

基于语义和用户偏好的网络电视直播推荐方法

张 翔¹,袁 政²,蒋 旦¹,朱 明¹

(1 中国科学技术大学 信息科学技术学院,安徽 合肥 230022,2 上海文广互动电视有限公司,上海 200072)

摘 要: 提出一种基于语义和用户偏好的网络电视直播实时推荐方法,该方法首先基于用户的历史记录构建用户偏好模型,然后使用基于词向量的语义相似度计算方法,分别计算待推荐节目和用户记录或待推荐节目和用户当前观看节目间的相似度,再结合该相似度和用户偏好求取用户对待推荐节目的虚拟兴趣,最后选出虚拟兴趣较高的一组节目作为对用户的实时推荐.实验结果表明,此方法的命中率在实时预测推荐的场景下较对比方法提高了10%以上,且在实时节目推荐的场景下有更好的推荐效果.

关键词: 网络电视直播;实时推荐;用户偏好;语义相似度

中图分类号: TP37

文献标识码: A

文章编号: 1000—7180(2016)12—0052—05

A Recommended Method Based on Semantics and User Preferences in Network Live Television

ZHANG Xiang¹,YUAN Zheng²,JIANG Dan¹,ZHU Ming¹

(1 School of Information Science and Technology, University of Science and Technology of China, Hefei 230022, China; 2 Shanghai Interactive Television Co., Ltd., Shanghai 200072, China)

Abstract: This paper proposes a real-time recommendation method based on semantics and user preference in Live TV. Firstly, it constructs a user preference model based on users' watching history, and calculates the similarity between the to-be-recommended programs with users' watching history or users' watching programs respectively with the Semantic similarity calculation method based on word vector. Combined with user preference, the similarity is used to work out users' virtual interest about the to-be-recommended programs. Finally, it's the program which gets higher virtual interest that becomes users' real-time recommendation result. Experimental results show that the recall rate of this method is improved by about ten percent than the existing methods when it's used on the real-time predicted recommendation. And the quality of recommendation is improved significantly on real-time program recommendation.

Key words: network live television; real-time recommendation; user preference; semantic similarity

1 引言

当前的网络电视直播环境中,直播频道的项目信息较少,例如一般的直播节目只有节目名称、节目类别以及简短的节目简介,这就为基于节目字面层、句法层的内容相似性进行推荐带来了困难.同样的,由于电视直播节目具有时效性,以前播放过的节目可能不会再次播放,整个节目库也是不断更新的,所

以基于协同过滤的推荐也很难取得较好的推荐效果.

针对以上问题,本文考虑通过语义层对节目描述信息进行分析比较,同时根据用户历史记录挖掘用户兴趣规律、构建用户模型,最后基于用户偏好和节目语义相似度对用户进行实时推荐.简而言之,即是提出了一种基于语义和用户偏好的网络电视直播实时推荐方法(SAUPRM).

2 数据准备与预处理

本文算法中所用到的数据主要包括用户观看网络电视直播产生的观看记录和央视、卫视、地方等电视直播频道当天的电子节目单。

用户的观看记录可以通过用户日志获取,为了消除用户“兴趣漂移”带来的影响,通常选取用户最近一段时间(如 30 天)内的 LOG,然后对其进行预处理,获取相应字段,可以得到如表 1 所示的用户记录基本数据格式。

表 1 用户记录数据格式

name	content	category	duration	atime	stime	etime
江苏卫视	继承者们	电视剧	44	180	93	101
CCTV-1	朝闻天下	新闻	100	150	72	92
CCTV-3	非常 6+1	综艺	80	80	234	250
北京卫视	半月传	电视剧	42	100	124	132
湖南卫视	旋风孝子	综艺	85	100	264	281

其中 name 是频道名称,content 是节目名称,category 是节目类别,duration 是观看时长(分钟),atime 是节目总时长(分钟),stime 是用户观看的起始时间分段,etime 是用户观看的结束时间分段。

对于时间分段,解释如下:由于用户在观看电视节目时一般会有一个试看选台的过程,根据经验,一般认为 5 分钟以内的观看属于用户选台阶段,此时的观看记录属于用户记录中的噪声,在预处理时常常予以祛除,而超过 5 分钟的观看记录则属于有效记录。因此,本文考虑将一天的时间以 5 分钟每段进行划分,定义时间分段为按该划分后的每段时间,则此时一天共有 288 个时间分段。如表 1 中,节目“朝闻天下”用户观看的起始时间是 06:00,则其起始时间分段对应为 $06 \times 60/5=72$,结束时间分段对应为 $72+100/5=92$ 。所以,用时间分段可以描述用户一天内各个时间段的有效行为,以此来表示用户一天中各时间段内的行为兴趣。

