文章编号:1007-5321(2019)01-0114-06 **DOI**:10.13190/j.jbupt.2018-067

基于 VSM 和 Bisecting *K*-means 聚类的新闻推荐方法

袁仁进, 陈 刚, 李 锋, 魏双建
(信息工程大学 地理空间信息学院, 郑州 450052)

摘要: 针对海量新闻数据给用户带来的困扰,为提升用户阅读新闻的个性化体验,提出了融合向量空间模型和 Bisecting *K*-means 聚类的新闻推荐方法. 首先进行新闻文本向量化,使用向量空间模型和 TF-IDF 算法构建出新闻特征向量;采用 Bisecting *K*-means 聚类算法对新闻特征向量集进行聚类;然后将已聚类的新闻集分为训练集和测试集,根据训练集构建"用户—新闻类别—新闻"三层层次结构的用户兴趣模型;最后采用余弦相似度方法得出新闻推荐结果,并与测试集进行对比分析. 实验以基于用户的协同过滤算法、基于物品的协同过滤算法、结合向量空间模型和 *K*-means 聚类的推荐方法为基准,实验结果表明,该方法具有可行性,在准确率、召回率和 *F* 值上都有所提高.

关 键 词: 个性化推荐; 向量空间模型; Bisecting *K*-means 聚类算法; 用户兴趣模型
中图分类号: TP391.3 **文献标志码**: A

A News Recommendation Method Based on VSM and Bisecting *K*-means Clustering

YUAN Ren-jin, CHEN Gang, LI Feng, WEI Shuang-jian
(Institute of Geospatial Information, Information Engineering University, Zhengzhou 450052, China)

Abstract: Personalized recommendation technology is a good solution to the problem of information overload. In order to improve the user's personalized experience of reading news, a news recommendation method based on the vector space model and Bisecting *K*-means clustering is proposed. Firstly, the news text vectorization is carried out: using the vector space model and TF-IDF algorithm to construct news feature vectors; then Bisecting *K*-means clustering algorithm is utilized to cluster the news feature vector set; after that, the clustered news set is divided into training set and test set, according to the training set, a "user - news category - news" three-level structure of the user interest model is built; finally, the cosine similarity method is used to calculate news recommendation results. The experiments are based on user-based collaborative filtering algorithm, item-based collaborative filtering algorithm, combined vector space model and *K*-means clustering recommendation method, and the results show that the proposed method is feasible, and the accuracy rate, recall rate and *F* value all have been improved.

Key words: personalized recommendation; vector space model; Bisecting *K*-means clustering algorithm; user interest model

信息的爆炸反而使得信息的利用率降低,造成了信息超载现象. 信息的增长使得人们难以从海量

收稿日期: 2018-04-16
基金项目: 国家自然科学基金项目(41301428)
作者简介: 袁仁进(1994—), 男, 硕士生.
通信作者: 陈 刚(1971—), 男, 教授, 博士生导师, E-mail:chengang_vge@sina.com.

资讯中选择自己感兴趣的新闻. 新闻个性化推荐系统被广泛用于解决信息超载问题,为不同的用户提供个性化体验来提高用户阅读满意度.

协同过滤算法是推荐系统的经典算法之一,目前已有学者将协同过滤算法应用到新闻推荐中,并取得了一些成果[1-3]. 但基于协同过滤算法的新闻推荐系统仍存在冷启动问题,同时可解释性较差[4-5]. 在基于内容的推荐算法中,解决了冷启动问题并有较强的解释性[6-7],但其对新闻的分类上采用的是编辑分类,导致存在 2 个问题:一是编辑很难控制分类的粒度;二是编辑的意见不能代表各种用户的意见. 针对新闻分类问题,古万荣等[8]提出了一个基于二次聚类的新闻推荐算法,李佳珊[9]对目前数据挖掘领域的聚类算法进行了总结与分析,但这些方法仅对新闻聚类问题进行了分析与改进.

