文章编号：2095-2163(2020)01-0061-06 中图分类号：TP391.1 文献标志码：A

基于多尺度的 n-grams 特征选择加权及匹配算法

刘世兴
（辽宁机电职业技术学院 信息工程系，辽宁 丹东 118009）

摘 要：n-grams 语言模型作为文本分类中常用的特征，具有结构简单、易筛选、携带语义量大以及对分类贡献值高等优点。但由于其固有的结构特点，在使用普通的选择加权及匹配算法时会造成权值区分不明显，并产生大量稀疏数据，使得建立的分类模型不准确，进而导致最终分类结果的偏差。为解决上述问题，根据词性、语义及词汇的内在偏序关系，提出一种结合词汇、词性和语义的特征选择加权及匹配算法，使 n-grams 特征权值区分明显的同时避免在训练和测试过程中产生大量稀疏数据。在美国当代英语语料库和北京 BBC 汉语语料库中的实验结果表明，与传统的 n-grams 特征选择加权及匹配算法相比，基于多尺度的 n-grams 特征选择加权及匹配算法中得到的 n-grams 特征权值区分明显且稀疏数据大幅减少，在支持向量机（Support Vector Machine，SVM）中的分类效果更好。

关键词：n-grams；特征选择；特征加权；偏序集；词性；语义近似度

multi-scale-based n-grams feature selection weighting and matching algorithm

LIU Shixing
（Department of Information Engineering, Liaoning Mechatronics College, Dandong Liaoning 118009, China）

【Abstract】 As a common feature type in text categorization, n-grams language model has the advantages of simple structure, easy screening, large carrying semantics and high contribution to classification. However, due to its inherent structural characteristics, when using the common selection weighting and matching algorithm, the weight distinction is not obvious, and a large amount of sparse data is generated, which results in an inaccurate classification model, and leads to deviation of the final classification result. In order to solve the above problems, according to the intrinsic partial order relationship between part of speech, semantics and vocabulary, a feature selection weighting and matching algorithm is proposed. It avoids the large amount of sparse data in the n-grams feature during training and testing.The experiment results in the American Contemporary English Corpus and the Beijing BBC Chinese Corpus show that compared with traditional n-grams feature selection weighting and matching algorithm, the sparse data is significantly reduced in n-grams feature selection weighting and matching algorithm based on multi-scale, and it has better classification results in Support Vector Machine(SVM).

【Key words】 n-grams; feature selection; feature weighting; partial order set; part of speech; semantic similarity

0 引 言

n-grams 语言模型的基本思想是将文本按照词汇大小为 n 的滑动窗口进行操作，形成长度为 n 的词汇片段序列，具有拼写容错能力强、语种无关性[1-2]、不需词典规则[3]及特征维度低[4]等优点，因而广泛应用于大数据及人工智能领域，如：文本分类[5-6]、聚类[7]、拼音校验[8]、预测[9]、机器翻译[10]、语音识别[11]、情感分析[12]及恶意软件检测[13]等。

作为 n-grams 语言模型的应用前提，其选择、加权及匹配算法一直是研究人员关注的重点。Maipradit 等人[14]提出一种基于逆文本频率的 n-grams 特征选择算法，用于解决情感分类问题。该算法仅注重特征的选择，没有加权过程，分类性能提升幅度较小。Zhou 等人[15]将 n-grams 模型进行重构，以解决特征选择及加权问题。但重构后的 n-grams 特征语义结构遭到破坏，减少了原有特征包含的文本信息。Hwang 等人[16]将 n-grams 特征的长度扩大到 5，用于长文本及海量文本的分类。但该方法只能用于特定文本，在中、短文本中，5-grams 特征会产生更多的稀疏数据。

为优化 n-grams 特征的选择、加权及匹配算法，在不破坏 n-grams 结构的前提下，增强 n-grams 特征的权值区分度，减少稀疏数据的产生，本文以词性、语义及词汇的内在偏序关系为基础，提出利用词性的类分布和语义近似度结合传统 n-grams 特征加权，达到区分特征权值的目的，通过权值过滤并选择 n-grams 特征，在匹配过程中引入语义近似度，减少匹配过程中产生的稀疏数据。在美国当代英语语料

