DOI: 10.11992/tis.201701002
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20170703.1601.004.html

个性化信息推荐方法研究

姜信景, 齐小刚, 刘立芳
（西安电子科技大学 数学与统计学院, 陕西 西安 710071）

摘　要: 随着信息技术和互联网的发展, 人们进入了信息过量且愈发碎片化的时代。当前, 个性化信息推送是用户获取网络信息的有效渠道。由于信息的更新速度快和用户兴趣更新等问题, 传统的推荐算法很少关注甚至忽略上述因素, 造成最终的推荐结果欠佳。为了给用户更好的个性化推荐服务, 论文首次引入截取因子, 提出了组合推荐算法(CR 算法)。该算法的实质是将截取因子引入到基于内容的推荐算法与基于用户的协同过滤算法中, 进而生成混合推荐算法。在推荐列表中, CR 算法产生的推荐结果由两部分组成: 一部分由混合推荐算法生成, 另一部分由基于用户的协同过滤算法生成。根据信息的发布时间, 决定该信息由哪类算法产生推荐: 当浏览时间与当前时间的间隔不大于某个值时, 采用混合推荐算法; 否则, 直接采用基于用户的协同过滤算法。基于真实数据的实验结果表明, CR 算法优于同类算法。

关键词: 网络信息; 截取因子; 信息推送; 基于内容的推荐; 基于内容相似的协同过滤; 基于行为相似的协同过滤; 混合推荐; 组合推荐
中图分类号: TP18; O29　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)02−0189−07

中文引用格式: 姜信景, 齐小刚, 刘立芳. 个性化信息推荐方法研究[J]. 智能系统学报, 2018, 13(2): 189–195.
英文引用格式: JIANG Xinjing, QI Xiaogang, LIU Lifang. Research on the recommendation method of personalized information[J]. CAAI transactions on intelligent systems, 2018, 13(2): 189–195.

Research on the recommendation method of personalized information

JIANG Xinjing, QI Xiaogang, LIU Lifang
(School of mathematics and statistics, Xi dian University, Xi'an 710071, China)

Abstract: It's an excessively informational and more fragmented era that is contributed to the development of information technology and the Internet. At present, personalized recommendation is a relatively effective way to help users gain various network information. Recommendations may not be ideal as the traditional algorithms rarely focus on the fast speed of information updating and change of users interests. We propose a combined recommendation algorithm by introducing an interception factor and calls it the CR algorithm. The core idea of it is to introduce the interception factor to the content-based recommendation algorithm and user-based collaborative filtering algorithm. The mixed recommendation consists of the content-based recommendation algorithm and user-based collaborative filtering algorithm. Recommending results of CR algorithm are divided into the outcomes produced by mixed recommendation algorithm and the user-based collaborative filtering algorithm. It is the publishing time of information that decides which algorithm should be chosen to produce recommendations: the mixed recommendation algorithm is selected when the difference between browsing time and message publishing time does not exceed some threshold, or directly chooses the user-based collaborative filtering. Simulation results based on real data show the algorithm we proposed is superior to other existing algorithms.

Keywords: network information; interception factor; information push; content-based recommendation; behavior-based similarity collaborative filtering; content-based similarity collaborative filtering; mixed recommendation; combined recommendation

收稿日期: 2017−01−04.　**网络出版日期:** 2017−07−02.
基金项目: 国家自然科学基金项目 (61572435, 61472305); 陕西省自然科学基金项目 (2015JZ002, 2015JM6311); 浙江省自然科学基金项目 (LZ16F020001); 宁波市自然科学基金项目 (2016A610035).
通信作者: 齐小刚. E-mail: xgqi@xidian.edu.cn.

随着互联网的迅速发展, 海量的网络信息大大超过用户的想象。面对如此浩瀚的信息, 用户如何从中能够阅读到满足其需求的信息是迫待解决的关键问题。个性化信息推荐主要处理消息和用户的匹

配问题，即对于一个信息而言，通过个性化推荐算法能够从众多用户中找到需要了解它的用户集；对于用户而言，通过个性化信息推荐能够从众多的网络消息中快速地发现其需求的信息集。目前，针对信息的推荐方法主要包括：基于内容的推荐[1-3]、基于知识的推荐[4-5]、协同过滤推荐[6-7]、混合推荐[7-9]以及其他推荐[10-15]。