针对上述问题,为了能保留基于内容的推荐算法的可解释性的优势,同时尽量避免编辑分类存在的问题,提出了一种基于向量空间模型(VSM, vector space model)和 Bisecting *K*-means 聚类的新闻推荐方法. 实验结果表明,与传统的协同过滤算法相比,其预测准确度得到了提高.

1　算法思想和流程框架

推荐系统的经典构建流程主要包括用户建模模块、推荐对象建模模块、推荐算法模块[9]. 推荐算法(见图 1)主要包括:新闻文本向量化、构建用户兴趣模型、相似度计算 3 个流程.

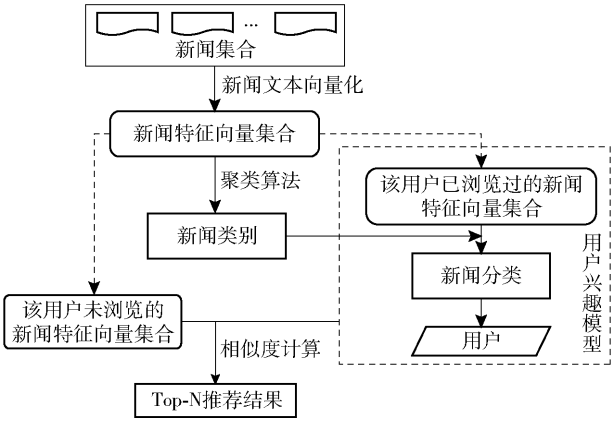

图 1　笔者提出的推荐系统流程框架

1) 新闻文本向量化(见第 2 节). 首先将新闻文本向量化,即使用一个多维向量来表示新闻的内容($d=(w_1,w_2,\cdots,w_m)$),将新闻文本向量化的结果称为新闻特征向量.

2) 构建用户兴趣模型(见第 3 节). 用户兴趣模型构建是本文方法中最关键的一步,首先对已进行向量化处理的新闻集进行自动聚类,接着对每个用户和其已经浏览过的新闻构建三层结构的用户兴趣模型:用户—新闻类别—新闻.

3) 相似度计算. 相似度计算方法很多,最常用的主要有 Pearson 相似性和余弦相似性. 余弦相似性通过计算 2 个向量之间的夹角余弦来衡量两者之间的相关性,由于该方法计算简单,本文中用户兴趣模型和新闻最终都采用向量表示,所以采用余弦相似性方法,如式(1)所示. 计算用户兴趣模型和候选新闻集之间的相似度,选取相似度靠前的 N 条新闻推荐给用户.

$$\mathrm{sim}(i,j)=\cos(i,j)=\frac{ij}{|i||j|} \tag{1}$$

2　新闻文本向量化

向量空间模型是信息检索领域最为经典的计算模型,在该模型中,每个文档用一个特征向量来表示该文档中的多维信息. 考虑到新闻数据的高维性以及为便于新闻聚类从而构建用户兴趣模型,笔者采用向量空间模型来表示新闻特征向量.

给定新闻集 $D=\{d_1,d_2,\cdots,d_n\}$,新闻集的 VSM 表示为

$$\boldsymbol{M}=\begin{bmatrix} w_{11} & \cdots & w_{1m} \\ \vdots & & \vdots \\ w_{n1} & \cdots & w_{nm} \end{bmatrix} \tag{2}$$

其中 w_{ij} 表示为关键词 j 在新闻 i 中的权重. 构建 VSM 的关键有 2 个方面:一是确定关键词集的维度 m;二是权重 w_{ij} 的计算.

1) 关键词集的维度 m:关键词用于表征该文档特性,文献[10]对比分析了多种关键词提取方法,实验得出当关键词个数在(5,8)之间时效果最好. 但关键词数量增加会导致 $\boldsymbol{M}$ 的维度 m 增大,从而引起时间复杂度增加. 在保证文档表征效果前提下,为减少时间开销,提取每篇新闻中的前 5 个关键词来表征该新闻特性,接着采用 TF-IDF 算法得出新闻集中关键词集的维度 m.