基金项目：校企协作基于现代学徒制的校内实训基地的研究与建设(JYLX2019024)。

作者简介：刘世兴(1990-)，男，硕士，助教，主要研究方向：大数据、模式识别、人工智能。

收稿日期：2019-11-10

库和北京 BBC 汉语语料库中的实验结果表明,相比于引言中提到的 3 种算法,基于多尺度的 n-grams 特征选择加权及匹配算法得到的特征质量更高,并在分类器中的准确率大幅提升。

1 多尺度 n-grams 特征选择加权

传统特征加权一般建立在文本中特征出现的频率或概率的基础上,如 tf-idf、卡方检验等。这类加权方法对于单独的词汇特征较为有效,但 n-grams 特征由多个词汇组成,假设 n 个词汇在文本中的出现概率为:$p_1, p_2, \cdots\cdots, p_n$,则这些词汇同时出现在同一文本的概率为:

$$P = p_1 \cdot p_2 \cdot \cdots\cdots \cdot p_n. \quad (1)$$

由于每个词汇的出现概率 $p_k(1 \leqslant k \leqslant n)$ 值域为(0,1),故概率 P 会随着 n 的增加趋近于 0,此时 P 为 n 个特征同时出现在同一文本中,并不考虑每个词汇的顺序和是否连续的情况,而 n-grams 特征是 n 个词汇连续出现且顺序不变,所以真正的 n-grams 特征出现概率要远远小于 P 甚至为 0,而建立在此基础上的传统加权方法在训练集中会导致 2 个问题:

(1)特征权值都为趋于 0 的值,区分度不够,难以选择;

(2)在测试集中匹配时产生大量的 0 数据,即稀疏数据。

1.1 n-grams 词性分布加权

性质 1 在文本中,n-grams 特征在词性、语义及词汇三种状态下包含的文本信息量存在如下的偏序关系:

$$\{\text{词性} \prec \text{语义} \prec \text{词汇}\}$$

证明 以“我爱香蕉”为例,首先,经过分词后会被切分为“我”、“爱”和“香蕉”,属于 3-grams 特征。当该特征以词性形式出现时,形式为“代词-动词-名词”,仅从该特征的词性结构来看,很难得到有用的文本信息。当该特征以近似的语义词出现时,以“我喜欢水果”为例,该 3-grams 特征与原特征近似度约为 0.8,从中可以获得一定量的文本信息,但具体喜欢什么水果难以从中得出结论,还需通过原特征中的具体词汇做出准确判断,故词性、语义及词汇存在偏序关系,性质得证。

为增强 n-grams 特征权值区分度并减少稀疏数据,充分利用性质 1 中的偏序关系,提出多尺度加权的第一个尺度,即词性分布加权。虽然词性包含的文本分类信息非常有限,但当其作为 n-grams 特征的词性组合时,提供的文本分类信息要多于其单独存在的情况。例如在新闻分类中,体育类新闻经常出现某名运动员或球队的名字后面接诸如:射门、投篮、运球、击打、获胜、失利、退赛等词汇,组成 2-grams 特征。而转换为词性是“名词-动词”的形式,相比较于财经类新闻和科技类新闻中常见的“某公司股价或财报”、“某省房地产”、“某国央行”及“某形容词+某品牌手机”这类“名词-名词”或“形容词-名词”的形式,可见“名词-动词”的 n-grams 词性组合在体育类新闻的分布更为广泛。综上所述,使用标准差对词性分布进行衡量,如式(2)所示:

$$Weight(POS) = \sqrt{\frac{1}{n}\sum_{1}^{n}(freq_i - freq_{avg})^2}. \quad (2)$$

其中,n 为待分类文本类别数;$freq_i$ 为该词性组合在第 i 个类中的出现频率;$freq_{avg}$ 为该词性组合在 n 个类别中出现的平均频率。由此给出尺度一,词性分布加权算法,详述如下。