基于内容的信息推荐算法[1]是根据对用户的历史行为分析进行建立用户模型，并向用户推荐与其模型比较匹配的信息。该推荐算法的核心就是挖掘用户的历史行为数据，找到与其相似的信息进行推荐，所以基于内容的推荐算法能够准确捕获用户的兴趣，能够为其推荐新出现的信息。但是，由于用户的兴趣随着时间快速变化，以及该方法仅仅推荐与其模型比较匹配的信息，所以该方法在获取用户的潜在兴趣以及推荐列表多样性方面存在不足。基于知识的推荐算法[5, 16]是针对特定领域建立规则，利用基于实例和规则的推理，实现对用户推荐。比如，效用知识是指一个项目为何满足某一特定用户的知识，其既能产生推荐也可以解释产生该推荐的原因。该方法的优点是把用户的需求直接映射到产品上以及考虑非产品属性，但是其缺点为知识难以获得并且推荐是静态的。协同过滤推荐算法[2, 6–7]是推荐系统中最基本的算法，其包括基于用户的协同过滤算法和基于物品的协同过滤算法。基于用户的协同过滤算法的思想是根据目标用户的历史行为找到与其相似的用户，然后将它们比较喜欢的但目标用户没有发现的东西推荐给目标用户。基于物品的协同过滤的思想与其类似。该方法的优点在于不需要领域知识、推荐多样性好以及可以挖掘用户的潜在兴趣，但是其缺点包括存在冷启动问题、系统开始时推荐质量差、可扩展性差以及质量取决于历史数据集等。

由于信息的实时性与用户兴趣的不固定性，在上述推荐方法的启发下，论文提出了组合推荐算法——CR 算法。该算法的基本思想是：首先是对目标用户历史行为日志进行发掘处理，根据基于内容的推荐算法生成用户的现存配置文件与当前兴趣配置文件；然后，由基于用户行为的协同过滤算法与基于用户内容的协同过滤算法共同生成用户的潜在配置文件；紧接着由现存用户配置文件与潜在配置文件共同产生用户的混合配置文件；最后根据信息集中信息的发布时间决定其有哪种方法产生推荐。当信息发布时间与当前时间的差小于某个阈值时，采用混合推荐算法；当消息发布时间与当前时间的差不小于上述阈值时，采用基于用户的协同过滤算法。

1　个性化推荐方法

1.1　问题定义

定义 1　主要特征词：设$F=(f_1,f_2,\cdots,f_n)$为信息集，我们把表示信息内容的词称为主要特征词，把有序序列$K=(k_1,k_2,\cdots,k_l)$称为主要特征词序列，其中$k_1,k_2,\cdots,k_l$表示主要特征词，l 表示主要特征词的数目。

定义 2　用户现存配置文件：对于任何用户，把其阅读过的信息生成的文件称为用户现存配置文件，并将用户现存配置文件表示成向量形式$\mathbf{UCF}=(\mathrm{wc}_1,\mathrm{wc}_2,\cdots,\mathrm{wc}_i,\cdots,\mathrm{wc}_l)$，其中$\mathrm{wc}_i$表示在用户现存配置文件中主要特征词$k_i$的权重。

定义 3　用户当前兴趣配置文件：对于用户u，把其最新阅读过的s个信息生成的文件称为用户u的当前兴趣配置文件，并将用户u的当前兴趣配置文件表示为$\mathbf{UCF}_{us}=(\mathrm{wc}_{u^s1},\mathrm{wc}_{u^s2},\cdots,\mathrm{wc}_{u^si},\cdots\mathrm{wc}_{u^sl})$，其中$\mathrm{wc}_{u^si}$表示在用户$u$的当前兴趣文件中主要特征词$k_i$的权重。

定义 4　用户潜在配置文件：对于任何用户，利用协同过滤的方法预测主要特征词的权重，进而获得用户潜在配置文件，其能够被表示为向量形式$\mathbf{UMF}=(\mathrm{wm}_1,\mathrm{wm}_2,\cdots,\mathrm{wm}_i,\cdots,\mathrm{wm}_l)$，其中$\mathrm{wm}_i$表示在用户潜在配置文件中主要特征词$k_i$的权重。

定义 5　用户混合配置文件：对于任何用户，融合上述的用户当前兴趣配置文件和用户潜在配置文件，获得其用户混合配置文件，其能够被表示成向量形式$\mathbf{UBF}=(\mathrm{wb}_1,\mathrm{wb}_2,\cdots,\mathrm{wb}_i,\cdots,\mathrm{wb}_l)$，其中$\mathrm{wb}_i$表示在用户混合配置文件中主要特征词$k_i$的权重。