2) 权重 w_{ij} 的计算:最常用和有效的权重的计算方法为 TF-IDF 表示法,该方法是信息检索领域的成熟技术.

TF-IDF 算法的计算可以分成词频(TF, tevm frequency)和逆文档频率(IDF, inverse document fre-

quency)两部分,由这两部分的乘积共同决定文档词语的权重. TF-IDF 算法有多种变种形式,笔者采用的计算方法为

$$\mathrm{TF}(i,j)=\frac{\mathrm{count}(i,j)}{\mathrm{size}(j)} \tag{3}$$

$\mathrm{count}(i,j)$表示关键词 i 在新闻 j 中的频数,$\mathrm{size}(j)$表示新闻 j 的总数. IDF 计算公式为

$$\mathrm{IDF}(i)=\log\frac{N}{n(i)} \tag{4}$$

N 表示新闻集总数,$n(i)$表示关键词 i 出现过的新闻数量. 权重由 TF 和 IDF 计算为

$$w_{(i,j)}=\mathrm{TF}(i,j)\,\mathrm{IDF}(i) \tag{5}$$

为使权重值处于[0,1]区间内且新闻能够用等长向量表示,使用余弦归一化的方式对权重进行归一化处理,词语的权重计算公式为

$$w_{(i,j)}=\frac{\text{TF-IDF}(i,j)}{\sqrt{\sum_{i=1}^{T}\text{TF-IDF}(i,j)^2}} \tag{6}$$

3 构建用户兴趣模型

3.1 基于 Bisecting *K*-means 算法的新闻分类

目前在数据挖掘领域的聚类算法主要包括基于模型的算法、基于网格的算法、基于密度的算法、基于距离的算法 4 种[9]. 笔者研究的数据为新闻数据,具有海量、高维等特点,同时采用了向量空间模型来表示新闻的文本特征,因此基于以上考虑,采用基于距离的算法作为新闻聚类方法. Abuaiadah D[11]研究得出 *K*-means 聚类算法的改进算法——Bisecting *K*-means 聚类算法收敛速度更快,聚类效果更优. 综上,在向量空间模型基础上,采用 Bisecting *K*-means 聚类算法可实现新闻的分类.

Bisecting *K*-means 聚类算法类似于一种层次聚类算法,具有不需要先验簇数量的优点,改善了 *K*-means聚类算法中初始质心位置对聚类结果产生影响的问题. 该算法中采用误差平方和 S 来衡量簇的质量,S 的计算如

$$S=\sum_{m=1}^{k}\sum^{p_i\in C_i}\mathrm{distance}(p_i,c_i)^2 \tag{7}$$

其中:c_i 为 C_i 簇的质心点,p_i 为 C_i 簇中的其他数据点,$\mathrm{distance}(p_i,c_i)^2$ 为 p_i 与 c_i 之间的欧氏距离. 在该算法中,关键一步在于选取需要进行二分的簇,该判断规则为:对目前存在的 k 个簇,从第一个簇开始运用 *K*-means 聚类算法进行二分,计算该簇的总误差 S_1 值以及剩余簇的总误差 S_2 值,得出 S_1+S_2 的误差和,接着循环遍历这 k 个簇,最终选取 S_1+S_2 的误差和最小的簇进行二分.

3.2 用户兴趣模型构建

通过用户已经浏览过的新闻数据构建用户兴趣模型,用户兴趣模型采用 3 层层次结构表示:用户–新闻类别–新闻. 如图 2 所示,第 1 层节点为用户,第 2 层节点为用户浏览的新闻类别,第 3 层节点为用户浏览过的新闻.