算法 1 词性分布加权算法

输入:待分类文本集合 $T\{t_1, t_2, \ldots\ldots, t_m\}$,待分类文本类别数 n

输出:n-grams 特征词性组合权值集合 $W_{pos}\{pos_1, pos_2, \ldots\ldots, pos_k\}$

begin

使用 ICTCLAS 分词接口对文本集合 T 进行分词

使用滑动窗口法得到 n-grams 特征集合,进行词性化处理得到 W_{pos}

for each i: 1 to k

$$Weight(pos_i) = \sqrt{\frac{1}{n}\sum_{1}^{n}(freq_i - freq_{avg})^2}$$

$$W_{pos}[i] = Weight(pos_i)$$

end for

end

1.2 n-grams 语义加权及匹配

语义作为性质 1 中的偏序集元素,可以作为另一个尺度引入 n-grams 特征加权或匹配。在进行语义加权及匹配时遵循如下原则,即相同长度的 n-grams 特征进行语义加权或匹配。匹配过程中首先得到待加权的 n-grams 特征,如下所示:

$$word_1 - word_2 - \cdots\cdots - word_n$$

从文本头部开始,以长度 n 为窗口,得到与待加权特征长度相同的文本段,如下所示:

$$word_1' - word_2' - \cdots\cdots - word_n'$$

进行语义加权时,使用 ICTCLAS 语义计算接

口，将 $word_1$ 与文本段中每个词汇做语义近似度计算，取语义近似最大值 $sim_{\max}(word_1)$。当出现如下两种情况时将 $sim_{\max}(word_1)$ 值赋 0：

(1) $sim_{\max}(word_1)$ 小于 0.5，表明 n - grams 特征中第一个词汇与文本段不具备语义近似；

(2) $word_1$ 为语义无关词集合中的词，如“的”、“了”、“地”等。

除上述两种情况外，将 $sim_{\max}(word_1)$ 保留，继续计算 n - grams 特征中第二个词汇 $word_2$ 的语义近似值 $sim_{\max}(word_2)$。n-grams 特征与第一个文本段的语义近似度计算如式(3)所示：

$$Similarity(n-grams)=\frac{\sum_{i=1}^{n} sim_{\max}(word_i)}{n}. \quad (3)$$

其中，n 为 n-grams 特征中的词汇个数，将特征中每个词的最大语义近似值求和后平均到每个词汇中，得到与文本段的最终语义近似结果。若文本 t 含有 m 个词汇，即长度为 m，则其在与长度为 n 的 n-grams特征加权时，会分为 $m-n+1$ 个文本段，则在该文本中得到的加权值即为 $m-n+1$ 个文本段权值和。若训练集或测试集 T 中存在 k 个文本，则 n-grams 特征在 k 个文本中重复上述操作，得到的结果 $Similarity(n-grams,t)$ 既可以作为训练集的特征加权，用于后续特征选择，也可以作为测试集的匹配结果，避免稀疏数据。综上所述给出尺度二，即语义加权及匹配算法，详述如下。

算法 2　语义加权及匹配算法

输入：训练集或测试集文本 $T\{t_1,t_2,\ldots\ldots,t_m\}$

输出：语义加权或匹配后的 n-grams 特征值集合 $W_{sim}\{w_1,w_2,\ldots\ldots,w_k\}$

```
begin
    对文本 T 使用 ICTCLAS 分词接口进行分词
    使用滑动窗口法得到 n-grams 特征集合 F
  for each i: 1 to F.size
      for each j: 1 to T.size
          W_sim(F[i]) = W_sim(F[i]) + Similarly(F[i],T[j])
      end for
  end for
end
```

1.3　多尺度的 n-grams 特征选择加权及匹配算法

在给出前两种尺度的 n-grams 特征加权方法后，结合传统的 tf-idf 加权特征值，将 3 种尺度得到的权值求和，若结果小于阈值 β，则舍弃该 n-grams 特征，否则保留。综上所述，给出基于多尺度的 n-grams 特征选择加权算法，详述如下。

算法 3　多尺度的 n-grams 特征选择加权算法

输入：训练集或测试集文本 $T\{t_1,t_2,\ldots\ldots,t_m\}$

输出：语义加权或匹配后的 n-grams 特征值集合 $W\{w_1,w_2,\ldots\ldots,w_k\}$

```
begin
    对文本 T 使用 ICTCLAS 分词接口进行分词
    使用滑动窗口法得到 n-grams 特征集合 F
    根据词性分布加权，得到 n-grams 词性组合权值 W_pos{pos_1,pos_2,.....,pos_k}
    根据语义加权，得到 n-grams 权值集合 W_sim{sim_1,sim_2,.....,sim_k}
    根据传统 tf-idf 加权方法得到权值集合 W_tfidf{td_1,td_2,.....td_k}
    for each i: 1 to k
        W[i] = W_pos[i] + W_sim[i] + W_tfidf[i]
        if W[i] < β
          delete W[i]
        end if
    end for
end
```