通过上面对一些概念的定义，下面给出论文的设计思路，如图 1 所示。

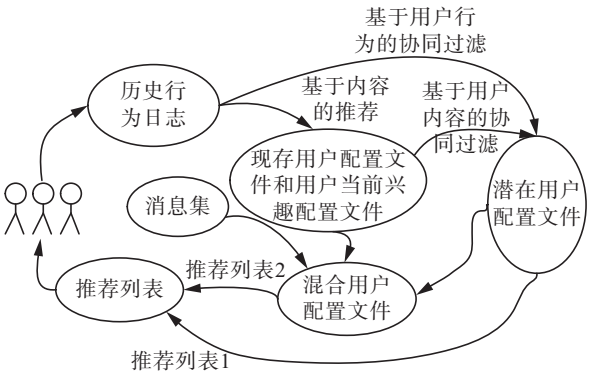

图 1　方案框架

Fig. 1　Scheme framework

1.2　现存用户配置文件

由于信息时效性强与用户的浏览兴趣并不是永久的，而是跟随社会流行和热点话题变化而变化，所以在进行信息推荐时需要考虑到用户的兴趣偏好

变化。为此, 论文引进截取因子、时间因子以及对用户的历史数据进行处理。

1.2.1　向量空间模型

给定信息集$F=(f_1,f_2,\cdots,f_i,\cdots,f_n)$和主要特征词序列$K=(k_1,k_2,\cdots,k_i,\cdots,k_l)$, f_i 能够被表示为向量空间模型$f_i=(w_{i1},w_{i2},\cdots w_{il})$, 其中$w_{ij}$表示特征词$k_j$在信息$f_i$中的权重。$w_{ij}=0$表示$k_j$不在$f_i$中出现。论文利用TF – IDF[17]方法对文本信息进行处理。计算w_{ij}的公式如下:

$$w_{ij}=tf(i,j)\times\log[1+n/n(j)]/\max\text{Other}(i,j) \quad (1)$$

式中: $tf(i,j)$是出现在f_i中的k_j的数目, $n(j)$表示出现k_j的信息数量,$\max\text{Other}(i,j)$是出现在f_i的其他特征词的最大数目。可以看出, 信息集 F 可以表示成一个权重矩阵。

1.2.2　用户现存配置文件、时间因子以及用户当前兴趣配置文件

鉴于用户的兴趣会随着时间的变化而快速变化, 而且用户的浏览兴趣往往和刚刚浏览过的前几条信息有很大的关联。所以论文在处理文本信息时首先对用户已阅读消息的浏览时间进行升序排序, 进而生成现存用户配置文件UCF, 然后选取最后浏览的 s 个信息用于生成用户u的当前兴趣配置文件UCF_{us}。设用户u已阅读的按浏览时间降序排列的信息集表示为$F_u=(f_{u1},f_{u2},\cdots,f_{ui},\cdots,f_{un_u})$, 所以最新浏览的 s 个信息集合为$F_{us}=\{f_{u1},f_{u2},\cdots,f_{un_{us}}\}$, t_i是用户u阅读信息f_{ui}的时间。时间因子能够被定义为

$$u_{i+1}=1/(1+\alpha|t_{t+1}-t_i|) \quad (2)$$

式中: α是时间衰减参数, 通过实验确定; F_u、F_{us}是F的子集。所以F_u、F_{us}也可以表示为一个权重矩阵。获得用户u的现存配置文件UCF和当前兴趣配置文件UCF_{us}的过程如算法 1。

算法 1

输入　F_u, F_{us}, 用户 u 阅读消息 F_{ui} 的时间 t_i, α;

1)　$\text{UCF}_u=(\text{wc}_{u1},\text{wc}_{u2},\cdots,\text{wc}_{ul})$, $\text{UCF}_{\text{us}}=(\text{wc}_{u^s1},\text{wc}_{u^s2},\cdots,\text{wc}_{u^sl})$