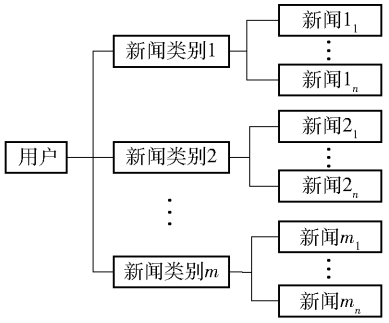

图 2 基于层次结构的用户兴趣模型

若用户浏览过 m 个不同的新闻类别,则用户兴趣模型可用如下模型表示:

$$\mathrm{user}=\{(\boldsymbol{T}_1,w_1,n_1),(\boldsymbol{T}_2,w_2,n_2),\cdots,(\boldsymbol{T}_m,w_m,n_m)\} \tag{8}$$

其中:$\boldsymbol{T}_i$ 表示第 i 个新闻类别特征向量,w_i 表示第 i 个新闻类别的权重,n_i 表示第 i 个新闻类别包含的用户浏览过的新闻的数量.

某个新闻类别的特征向量根据该类别所包含的所有已浏览过的新闻特征向量根据兴趣度加权平均求出,即第 i 个新闻类别特征向量 $\boldsymbol{T}_i$ 的计算公式为

$$\boldsymbol{T}_i=\frac{\sum_{e_j\in E_j}\boldsymbol{e}_jI_j}{\sum_{e_j\in E_j}I_j} \tag{9}$$

其中:E_j 为新闻类别 i 中的用户浏览过的新闻集合,$\boldsymbol{e}_j$ 为新闻特征向量,I_j 为该类别中第 j 个新闻的用户兴趣度,用户浏览过某新闻即表示用户对该新闻有兴趣,因此将 I_j 设为 1,则式(9)可简化为

$$\boldsymbol{T}_i=\frac{\sum_{e_j\in E_j}\boldsymbol{e}_j}{n_i} \tag{10}$$

w_i 的值根据第 i 个新闻类别中用户浏览过的新闻数量占总共浏览过的新闻数量的权重来计算,如

$$w_i = \frac{n_i}{\sum_{i=1}^{m} n_i} \tag{11}$$

在计算时，用户兴趣模型表示为

$$\boldsymbol{V}_{\text{user}} = (w_1\boldsymbol{T}_1, w_2\boldsymbol{T}_2, \cdots, w_m\boldsymbol{T}_m)^{\mathrm{T}} \tag{12}$$

最终，使用余弦相似度计算候选新闻 d_i 与用户之间的相似性，计算公式如

$$\text{sim}(\text{user}, \boldsymbol{V}_{d_i}) = \cos(w_i\boldsymbol{T}_i^{\mathrm{T}}, \boldsymbol{V}_{d_i}) \tag{13}$$

$w_i\boldsymbol{T}_i^{\mathrm{T}}$ 为候选新闻 d_i 所属新闻类别的特征向量，$\boldsymbol{V}_{d_i}$ 为 d_i 的特征向量.

4　实验及对比分析

4.1　实验环境与数据

使用 DataCastle 提供的用户浏览新闻数据集，该数据集从国内某著名财经新闻网站——财新网随机采集，总共包括 1 万名用户在 2014 年 3 月的所有新闻浏览记录，共 116 225 条. 每条浏览记录包括用户编号、新闻编号、浏览时间以及新闻文本内容等. 实验采用 python 的第三方库——jieba 分词器进行分词，根据实际新闻内容采用改进的哈尔滨工业大学信息检索中心的停用词表去除停用词.

在实验数据预处理阶段，首先将数据集中用户浏览记录少于 40 条的用户删除，共获得 279 名用户的 30 865 条浏览记录；接着使用 python 中的 scikit-learn 库将数据集切分为训练集和测试集，训练集和测试集之间的比例选择为 5∶1. 为增强结果的可靠性，每种方法都重复试验 5 次取平均值作为实验结果.

4.2　实验评估指标

采用准确率、召回率和 F 值作为实验的评估指标，准确率和召回率使用混淆矩阵表示，见表 1.