2　实验

本文研究采用的实验数据集分中英文两种，即北京 BBC 汉语语料库和美国当代英语语料库。使用支持向量机(Support Vector Machine，SVM)进行分类性能评价，评价指标包括准确率(P)、召回率(R)和 F 值(F)。

2.1　北京 BBC 汉语语料库实验

北京 BBC 汉语语料库实验中，n-grams 文本分类特征采用 5-grams。在特征数分别为 200、500、1 000、1 500、2 000、2 500 和 3 000 时，对比引言中提到的 3 种 n-grams 特征选择加权方法和本文方法的分类准确率、召回率及 F 值，如图 1~图 3 所示。从图 1~图 3 中可以看到，4 种方法在 5-grams 特征数不超过 1 000 时，准确率、召回率和 F 值都呈上升趋势。当 5-grams 特征数量超过 1 000 时，逆文本频率法和重构法的分类性能呈现断崖式下降，究其原因在于当 5-grams 特征数量增多时，产生的稀疏数据也在增加，且增加速率相比于非稀疏数据要快，导致在特征数量达到3 000时，二者的准确率不足 70%。扩展的5-grams方法由于其针对 5-grams 特征设计，在数量超过1 500后分类性能逐渐下降，但

相比于逆文本频率和重构法,性能下降幅度较小,在5-grams 特征数达到 3 000 时准确率为 82%。而本文方法由于从词性、语义和词汇三个尺度对 n-grams 特征进行选择加权及匹配,使得稀疏数据大幅减少,从图 3 中可以看到,特征数量超过2 000时分类性能才开始下降,在特征数达到 3 000 时准确率达到 89%。

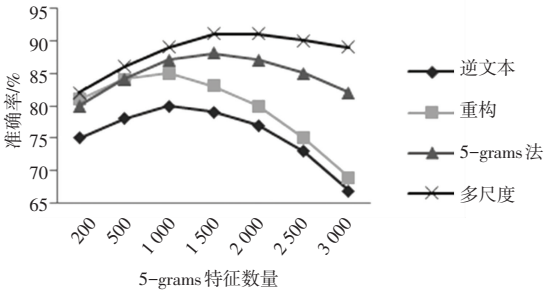

图 1 3 种引言方法与本文方法在 SVM 中的准确率

Fig. 1 Accuracy of three introduction methods and the proposed methods in SVM

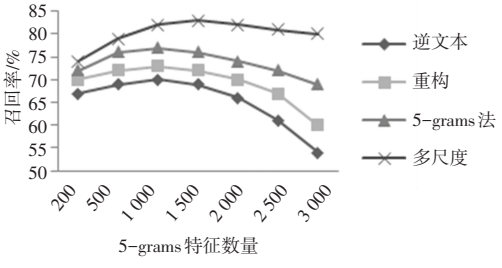

图 2 3 种引言方法与本文方法在 SVM 中的召回率

Fig. 2 The recall rate of three introduction methods and the proposed method in SVM

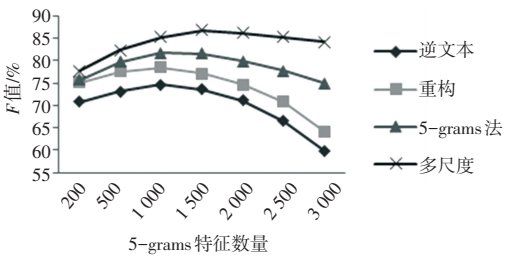

图 3 3 种引言方法与本文方法在 SVM 中的 *F* 值

Fig. 3 The *F* value of three introduction methods and the proposed method in SVM