2)　for each $k_j\in K$ do

3)　　$\text{wc}_{uj}=0$, $\text{wc}_{u^sj}=0$

4)　end for

其中, 对用户浏览的信息集从最早阅读的消息开始, 依次到最新阅读的信息进行下述 5) ~ 13) 的操作。

5)　for each $f_{ui}\in F_{\text{us}}$ do

6)　　if $i==s$ then

7)　　$\text{wc}_{u^sj}=w_{uij}$

8)　　else

9)　　$\mu_i=\dfrac{1}{1+\alpha|t_i-t_{i-1}|}$

10)　　$\text{wc}_{u^sj}=\mu_i*\text{wc}_{u^sj}+w_{uij}$

11)　　end for

12)　end for

13)　for each $f_{ui}\in F_u$ do

14)　　if $s==n_u$ then

15)　　$\text{wc}_{ui}=w_{uij}$

16)　　else

17)　　$\mu_i=\dfrac{1}{1+\alpha|t_i-t_{i-1}|}$

18)　　$\text{wc}_{uj}=\mu*\text{wc}_{uj}+w_{uij}$

19)　　end if

20)　end for

输出　UCF_u, UCF_{us}。

1.3　潜在配置文件

由于用户的浏览兴趣并不是永久的, 是跟随社会流行和热点话题变化而变化, 所以推荐信息的列表不应该仅仅包括用户现存兴趣, 也应该包括用户的潜在兴趣。考虑到信息的特殊性, 本文利用同时考虑行为相似和内容相似的基于用户的协同过滤方法来寻找目标用户的相似用户和潜在兴趣。

1.3.1　混合相似性的计算

由于信息的特殊性, 基于信息的协同过滤应考虑: 行为相似$\text{simAct}(u,v)$和内容相似$\text{simCon}(u,v)$的计算。

给定信息集F_{us}和F_v, 用户u的当前兴趣文件$\text{UCF}_{\text{us}}=(\text{wc}_{u^s1},\text{wc}_{u^s2},\cdots,\text{wc}_{u^si},\cdots\text{wc}_{u^sl})$, 用户$v$的现存配置文件$\text{UCF}_v=(\text{wc}_{v1},\text{wc}_{v2},\cdots,\text{wc}_{vl})$。则用户$u$与用户$v$的行为相似和内容相似的计算如下:

$$\text{simAct}(u,v)=|F_{us}\cap F_v|/\sqrt{|F_{us}|\times|F_u|} \quad (3)$$

$$\text{simCon}(u,v)=\left(\text{CUF}_{us}\cdot\text{CUF}_v^{\text{T}}\right)/\sqrt{|\text{CUF}_{us}|\times|\text{CUF}_v|} \quad (4)$$

根据式 (3) 和 (4), 混合相似度计算公式如下:

$$\text{sim}(u,v)=\beta\times\text{simAct}(u,v)+(1-\beta)\times\text{simCon}(u,v) \quad (5)$$

式中: 系数$\beta\in[0,1]$, 通过实验来决定。获得 u 和 v 相似性的过程如算法 2。

算法 2

输入　F_{us}、F_v、UCF_{us} 和 UCF_V, 系数 β;

1)　$\text{numSimNews}=0$, $\text{innerPro}=0$, $\text{norm}_u=0$, $\text{norm}_v=0$

2)　for each $f_{ui}\in F_{us}$ do

3)　　for each $f_{uj}\in F_v$ do

4)　　if $f_{ui}=f_{uj}$ then

5)　　$\text{numSimNews}=1+\text{numSimNews}$

6)　　end if

7)　　end for

8)　end for

9)　　$\text{simAct}(u,v)=\text{numSimNews}/\text{sqrt}(\text{len}(F_u)*$

len(F_v))

10) for each $wc_{u^s j} \in \mathrm{UCF}_{us}$ do

11) innerPro = innerPro + $\mathrm{wc}_{u^s j} * \mathrm{wc}_{vj}$

12) $\mathrm{norm}_{us} = \mathrm{norm}_u + \mathrm{wc}_{u^s j} * \mathrm{wc}_{u^s j}$

13) $\mathrm{norm}_v = \mathrm{norm}_v + \mathrm{wc}_{vj} * \mathrm{wc}_{vj}$

14) end for

15) $\mathrm{simCon}(u,v)$ = innerPro/(sqrt(norm_{us}) * sqrt (norm_v))

16) $\mathrm{sim}(u,v)=\beta*\mathrm{simAct}(u,v)+(1-\beta)*\mathrm{simCon}(u,v)$

输出 用户 u 和 v 的相似性$\mathrm{sim}(u,v)$。

1.3.2 潜在用户配置文件和相似用户文件的生成

目标用户u和其他用户的相似性通过算法 2 计算。选择相似性最大的h个用户构造相似用户文件。然后通过加权计算获得目标用户u的潜在用户配置文件UMF。

给定相似用户集$U_u = \{v_1, v_2, \cdots, v_h\}$, 用户$v_i$的现存配置文件$\mathrm{UCF}_{v_i} = (\mathrm{wc}_{v_i 1}, \mathrm{wc}_{v_i 2}, \cdots, \mathrm{wc}_{v_i l})$, 用户$u$和用户 v_i 的相似性为$\mathrm{sim}(u,v_i)$。利用式 (6) 计算在MUF_u中的k_j的权重。获得潜在用户配置文件的过程如算法 3。

$$\mathrm{wm}_{uj} = \sum_{v_i j}\left[\mathrm{wc}_{v_i j} \times \mathrm{sim}(u,v_i) / \sum_{v_a \in U_u} \mathrm{sim}(u,v_a)\right] \quad (6)$$