下一步根据用户观看节目时长计算出用户对每个观看过节目的感兴趣程度,即兴趣度 i ,定义为用户观看该节目的时长与该节目总时长的比值,计算公式为

$$i = \frac{\text{duration}}{\text{atime}}$$

(1)

式中, $i \in (0,1]$,对于每条用户记录的 stime 和 etime,我们认为用户在 $(\text{stime}, \text{etime}]$ 时间分段内

其兴趣度都是 i 。

据此可生成用户在一段时间内关于(观看时间的分段 t 、节目类别 c 、兴趣度 i) 的兴趣矩阵,表示如下:

$$\mathbf{I} = [t_m, c_m, i_m]$$

(2)

式中, t_m, c_m, i_m 均为 m 维列向量, m 表示用户所有观看记录各时间分段数目之和,设 L 为用户记录总数,则

$$m = \sum_1^L (\text{etime-stime})$$

(3)

对于每条用户记录,分段 t 的取值是多个,即 $t \in (\text{stime}, \text{etime}]$, $t \in \mathbb{Z}$,而对应的 c 和 i 不变。

节目单库中央视、卫视等电视直播频道当天的电子节目单可以通过互联网获取,进行预处理后,可获得类似表 1 的数据格式,主要属性包括播出时间 stime、节目结束时间 etime、频道名 name、节目名 content、节目类别 category,然后将其预处理为关于(播出时间 t_s 、结束时间 t_e 、节目名 d 、节目类别 c') 的节目库矩阵 $\mathbf{E}$,表示如下:

$$\mathbf{E} = [t_{s_n}, t_{e_n}, d_n, c'_n]$$

(4)

式中, $t_{s_n}, t_{e_n}, d_n, c'_n$ 均为 n 维列向量, n 表示节目库中节目总数。

3 用户偏好模型构建

在获取了用户一段时间内的兴趣矩阵后,这一步开始为用户构建偏好模型。由于节目库中节目相关信息较少,且节目类别能较为直观地反映节目的属性,也能较为形式化地表达出用户的兴趣,而一天不同的时间段内用户对不同类别节目的喜好也是不一样的,因此我们使用第一步数据预处理中生成的时间分段和节目类别作为描述用户偏好的属性值,基于用户一段时间内的兴趣矩阵 $\mathbf{I}$ 为用户生成相应的偏好模型。

定义矩阵 $\mathbf{U}$ 是用户关于时间分段 t 和节目类别 c 的偏好模型,则 $\mathbf{U}(c, t)$ 可表示如下:

$$\mathbf{U}(c, t) = \begin{bmatrix} u_{t_1 c_1} & \cdots & u_{t_1 c_q} & \cdots & u_{t_1 c_k} \\ \vdots & & & & \\ u_{t_p c_1} & & u_{t_p c_q} & & \cdots \\ \vdots & & & \cdots & \\ u_{t_s c_1} & \cdots & & & u_{t_s c_k} \end{bmatrix}$$

(5)

式中, $u_{t_p c_q}$ 表示用户在时间分段 t_p 时对节目类别 c_q 的兴趣度。其中 $1 \leq p \leq s (p \in \mathbb{Z})$, s 表示时间分段所有可能取值的总数; $1 \leq q \leq k (q \in \mathbb{Z})$, k 表示节目类别总数;则 $\mathbf{U}(c, t)$ 是一个 $s \times k$ 矩阵。

对于 u_{p,c_q} , 其计算依赖于式(2)中的用户兴趣矩阵 $\mathbf{I}$, 计算公式如下:

$$u_{p,c_q} = \frac{\sum_{0 \leq j \leq m} (\mathbf{I}(j,3) | \mathbf{I}(j,1) = t_p, \mathbf{I}(j,2) = c_q)}{r} \quad (6)$$

式中, r 是所有满足 $\mathbf{I}(j,1) = t_p, \mathbf{I}(j,2) = c_q$ 的 $\mathbf{I}(j,3)$ 总数; r 为 0 时, u_{p,c_q} 也为 0.

这样, 就构建出了用户偏好模型 $\mathbf{U}$, 该模型可描述出用户在过去某一段时间内对所有类型节目在每天不同时间分段的喜好程度.

4 基于语义的相似度计算方法

传统的基于内容的推荐一般是基于项目内容的字面层、句法层进行项目间相似性的计算并推荐, 然而这种计算并没有考虑到语义级别的项目含义.