表 1　混淆矩阵

喜好	被推荐	未被推荐
喜欢	True Positive(TP)	False Negative(FN)
不喜欢	False Positive(FP)	True Negative(TN)

准确率 P、召回率 R 的计算公式为

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

$$F = \frac{2PR}{P + R} \tag{16}$$

4.3　实验结果与分析

为验证提出方法的可行性和推荐性能，实验以基于用户的协同过滤算法(user-based CF)、基于物品的协同过滤算法(item-based CF)、结合 VSM 和 *K*-means聚类推荐算法(简称 VSM + *K*-means算法)为 Baseline，对实验结果进行对比分析，推荐结果数量 N 考虑了(10,15,20,25,30)共 5 种情况.

1) user-based CF 和 item-based CF 最优评估指标确定

基于用户的协同过滤算法和基于物品的协同过滤算法作为 2 种经典的个性化推荐算法，其推荐结果的准确度与最近邻的个数紧密相关，实验中，考虑的最近邻个数 K 为 5、10、15、20 共 4 种情况，对应的评估指标如表 2、表 3 所示.

表 2　user-based CF 在不同 *K* 参数下的性能　%

K	$N=10$	$N=15$	$N=20$	$N=25$	$N=30$
5	$R=16.92$	$R=21.29$	$R=23.91$	$R=25.75$	$R=27.42$
	$P=30.11$	$P=25.26$	$P=21.27$	$P=18.32$	$P=16.26$
	$F=21.67$	$F=23.11$	$F=22.51$	$F=21.41$	$F=20.41$
10	$R=17.28$	$R=21.47$	$R=24.09$	$R=26.19$	$R=27.78$
	$P=30.75$	$P=25.47$	$P=21.43$	$P=18.64$	$P=16.48$
	$F=22.13$	$F=23.30$	$F=22.68$	$F=21.78$	$F=20.69$
15	$R=17.08$	$R=20.91$	$R=23.63$	$R=25.54$	$R=27.60$
	$P=30.39$	$P=24.80$	$P=21.02$	$P=18.18$	$P=16.37$
	$F=21.87$	$F=22.69$	$F=22.25$	$F=21.24$	$F=20.55$
20	$R=16.52$	$R=20.67$	$R=23.31$	$R=25.28$	$R=27.14$
	$P=29.39$	$P=24.52$	$P=20.73$	$P=17.99$	$P=16.09$
	$F=21.15$	$F=22.43$	$F=21.94$	$F=21.02$	$F=20.20$

表 3　item-based CF 在不同 *K* 参数下的性能　%

K	$N=10$	$N=15$	$N=20$	$N=25$	$N=30$
5	$R=15.21$	$R=19.34$	$R=21.97$	$R=23.67$	$R=25.86$
	$P=28.54$	$P=23.62$	$P=21.43$	$P=17.13$	$P=15.21$
	$F=19.84$	$F=21.27$	$F=21.70$	$F=19.88$	$F=19.15$
10	$R=15.36$	$R=19.87$	$R=22.19$	$R=24.20$	$R=26.34$
	$P=28.65$	$P=23.79$	$P=21.67$	$P=17.58$	$P=15.69$
	$F=20.00$	$F=21.65$	$F=21.93$	$F=20.37$	$F=19.67$
15	$R=15.48$	$R=20.03$	$R=22.26$	$R=24.51$	$R=26.64$
	$P=28.73$	$P=24.12$	$P=21.75$	$P=17.66$	$P=15.83$
	$F=20.12$	$F=21.89$	$F=22.00$	$F=20.53$	$F=19.86$
20	$R=15.26$	$R=19.05$	$R=22.13$	$R=23.87$	$R=26.15$
	$P=28.14$	$P=23.42$	$P=21.80$	$P=17.28$	$P=15.23$
	$F=19.79$	$F=21.01$	$F=21.96$	$F=20.05$	$F=19.25$

观察表 2 和表 3 可知，在 user-based CF 中，当最近邻 $K=10$ 时推荐效果更佳；在 item-based CF 中，当最近邻 $K=15$ 时推荐效果更佳.