算法 3

输入 $U_u = \{v_1, v_2, \cdots, v_h\}$, $v_i \in U_u$, $sim(u,v_i)$, UCF_{vi};

1) sumSim = 0, $\mathrm{UMF}_u = (\mathrm{wm}_{u1}, \mathrm{wm}_{u2}, \cdots, \mathrm{wm}_{ul})$

2) for each $k_j \in K$ do

3) $\mathrm{wm}_{ij} = 0$

4) end for

5) for each $v_a \in U_u$ do

6) sumSim = sumSim + $\mathrm{sim}(u,v_a)$

7) end for

8) for each $v_i \in U_u$ do

9) for each $k_j \in K$ do

10) $\mathrm{wm}_{uj} = \mathrm{wm}_{uj} + \mathrm{wc}_{v_i j} * \mathrm{sim}(u,v_i)$/sumSim

11) end for

12) end for

输出 UMF_u。

1.4 用户混合配置文件的生成

用户混合配置文件 UBF 能够在获得目标用户的当前兴趣配置文件 UCF_s 和潜在配置文件 UMF 后, 通过对 UCF, UMF 上的每个主要特征词加权得到。设用户u的$\mathrm{UCF}_{us} = (\mathrm{wc}_{u^s 1}, \mathrm{wc}_{u^s 2}, \cdots, \mathrm{wc}_{u^s l})$, $\mathrm{UMF}_u = (\mathrm{wm}_{u1}, \mathrm{wm}_{u2}, \cdots, \mathrm{wm}_{ul})$, $\mathrm{UBF}_u = (\mathrm{wb}_{u1}, \mathrm{wb}_{u2}, \cdots, \mathrm{wb}_{ul})$。利用式 (7) 计算$\mathrm{wb}_{uj}$。

$$\mathrm{wb}_{uj} = \gamma \mathrm{wc}_{u^s j} + (1-\gamma)\mathrm{wm}_{uj} \quad (7)$$

式中: $\gamma \in [0,1]$, 其值通过实验确定。获得用户u的混合用户配置文件UBF_u过程如算法 4。

算法 4

输 入 $\mathrm{UMF}_u = (\mathrm{wm}_{u1}, \mathrm{wm}_{u2}, \cdots, \mathrm{wm}_{ul})$, $\mathrm{UMF}_u = (\mathrm{wc}_{u1}, \mathrm{wc}_{u2}, \cdots, \mathrm{cm}_{ul})$, γ

1) $\mathrm{UBF}_u = (\mathrm{wb}_{u1}, \mathrm{wb}_{u2}, \cdots, \mathrm{wb}_{ul})$

2) for each $k_i \in K$ do

3) $\mathrm{wb}_{uj} = 0$

4) $\mathrm{wb}_{uj} = \gamma \mathrm{wc}_{uj} + (1-\gamma)\mathrm{wm}_{uj}$

5) end for

输出 UBF_u。

1.5 推荐结果的生成

由于信息的时效性和用户兴趣不固定等问题, 在推荐列表中, 信息由两部分组成: l_1、l_2。

l_1部分由混合配置文件生成, 即通过添加时间因子ε_1来限定消息是否采用混合推荐方法: 当消息的发布时间与当前时间的时间间隔小于ε_1, 若满足, 则该文件采用混合推荐方法, 否则将不采用。详细过程如下:

设用户 u 的$\mathrm{BUF}_u = (\mathrm{wb}_{u1}, \mathrm{wb}_{u2}, \cdots \mathrm{wb}_{ul})$, 新闻$d_0 = (\mathrm{wd}_1, \mathrm{wd}_2, \cdots \mathrm{wd}_l)$, 信息$d_0$的发布时间为$t_0$, 当前时间$t_{\mathrm{cur}}$, 阈值$\varepsilon_1$、$\varepsilon_2$。首先检查

$$t_{cur} - t_0 \leqslant \varepsilon_1 \quad (8)$$

若不等式 (8) 成立, 则检查

$$d_0 \cdot \mathbf{BUF}_u^{\mathrm{T}} \geqslant \varepsilon_2 \quad (9)$$

若式 (9) 成立, 则将信息d_0放入l_1中。

l_2部分直接由基于内容相似和行为相似的协同过滤算法生成。详细过程如下:

设用户u的相似用户集$U_u = \{v_1, v_2, \cdots, v_h\}$, 用户$u$和用户$v_i$的相似性为$\mathrm{sim}(u,v_i)$。对于信息$d_0$, 设该信息在用户$u$的相似用户集上的权重为$\{w_{v_1 d_0}, w_{v_2 d_0}, \cdots, w_{v_h d_0}\}$, 那么信息$d_0$相对于用户$u$的权重为

$$w_{ud_0} = \sum_{j=1}^{h} w_{v_j d_0} * \mathrm{sim}(u,v_j) \quad (10)$$

选出相对于用户u的权重较大的消息放入l_2部分。

2 实验和分析

实验数据来源于财新网站 2014 年 3 月份的一万个用户的所有浏览记录。每个浏览记录由用户编号、新闻编号、浏览时间、新闻标题、新闻内容以及发表时间组成。从数据集中抽取阅读超过 25 条的新闻用户作为训练集。令包含在网站给定的测试集中的训练集用户作为测试集, 其中测试集中的用户只有一个测试记录。论文采用F值、召回率(recall)、准确率(precision)和多样性(Diversity)作为评价指标。F值的定义为

$$F = \frac{1}{1/\mathrm{reacll} + 1/\mathrm{precision}} \quad (11)$$

式中recall和precision的定义如下:

$$\text{recall} = \frac{\sum_{u_i \in U} \text{hit}(u_i)}{\sum_{u_i \in U} T(u_i)} \tag{12}$$

式中: U为数据集中用户的集合, $\text{hit}(u_i)$表示推荐给用户u_i的新闻中, 确实在测试集中被该用户浏览的个数。$T(u_i)$为测试集中用户u_i真正浏览的新闻的数目。

$$\text{precision} = \frac{\sum_{u_i \in U} \text{hit}(u_i)}{\sum_{u_i \in U} L(u_i)} \tag{13}$$

式中: $\text{hit}(u_i)$的定义同上, $L(u_i)$表示用户u_i的新闻推荐列表的长度。在进行实验时, 对于消息$f_0 = (w_{01}, w_{02}, \cdots, w_{0l})$, 若$k_i$在$f_0$中出现的频率排在前 10, 则$w_{0i} = 1$, 否则$w_{0i} = 0$。设$s = 5$, $\alpha = 10^{-6}$, $\gamma = 0.5$, $\varepsilon_1 = 3\,600$, $\varepsilon_2 = 0.5$。

首先验证β的取值, 由于测试集中每个用户只有一个测试记录, 所以用F值不能获得好的效果。因此, 在实验仿真中, 论文采用recall。表 1 是推荐列表长度为 20 时, recall与β的关系。

表 1　recall 与β的关系

Table 1　Relationship between recall and β

β	recall	β	recall
0.0	0.607	0.5	0.778
0.1	0.657	0.6	0.788
0.2	0.723	0.7	0.791
0.3	0.745	0.8	0.791
0.4	0.776	0.9	0.807

通过实验数据显示, 当$\beta = 0.9$时, recall 最好。接着验证F值、recall 和 precision。

在图 2 中, 随着推荐列表长度的增加, 上述 6 种方法除 CBR(基于内容的推荐算法) 外, F 值都逐渐减少。在相同的推荐列表长度的情况下。*CR*(组合推荐) 的 F 值最大, 除个别点, ICFBBS(改进的基于行为相似的协同过滤)、ICFCBS(改进的基于内容相似的协同过滤)、MR(混合推荐)、CFBBS(基于行为相似的协同过滤)、CFCBS(基于内容相似的协同过滤) 依次减少。CBR 的 F 值最小。图 3 为 recall指标随推荐列表长度变化的情况。随着推荐列表长度的增加, 6 种方法的 recall 值都逐渐增加。在相同推荐列表长度的情况下, 除个别点, CR、ICFBBS、ICFCBS、MR、CFBBS、CFCBS 以及 CBR 的 recall值依次减少。图 4 为 precision 指标随推荐列表长度变化的情况。随着推荐列表长度增加, 6 种方法值都逐渐减少。在相同列表长度的情况下, 除个别点, CR、ICFBBS、ICFCBS、MR、CFBBS、CFCBS 以及 CBR 的 Precision 值依次减少。