由于近年来, 在自然语言处理领域, 词向量空间模型在语义挖掘和表示上取得了新的进展, 相比于传统的以《知网》《同义词林》为代表的基于词典的语义计算方法^[2,3], 这些方法基于大型语料库进行训练, 编码了词汇之间的统计关系, 涵盖了更广的词汇量, 因此在语义相似度计算上表现突出. 其中, 基于神经网络的语言模型 (Neural Network Language Model, NNLM) 表现尤为突出. NNLM 具有较多优点, 例如其能克服维度灾难、可以建立较远的上下文关系、能建立词语在语义维度上的联系等.

NNLM 最早是由 Bengio 提出^[4], 该模型是由一个线性映射层和一个非线性隐层构成的前向神经网络, Bengio 提到通过训练该语言模型还可以同时获得其副产品—词向量. 之后 Mikolov 研究并简化了 Bengio 提出的 NNLM, 又提出了 CBOW 模型和 Skip-gram 模型^[5,6], 这两个模型的目的并不是为了训练出一个语言模型, 而是为了得到其训练出的副产品—单词的语义向量. 经过 Mikolov 简化后的模型提高了训练速度, 而得到的词向量则可以被利用在各种不同的自然语言处理任务中. 而 Mikolov 经过研究也发现, CBOW 模型对于字符的推理效果更好, Skip-gram 模型则对于语义的推理效果更佳. 因此本节采用 Skip-gram 模型对节目进行语义层的相似性计算.

由于 Skip-gram 模型训练得到的词向量存在一定的线性推导关系^[5,6], 所以可以按如下方式计算两个节目基于语义的相似度. 首先对两个节目名称进行分词, 分词后采用由训练得到的单词向量的线

性加和来构成两个节目文本的向量表示, 在得到节目文本的向量表示后, 即可采用余弦相似度来计算两个节目文本基于词向量的语义相似度. 设节目间基于语义的相似度为 s , 则 $0 \leq s \leq 1$. 举例如下: 节目 d_1 : 动物世界, 分词后得到 $d_1 = (\text{动物}, \text{世界})$; 节目 d_2 : 自然传奇, 分词后得到 $d_2 = (\text{自然}, \text{传奇})$, 基于上述方法计算节目语义相似度, 可得 $s = 0.655\ 99$.

5 基于语义和用户偏好的实时推荐

根据实际应用场景的不同, 提出两种实时推荐方法. 第一种是实时预测推荐, 当用户还没有观看节目时, 针对用户过往的观看历史和当前时间, 为用户预测并推荐一组其当前可能感兴趣的实时节目. 第二种是实时节目推荐, 针对用户当前正在观看的节目, 为其推荐一组符合兴趣且和当前节目相似的实时节目.

5.1 实时预测推荐

实时预测是在用户还没有观看节目时, 根据用户过去的观看记录和当前时间 t_0 , 预测出用户此时可能感兴趣的节目, 推荐给用户. 具体步骤如下.

(1) 首先根据用户最近的观看记录和当前时间 t_0 过滤出用户记录中观看时段和 t_0 相似的部分. 由于用户观看节目的时间并不是固定不变的, 每天在相同时段都会有变化, 所以过滤规则如下: 只要 $(t_0 - 10)$ 或 $(t_0 + 10)$ 处于用户某观看记录的时段内, 即认为该记录符合过滤要求. 经过上述处理后得到关于 (观看时间 t 、节目 d) 的用户记录矩阵 $\mathbf{E}_u$, 如下:

$$\mathbf{E}_u = [\mathbf{t}_w, \mathbf{d}_w] \quad (7)$$

式中, $\mathbf{t}_w, \mathbf{d}_w$ 均为 w 维列向量, w 表示经过过滤后的用户记录数.

(2) 根据式(4)中节目单库矩阵 $\mathbf{E}$ 和当前播放时间 t_0 获取实时节目库, 用矩阵 $\mathbf{E}_c$ 表示:

$$\mathbf{E}_c = [\mathbf{t}_{s_n}, \mathbf{t}_{e_n}, \mathbf{d}_n, \mathbf{c}_n] \quad (8)$$

式中, $\mathbf{d}_n$ 表示各频道实时播放节目, $\mathbf{c}_n$ 表示各节目对应节目类别, n 为各频道实时节目总数.