2）VSM + *K*-means 算法与本文方法最优评估指标确定

VSM + *K*-means 算法与本文方法都使用了聚类算法，新闻集的聚类簇数 *M* 对最后算法推荐结果的评估指标会产生影响，笔者考虑的聚类簇数 *M* 为10、15、20、25、30 共5种情况，对应的评估指标结果如表4、表5所示.

表4 VSM + *K*-means 算法在不同 *M* 参数下的性能

%

M	*N* = 10	*N* = 15	*N* = 20	*N* = 25	*N* = 30
	R = 17.25	*R* = 22.06	*R* = 24.74	*R* = 26.31	*R* = 27.86
10	*P* = 30.24	*P* = 26.32	*P* = 23.26	*P* = 20.13	*P* = 18.21
	F = 21.97	*F* = 24.00	*F* = 23.98	*F* = 22.81	*F* = 22.02
	R = 19.31	*R* = 23.54	*R* = 25.68	*R* = 28.03	*R* = 30.12
15	*P* = 31.49	*P* = 29.51	*P* = 27.21	*P* = 25.24	*P* = 22.13
	F = 23.94	*F* = 26.19	*F* = 26.42	*F* = 26.56	*F* = 25.51
	R = 20.43	*R* = 24.21	*R* = 26.87	*R* = 29.56	*R* = 32.34
20	*P* = 33.30	*P* = 31.56	*P* = 28.31	*P* = 26.52	*P* = 24.08
	F = 25.32	*F* = 27.40	*F* = 27.57	*F* = 27.96	*F* = 27.61
	R = 18.54	*R* = 22.61	*R* = 25.43	*R* = 27.94	*R* = 29.86
25	*P* = 31.12	*P* = 30.01	*P* = 27.63	*P* = 26.06	*P* = 22.17
	F = 23.24	*F* = 25.79	*F* = 26.48	*F* = 26.97	*F* = 25.45
	R = 17.26	*R* = 21.30	*R* = 23.56	*R* = 25.67	*R* = 26.51
30	*P* = 30.58	*P* = 26.84	*P* = 23.51	*P* = 19.62	*P* = 17.85
	F = 22.07	*F* = 23.75	*F* = 23.53	*F* = 22.24	*F* = 21.33

表5 本文方法在不同 *M* 参数下的性能 %

M	*N* = 10	*N* = 15	*N* = 20	*N* = 25	*N* = 30
	R = 19.24	*R* = 24.56	*R* = 26.43	*R* = 28.31	*R* = 30.64
10	*P* = 32.51	*P* = 28.55	*P* = 26.31	*P* = 23.54	*P* = 20.87
	F = 24.17	*F* = 26.40	*F* = 26.37	*F* = 25.71	*F* = 24.83
	R = 21.35	*R* = 25.46	*R* = 27.64	*R* = 29.89	*R* = 31.26
15	*P* = 33.61	*P* = 31.59	*P* = 28.66	*P* = 26.85	*P* = 23.63
	F = 26.11	*F* = 28.20	*F* = 28.14	*F* = 28.29	*F* = 26.91
	R = 23.36	*R* = 26.68	*R* = 28.74	*R* = 31.02	*R* = 33.84
20	*P* = 35.69	*P* = 33.51	*P* = 31.69	*P* = 28.45	*P* = 26.53
	F = 28.24	*F* = 29.71	*F* = 30.14	*F* = 29.68	*F* = 29.74
	R = 21.46	*R* = 25.43	*R* = 26.96	*R* = 29.06	*R* = 30.97
25	*P* = 33.42	*P* = 31.06	*P* = 28.56	*P* = 26.09	*P* = 22.92
	F = 26.14	*F* = 27.96	*F* = 27.74	*F* = 27.50	*F* = 26.34
	R = 18.61	*R* = 24.05	*R* = 25.98	*R* = 27.46	*R* = 29.91
30	*P* = 31.62	*P* = 27.03	*P* = 25.53	*P* = 22.79	*P* = 19.86
	F = 23.43	*F* = 25.45	*F* = 25.75	*F* = 24.91	*F* = 23.87

观察表4、表5可知，当聚类簇数 *M* = 20 时，2种方法的推荐效果更优于其他聚类簇数.