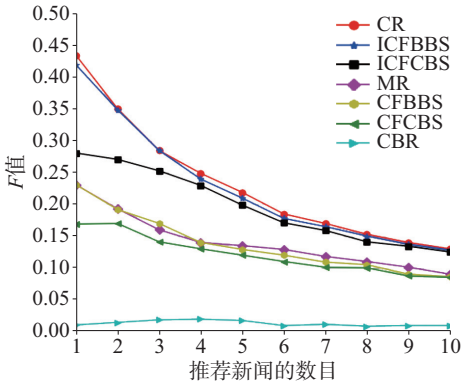

图 2　F 值比较

Fig. 2　Comparison of F

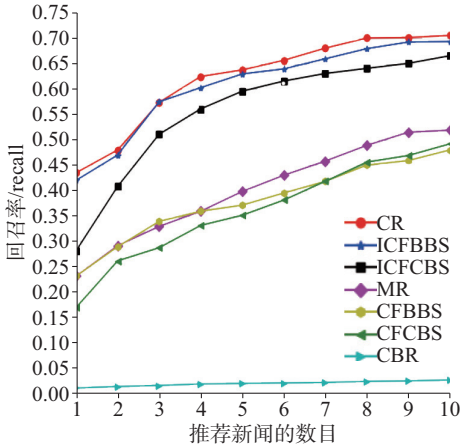

图 3　召回率比较

Fig. 3　Comparison of recall

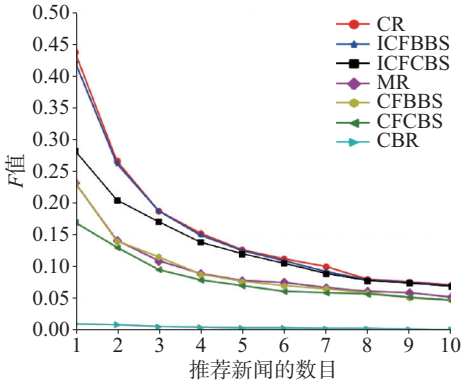

图 4　精确度比较

Fig. 4　Comparison of precision

多样性Diversity描述了推荐列表中物品两两之间的差异性。所以多样性和相似性是对应的, 假设$\text{sim}(i, j) \in [0, 1]$为消息 i 和 j 之间的相似度, 用户u的推荐列表$R(u)$的多样性定义如式 (14):

$$\text{Diversity} = 1 - \frac{\sum_{i,j \in R(u), i \neq j} s(i, j)}{0.5|R(u)|(|R(u)| - 1)} \tag{14}$$

而推荐系统的整体多样性可以定义为所有用户推荐列表多样性的平均值如式 (15):

$$\text{Diversity} = \frac{1}{|U|}\sum_{u\in U}\text{Diversity}(R(u)) \tag{15}$$

图 5 是上述 7 种方法在不同推荐长度下多样性。从图中可以看出，CBR 算法是通过对用户先前消息的内容进行分析，然后推荐与其内容相似的消息，所以在推荐列表中的消息内容相似性特别高，进而多样性很差。ICFBBS、ICFCBS、CFBBS、CFCBS 是目标用户通过找到与其行为相似或者内容相似的用户集，给目标用户推荐用户集中浏览最多的消息，所以多样性比 CBR 好。CR 是混合推荐和直接基于用户的协同过滤算法的组合，所以多样性比 CBR 好，比 ICFBBS、ICFCBS、CFBBS、CFCBS 差。MR 推荐的消息是与用户的兴趣模型相似度较高的消息，所以多样性与 CBR 相似。

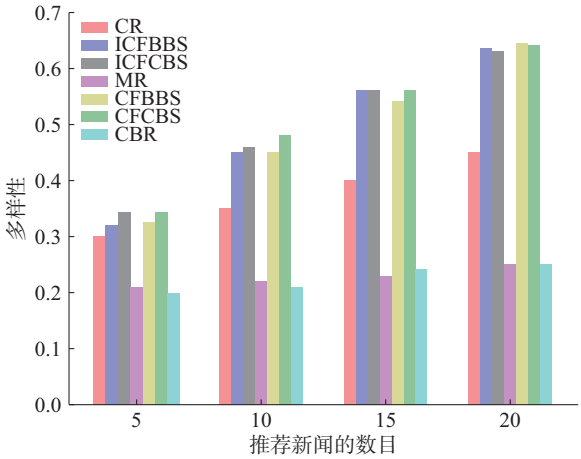

图 5　多样性比较

Fig. 5　Comparison of diversity

此外，CR 方法在进行推荐时，由于对消息的分类推荐，所以推荐所用的时间远远小于基于内容的算法和用户的协同过滤混合推荐算法。

3　结束语

本文首先介绍了个性化信息推荐的传统方法，对基于内容推荐算法和基于协同过滤算法进行了简单说明。针对信息的特点，本文提出了组合推荐算法 (CR 算法)。针对该算法设计实验并分析了实验结果。数据显示 CR 方法显著优于其他同类方法。但是随着信息属性和用户权限的细分，通用的推荐算法已不适应某些特殊的信息领域，下一步，可以试着通过改造上述算法的结构进行比较精准的推荐。