2.2 美国当代英语语料库

在美国当代英语语料库中,分别对比引言中的 3 种方法和本文方法在 2-、3-、4-及 5-grams 特征时产生的稀疏数据数量和占比,见表 1~表 4。从表 1~表 4 中可以明显看出,当特征为 2-grams 且特征数量不超过 2 000 时,引言中的 2 种方法稀疏数据并不严重,占比不超过 20%,当特征数量达到 3 000 时,稀疏数据逐渐增加,分别达到 32.7%和 28.4%。而本文方法在稀疏数据的抑制上效果较好,在 2-grams 特征数量达到 3 000 时,占比仅为 11.9%。当 n-grams 特征中 n 值分别为 3 时,逆文本频率法在特征数为 200 时,稀疏数据的占比就达到了 20%以上,重构法在特征数为 500 时稀疏数据占比超过 20%,而当特征数量达到 3 000 时,2 种引言方法的稀疏数据占比接近 50%,已经不适合作为分类特征。相比而言,本文方法在特征数达到 3 000 时,稀疏数据占比仍然保持在 20%以下。当 n 值达到 4 和 5 时,引言方法在特征数量超过 1 000 后都产生了占比接近于 50%的稀疏数据,无法作为分类特征进行分类建模。而本文方法在特征数达到 3 000 的情况下,4-grams 和 5-grams 时的稀疏数据占比分别为 23.3%和 29.4%,远低于引言方法。

表 1 3 种方法在 2-grams 特征时的稀疏数据量及所占比例

Tab. 1 The amount of sparse data and the proportion of the three methods in the 2-grams feature

特征数量	逆文本		重构		多尺度	
	稀疏数据	占比/%	稀疏数据	占比/%	稀疏数据	占比/%
200	36	18.0	30	15.0	7	3.5
500	95	19.0	84	17.0	24	4.8
1 000	186	18.6	179	18.1	61	6.1
1 500	301	20.0	291	19.5	114	7.6
2 000	425	21.2	418	21.0	182	9.1
2 500	690	27.6	621	24.9	267	10.7
3 000	981	32.7	852	28.4	357	11.9

表 2　3 种方法在 3-grams 特征时的稀疏数据量及所占比例

Tab. 2　The amount of sparse data and the proportion of the three methods in the 3-grams feature

特征数量	逆文本		重构		多尺度	
	稀疏数据	占比/%	稀疏数据	占比/%	稀疏数据	占比/%
200	43	21.5	30	15.0	11	5.5
500	120	24.1	101	20.2	35	7.0
1 000	269	26.9	238	23.8	92	9.2
1 500	445	29.7	409	27.3	166	11.1
2 000	676	33.8	602	30.1	268	13.4
2 500	977	39.1	867	34.7	400	16.0
3 000	1 416	47.2	1 242	41.4	567	18.9

表 3　3 种方法在 4-grams 特征时的稀疏数据量及所占比例

Tab. 3　The amount of sparse data and the proportion of the three methods in the 4-grams feature

特征数量	逆文本		重构		多尺度	
	稀疏数据	占比/%	稀疏数据	占比/%	稀疏数据	占比/%
200	62	31.5	57	28.5	21	10.5
500	180	36.0	167	33.4	52	10.4
1 000	419	41.9	381	38.1	139	13.9
1 500	708	47.2	666	44.4	243	16.2
2 000	1 102	55.1	1 006	50.3	362	18.1
2 500	1 535	61.4	1 445	57.8	510	20.4
3 000	2 064	68.8	1 941	64.7	699	23.3

表 4　4 种方法在 5-grams 特征时的稀疏数据量及所占比例

Tab. 4　The amount of sparse data and the proportion of the four methods in the 5-grams feature

特征数量	逆文本		重构		扩展 5-grams		多尺度	
	稀疏数据	占比/%	稀疏数据	占比/%	稀疏数据	占比/%	稀疏数据	占比/%
200	96	48.0	91	45.5	48	24.0	45	22.5
500	276	55.2	268	53.6	136	27.2	121	24.2
1 000	667	66.7	644	64.4	311	31.1	263	26.3
1 500	1 071	71.4	1 063	70.9	553	36.9	406	27.1
2 000	1 502	75.1	1 496	74.8	854	42.7	561	28.0
2 500	1 932	77.3	1 927	77.1	1 230	49.2	715	28.6
3 000	2 388	79.6	2 376	79.2	1 623	54.1	882	29.4