(3) 针对式(8)中的实时节目库矩阵, 计算每个实时节目 d_n 与(7)中经过过滤的用户记录基于语义的相似度. 计算方法如下, 首先利用 4 节中基于词向量的语义相似度计算方法, 计算出 d_n 和每个 d_w 的相似度, 同时为了消除记录中相似度较低的噪声记录的影响, 选取其中相似度最高的三组, 求其均值, 不满三组则求其全部相似度均值. 将计算出的相似度均值作为 d_n 与过滤后用户记录基于语义的相似度, 记为 s_n . 则此时实时节目库更新为

$$E_s = [t_{s_n}, t_{e_n}, d_n, c_n, s_n] \quad (9)$$

(4) 利用3节中构建的用户偏好模型,计算用户对实时节目库中每个节目相应类别的偏好度,记为 P 。同时,设 t_0 对应时间分段为 a ,设 t_0 到每个实时节目的结束时间 t_e 包含 b 个时间分段,则 P 的计算公式如下:

$$P_{c_q} = \frac{\sum_{i=0}^b u_{t_{a+i}+c_q}}{b} \quad (1 \leq q \leq k) \quad (10)$$

式中, k 表示节目类别总数, $b > 0$ 。

(5) 定义虚拟兴趣 v_n ,表示用户对实时节目库中节目的兴趣度,计算如下:

$$v_n = (s_n \cdot P_{c_q}) | c_n = c_q, P_{c_q} \neq 0 \quad (11)$$

当 $P_{c_q} = 0$ 时, $v_n = s_n \cdot \min\{P_{c_q} | P_{c_q} \neq 0\}$,此举是为了保证用户在当前时间段从未有过偏好的高相似度节目也能得到推荐。

(6) 根据步骤(5)计算出的虚拟兴趣 v_n 更新当前实时节目库,用矩阵 E_v 表示如下:

$$E_v = [d_n, c_n, s_n, v_n] \quad (12)$$

然后,对 E_v 按 v_n 进行降序排列,选取 v_n 较高的 K 个节目,即是对用户的实时预测推荐,一般地, K 常选取3,5,8,10等。

5.2 实时节目推荐

实时节目推荐方法针对用户当前正在观看的节目 d_0 ,为其推荐一组符合兴趣且和当前节目相似的实时节目。

该方法首先根据式(4)的节目单库矩阵 E 获取如式(8)所示的实时节目单库 E_c ,然后依据第4节中基于词向量的语义相似度计算方法,计算 E_c 中每个实时节目 d_n 和用户正在观看的节目 d_0 基于语义的相似度 s_n ,再根据式(10)计算实时节目相应类别 c_n 对应的用户偏好 P ,最后利用式(11)即可计算出用户在该推荐场景下对实时节目库中节目的虚拟兴趣 v_n 。同样的,对每个实时节目按 v_n 进行降序排列,选取 v_n 较高的 K 个节目,即是对用户的实时节目推荐,一般地, K 常选取3,5,6等。

6 实验与结果分析

本文的实验数据包括用户的观看记录和网络电视直播频道的电子节目单库。其中用户记录来源于10位志愿者用户使用播控平台30天产生的用户LOG,电子节目单库则来源于互联网上公开数据的抓取与聚合,如搜视网、电视猫等。

推荐算法常用的评价方式有在线评价和离线评

价两种^[7],本实验采用离线评价对5.1节提出的推荐方法进行评估,验证其有效性。

实验方法如下:首先获取10位志愿者用户使用播控平台30天产生的用户记录,以及这30天相应的电子节目单。选取前24天的用户记录作为训练集,剩下6天的用户记录作为测试集。利用训练集构建用户最近时间段内的用户偏好模型,测试时选取测试集中每条用户记录对应的观看时间作为实时预测推荐输入的当前时间 t_0 ,采用5.1节所述的推荐方法,为用户推荐一组 K 个节目,如果该组推荐节目中含有该条测试集记录,则本次推荐视为命中。对比方法采用基于内容的推荐和协同过滤推荐相混合的推荐方法(记为CBACFRM),该方法首先基于节目内容初步筛选实时节目组,再利用协同过滤计算相似用户,依据相似用户的观看记录进行预测推荐。推荐方法的有效性可使用命中率(也叫召回率)来衡量:

$$\text{命中率} = \frac{\text{推荐的命中次数}}{\text{测试集中记录总数}}$$

实验在每组推荐节目数 K 取不同值时(3,6,9,12,15),分别进行实验,实验结果如图1所示。

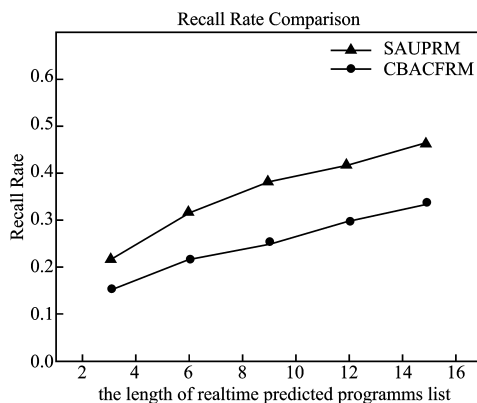


图1 命中率对比图

如图1所示,本文推荐方法的命中率在实时预测推荐节目列表长度不同时均较对比方法更高,命中率平均提升10%左右。同时,随着推荐列表长度的增长,命中率也在提升,然而列表较长的推荐可能会带给用户较差的体验,因此在实际应用中应该综合推荐命中率的高低和推荐列表的长短两个因素,给用户更好的推荐体验。