参考文献:

[1] 李佳珊. 个性化新闻推荐引擎中新闻分组聚类技术的研究与实现[D]. 北京: 北京邮电大学, 2013.
LI Jiashan. Research and implementation of text clustering for personalized news recommandation system[D]. Beijing: Beijing University of Posts and Telecommunications, 2013.

[2] 项亮. 推荐系统实践[M]. 北京: 人民邮电出版社, 2012.

[3] BALABANOVIĆ M, SHOHAM Y. Fab: content-based, collaborative recommendation[J]. Communications of the ACM, 1997, 40(3): 66–72.

[4] MANDL M, FELFERNIG A, TEPPAN E, et al. Consumer decision making in knowledge-based recommendation[J]. Journal of intelligent information systems, 2011, 37(1): 1–22.

[5] LI Xiaohui, MURATA T. A knowledge-based recommendation model utilizing formal concept analysis and association [C]//Proceedings of the 2nd International Conference on Computer and Automation Engineering. Singapore, 2010: 221–226.

[6] GARCIN F, ZHOU Kai, FALTINGS B, et al. Personalized news recommendation based on collaborative filtering[C]// Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology. Washington, DC, USA: IEEE, 2012: 437–441.

[7] DARVISHY A, IBRAHIM H, MUSTAPHA A, et al. New attributes for neighborhood-based collaborative filtering in news recommendation[J]. Journal of emerging technologies in web intelligence, 2015, 7(1): 13–19.

[8] YANG Wu, TANG Rui, LU Ling. A fused method for news recommendation[C]//Proceedings of the 2016 International Conference on Big Data and Smart Computing (BigComp). Hong Kong, China, 2016: 341–344.

[9] LU Zhongqi, DOU Zhicheng, LIAN Jianxun, et al. Content-based collaborative filtering for news topic recommendation[C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin, Texas, USA, 2015: 217–223.

[10] LIU Y, BAO L, GAO L. Trust-based new recommendation algorithm of collaborative filtering combination[J]. Information Japan, 2013, 16(7): 4555–4576.

[11] WANG Jingjin, LIN Kunhui, LI Jia. A collaborative filtering recommendation algorithm based on user clustering and slope one scheme[C]//Proceedings of the 2013 8th International Conference on Computer Science & Education (ICCSE). Colombo, Sri Lanka, 2013: 1473–1476.

[12] CAPELLE M, FRASINCAR F, MOERLAND M, et al. Semantics-based news recommendation[J]//Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. Craiova, Romania, 2012: 27.

[13] CUI Limeng, SHI Yong. A Method based on one-class SVM for news recommendation[J]. Procedia computer sci-

ence, 2014, 31: 281–290.

[14] REN Rui, ZHANG Lingling, CUI Limeng, et al. Personalized financial news recommendation algorithm based on ontology[J]. Procedia computer science, 2015, 55: 843–851.

[15] LOMMATZSCH A, KENTER T, DE VRIES A P, et al. Real-time news recommendation using context-aware ensembles[M]//DE RIJKE M. Advances in Information Retrieval. Cham, Germany: Springer, 2014.

[16] 杨博, 赵鹏飞. 推荐算法综述[J]. 山西大学学报: 自然科学版, 2011, 34(3): 337–350.
YANG Bo, ZHAO Pengfei. Review of the art of recommendation algorithms[J]. Journal of Shanxi university: natural science edition, 2011, 34(3): 337–350.

[17] 路永和, 李焰锋. 改进 TF—IDF 算法的文本特征项权值计算方法[J]. 图书情报工作, 2013, 57(3): 90–95.
LU Yonghe, LI Yanfeng. Improvement of text feature weighting method based on TF-IDF algorithm[J]. Library and information service, 2013, 57(3): 90–95.

作者简介:

姜信景, 男, 1988 年生, 硕士研究生, 主要研究方向为个性化信息推荐。

齐小刚, 男, 1973 年生, 教授, 博导, 博士, 主要研究方向为系统建模与故障诊断。

刘立芳, 女, 1972 年生, 教授, 博士, 主要研究方向为数据处理与智能计算。