3）不同算法下的评估指标比较

在图3中，4种算法的准确率都呈现由高至低的变化趋势，是由推荐结果中用户喜欢的新闻增长率低于新闻推荐结果数量增长率导致的. 图4中4种算法的召回率都呈现与图3相反的趋势，是因为随着推荐结果中新闻数量的增加，其包含的用户喜欢新闻数量也增加，根据召回率的计算规则，可知召回率将呈现增长的趋势. 准确率和召回率的排序由高到低依次为本文方法、VSM + *K*-means算法、user-based CF 和 item-based CF. 因为实验数据中新闻数量比用户数量多数倍，故可能导致 user-based CF 推荐效果要比 item-based CF 稍好一点；但可知 VSM + *K*-means算法明显优于前2种协同过滤算法；同时本文方法的准确率和召回率都要优于 VSM + *K*-means 算法，这应该是本文算法采用的 Bisecting *K*-means聚类算法聚类效果优于 *K*-means 算法导致的，实验结果也符合上文中的原理分析. 数据方面，在准确率上，本文方法要比 user-based CF 和 item-based CF 2种协同过滤算法平均优于6%，比 VSM + *K*-means 算法平均优于2%；在召回率上，本文方法要比 user-based CF 和 item-based CF 2种协同过滤算法平均优于7%，比 VSM + *K*-means 算法平均优于2%.

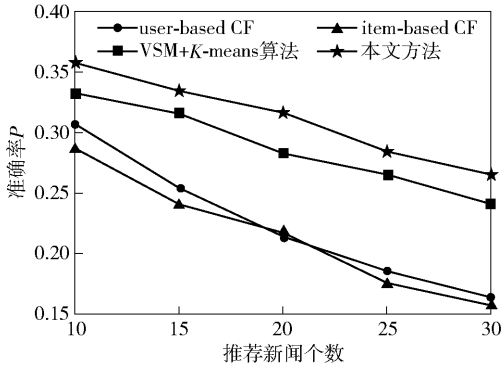

图3 不同算法下的准确率

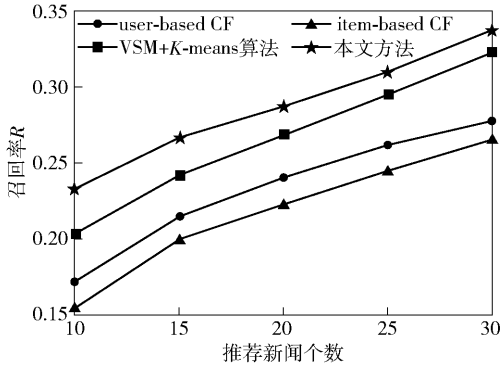

图4 不同算法下的召回率

图5展示了这4种算法在 *F* 值上的变化趋势，

F 值都表现为先高后低，*F* 值是综合准确率和召回率的一种评价指标，其变化趋势与算法中准确率和召回率的变化速率有关，具体内涵还需深入思考. 数据方面，从图中可以看出，当推荐结果个数在15 ~ 25 之间时推荐效果更好，在此期间本文方法要比 user-based CF 和 item-based CF 2 种协同过滤算法平均优于 7.8%，比 VSM + *K*-means 算法平均优于 2.2%.

从准确率、召回率和 *F* 值这 3 种评估指标分析，可见笔者提出的方法要比传统的 user-based CF 和 item-based CF 2 种协同过滤算法在推荐效果上有明显提高；与 VSM + *K*-means 算法相比略有改善.