为了验证5.2节提出的推荐方法的有效性,在播控平台上构建一个实时节目推荐的模拟系统,在收集了10位用户30天的LOG后,选择其后连续的7天作为实验时间,这7天每天在用户的使用时间

中随机选择 5 次为其进行实时节目推荐,每次推荐 2 组,一组 6 个实时节目.实验结果表明,本文推荐方法的推荐效果较对比方法更好.

7 结束语

本文针对当前网络电视直播中节目信息较少、节目单库更新频繁,且现有个性化推荐算法对此效果不佳的问题,提出了一种基于语义和用户偏好的网络电视直播实时推荐方法.该推荐方法在提升用户体验、优化直播资源配置上具有很好的现实意义.实验结果也表明,本文提出的方法在网络电视直播实时推荐的应用场景下较现有方法推荐效果更好,且可在一定程度上解决用户“兴趣漂移”的问题.然而,本文提出的方法同样存在一些不足,如在系统初期、用户记录稀疏时推荐效果不尽理想.因此,如何利用搜索引擎深入挖掘项目信息、如何采用更精确的属性构建用户模型以解决上述问题,将会是今后一段时间的研究重点.

参考文献:

[1] 许海玲,吴 潇,李晓东,等. 互联网推荐系统比较研究[J]. 软件学报, 2009, 20(2): 350-362.
[2] 刘群,李素建. 基于《知网》的词汇语义相似度计算[J]. 中文计算语言学, 2002, 7(2): 59-76.

[3] 田久乐,赵蔚. 基于同义词词林的词语相似度计算方法[J]. 吉林大学学报(信息科学版), 2010 ,28(6): 602-608.
[4] Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. The Journal of Machine Learning Research, 2003, 3(6): 1137-1155.
[5] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. Computer Science, 2013.
[6] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems. 2013; 26(1):3111-3119.
[7] 朱郁筱,吕琳媛. 推荐系统评价指标综述[J]. 电子科技大学学报, 2012, 41(2): 163-175.

作者简介:

张 翔 男,(1993-),硕士研究生.研究方向为推荐算法、数据挖掘. E-mail:zx1064@mail.ust.edu.cn.
袁 政 男,(1964-),高级工程师.研究方向为广播电视、新媒体技术.
蒋 旦 男,(1992-),硕士研究生.研究方向为自然语言处理.
朱 明 男,(1963-),教授,博士生导师.研究方向为视频大数据分析、智能机器人、未来网络.

(上接第 51 页)

[3] Park I Y, Choi Y K, Ko K Y, et al. BD180 - a new 0.18 μm BCD (Bipolar-CMOS-DMOS) Technology from 7V to 60V[C]// Power Semiconductor Devices and ICs, 2008. ISPSD '08. 20th International Symposium on. [s. l.]: IEEE, 2008;64-67.
[4] De la Rosa, José M, et al. A CMOS 110-dB@40-kS/s programmable-gain chopper-stabilized third-order 2-1 cascade sigma-delta Modulator for low-power high-linearity automotive sensor ASICs[J]. Solid-State Circuits, IEEE Journal of, 2005,40(11):2246-2264.
[5] Xu L, Gonen B, Fan Q, et al. 5.2 A 110dB SNR ADC with $\pm 30\text{V}$ input common-mode range and $8\mu\text{V}$ Offset for current sensing applications[C]// Solid-State Circuits Conference - (ISSCC), 2015 IEEE International. Msnfield, Texas,USA;IEEE, 2015;1-3.

[6] Chen C L, Hu Y, Luo W, et al. A high voltage analog multiplexer with digital calibration for battery management systems [C]// IC Design & Technology (ICIDT), 2012 IEEE International Conference on. Washington,DC, USA; IEEE, 2012;1-4.

作者简介:

满雪成 男,(1989-),硕士研究生,研究方向为新能源汽车电池管理系统核心芯片.
乌力吉(通讯作者) 男,(1965-),博士,博士生导师,研究方向为汽车电子与信息安全.
E-mail:ljiwu@tsinghua.edu.cn.
张向民 男,(1966-),硕士,助理研究员,研究方向为汽车电子与信息安全.