你能给我介绍吗
基于协同过滤的话题推荐聊天机器人	Q: 你觉得 c 罗和梅西谁更厉害 A: c 罗愿意降薪三分之一 Q: c 罗目前的薪水多少 A: 不知道呢!	Q: 因为这次疫情我在家待的快发霉了 A: 美国累计病例破百	Q: 你觉得马云是外星人吗 A: 马云出席乡村教师颁奖大会 Q: 我觉得马云是个了不起的商人 A: 嗯嗯!	Q: 好无聊啊, 不能让我不那么无聊啊 A:	Q: 周星驰是最好的喜剧演员吗 A: 周星驰生日低调庆生, 少量艺人送祝福! Q: 感觉星爷好孤单啊 A: 嗯嗯, 是的

以上是从 100 个聊天对比实验中截取的五个对话片段, 从中可以看出有的情况下可以推荐出用户可能感兴趣的话题, 并让聊天持续下去, 由于语料库中的热点新闻是最近的, 导致热点新闻涵盖不全, 并且受限于语料库质量和技术本身, 有时推荐的话题用户并没有兴趣, 因此聊天中断或者用户换了话题。以下是原始聊天机器人对话轮数是一轮, 加入协同过滤话题推荐后用户对话轮数变化情况, 如表 3 所示。

Table 3. Comparison of the number of dialogue rounds

表 3. 对话轮数对比

一轮保持不变	一轮变两轮	一轮变三轮
67	30	3

因此可以表明本文提出的基于协同过滤话题推荐方式可以在机器人生成不满意回复后给用户推荐可能感兴趣的话题并且提高对话轮数。

4. 结语

本文提出了一种在聊天机器人中的基于协同过滤的话题推荐方法, 旨在给用户推荐可能感兴趣的热点话题, 同时在机器人生成安全回复时使用推荐可能感兴趣话题的方法使得话题持续下去, 增强用户的体验, 经过实验验证本文提出的方法是可行的。

参考文献

- [1] 曹东岩. 基于强化学习的开放域聊天机器人对话生成算法[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2017.
- [2] 张世尧. 基于用户聚类的微博话题推荐方法研究[D]: [硕士学位论文]. 安徽: 安徽理工大学, 2017.
- [3] 王广新. 基于微博的用户兴趣分析与个性化信息推荐[D]: [硕士学位论文]. 上海: 上海交通大学, 2013.
- [4] 杨晶. 用户兴趣模型及实时个性化推荐算法研究[D]: [硕士学位论文]. 南京: 南京邮电大学, 2013.