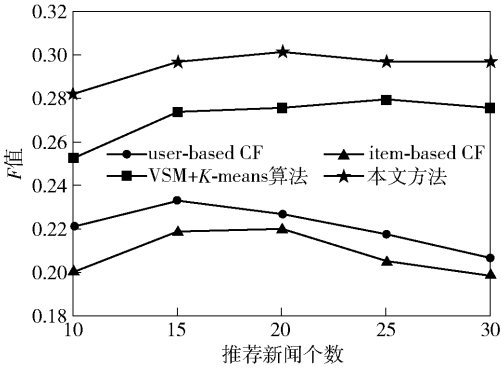

图 5　不同算法下的 *F* 值比较

5　结束语

个性化新闻推荐技术有了一定的研究基础，笔者在之前研究的基础上，提出了一种融合 VSM 和 Bisecting *K*-means 聚类算法的新闻推荐方法，使用 VSM 能较好地表征新闻文本特征，根据 Bisecting *K*-means 聚类算法对新闻聚类避免产生人为对新闻分类的主观缺陷. 实验结果表明，所提出的方法在准确率、召回率和 *F* 值等方面与传统的基于用户的协同过滤算法、基于物品的协同过滤算法相比有较好的优越性. 与 VSM 和 *K*-means 聚类相结合的推荐方法进行了比较，结果表明，Bisecting *K*-means 聚类算法得出的推荐结果效果更佳. 下一步将对用户兴趣模型更新、结合地理位置信息的新闻推荐方法进行研究，进一步提高推荐性能，增强用户满意度.

参考文献：

[1] 冷亚军，陆青，梁昌勇. 协同过滤推荐技术综述[J]. 模式识别与人工智能，2014，27(8)：720-734.
Leng Yajun，Lu Qing，Liang Changyong. Survey of recommendation based on collaborative filtering[J]. PR and AI，2014，27(8)：720-734.

[2] Das A S，Datar M，Garg A，et al. Google news personalization：scalable online collaborative filtering [C] // International Conference on World Wide Web. [S. l.]：ACM，2007：271-280.

[3] Garcin F，Zhou K，Faltings B，et al. Personalized news recommendation based on collaborative filtering [C] // IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. [S. l.]：IEEE，2013：437-441.

[4] Wu X，Xie F，Wu G，et al. Personalized news filtering and summarization on the web[C] // IEEE International Conference on TOOLS with Artificial Intelligence. [S. l.]：IEEE，2011：68-76.

[5] 曹一鸣. 基于协同过滤的个性化新闻推荐系统的研究与实现[D]. 北京：北京邮电大学，2013.

[6] 周由，戴牡红. 语义分析与 TF-IDF 方法相结合的新闻推荐技术[J]. 计算机科学，2013，40(S2)：267-269，300.
Zhou You，Dai Muhong. News recommendation technology combining semantic analysis with TF-IDF method[J]. Computer Science，2013，40(S2)：267-269，300.

[7] 郝水龙，吴共庆，胡学钢. 基于层次向量空间模型的用户兴趣表示及更新[J]. 南京大学学报(自然科学)，2012，48(2)：190-197.
Hao Shuilong，Wu Gongqing，Hu Xuegang. Presentation and updation for user profile based on hierarchical vector space model[J]. Journal of Nanjing University (Natural Sciences)，2012，48(2)：190-197.

[8] 古万荣，董守斌，何锦潮，等. 基于二次聚类的新闻推荐方法[J]. 华南理工大学学报 (自然科学版)，2014(7)：15-20.
Gu Wanrong，Dong Shoubin，He Jingchao，et al. News recommendation method based on secondary clustering [J]. Journal of South China University of Technology (Natural Science Edition)，2014(7)：15-20.

[9] 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，2013：20-29.

[10] 刘啸剑. 基于主题模型的关键词抽取算法研究[D]. 合肥：合肥工业大学，2016.

[11] Abuaiadah D. Using bisect *k*-means clustering technique in the analysis of Arabic documents[J]. ACM Transactions on Asian and Low-Resource Language Information Processing，2016，15(3)：1-13.