3　结束语

针对 n-grams 特征选择加权及匹配过程中的权值区分度低和稀疏数据多的问题，本文利用词性、语义和词汇的内在偏序关系，提出使用词性分布、语义和词汇的多尺度 n-grams 特征选择加权及匹配算法。该算法在有效降低稀疏数据量的同时，能够在特征加权过程中增强权值区分度，便于选择优质特征。对比实验结果表明，本文方法得到的 n-grams 特征在分类性能和稀疏数据控制上均有大幅提升，有利于进一步推广 n-grams 特征的应用领域。

参考文献

[1] GAJENDRAGADKAR U, JOSHI S. Anatomy of building Marathi n-grams [C]//2016 International Conference on Computing, Analytics and Security Trends (CAST). Pune, India: IEEE, 2016:362-365.

[2] SARKAR K. Using character N-gram features and multinomial Naive Bayes for sentiment polarity detection in Bengali Tweets [C]//2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT). Kolkata, India: [s.n.], 2018.

[3] 于津凯，王映雪，陈怀楚.一种基于 N-Gram 改进的文本特征提取算法[J]. 图书情报工作，2004，48(8)：48-50，43.

[4] PENAGARIKANO M, VARONA A, RODRIUEZ-FUENTES L J. Dimensionality reduction for using high-order n-Grams in SVM-based phonotactic language recognition [C]// Proc of the 12th Annual Conference of the International Speech Communication Association.Florence,Italy,2011: 853-856.
[5] ZAKI T,ES-SAADY Y,MAMMASS D,et al. A hybrid method N-Grams-TFIDF with radial basis for indexing and classification of Arabic document[J].International Journal of Software Engineering and Its Applications,2014,8(2): 127-144.
[6] SIDOROV G,VELASQUEZ F,STAMATATOS E,et al. Syntactic dependency-based N-Grams as classification features [M]// BATYRSHIN I, MENDOZA M G. Advances in Computational Intelligence. MICAI 2012. Lecture Notes in Computer Science. Berlin/Heidelberg:Springer,2013,7630:1-11.
[7] KUMAR R, MATHUR R P. Short text clustering using numerical data based on n-gram[C]//2014 5th International Conference-Confluence The Next Generation Information Technology Summit (Confluence). Noida, India:IEEE,2014:274-276.
[8] PRIYA M, KALPANA R, SRISUPRIYA T. Hybrid optimization algorithm using N gram based edit distance [C]//2017 International Conference on Communication and Signal Processing (ICCSP). Chennai, India:IEEE,2017:216-221.
[9] NAGALAVI D, HANUMANTHAPPA M. N-gram Word prediction language models to identify the sequence of article blocks in English e-newspapers [C]//2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS). Bangalore, India: IEEE, 2016.
[10] RAHMAN M M, KABIR M F, HUDA M N. A corpus based N-gram hybrid approach of Bengali to English machine translation [C]//2018 21st International Conference of Computer and Information Technology (ICCIT). Dhaka, Bangladesh:[s.n.], 2018:1-6.
[11] WANG Bin, OU Zhijian, LI Jian. Joint-character-POC N-gram language modeling for Chinese speech recognition[C]//The 9th International Symposium on Chinese Spoken Language Processing. Singapore: Chinese and Oriental Languages Information Processing Society, 2014:1-5.
[12] CABANLIT M A, ESPINOSA K J. Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons[C]// 2014 5th International Conference on Information, Intelligence, Systems and Applications (IISA). Chania, Greece:IEEE,2014:1-4.
[13] MIRA F, HUANG Wei, BROWN A. Improving malware detection time by using RLE and N-gram [C]// 2017 23rd International Conference on Automation and Computing (ICAC). Huddersfield, UK:IEEE,2017:1-5.
[14] MAIPRADIT R, HATA H, MATSUMOTO K. Sentiment classification using N-Gram inverse document frequency and automated machine learning[J]. IEEE Software, 2019,36(5): 65-70.
[15] ZHOU Zhengyu, MENG H. Pseudo-conventional N-Gram representation of the discriminative N-Gram model for LVCSR [J]. IEEE Journal of Selected Topics in Signal Processing, 2010, 4(6):943-952.
[16] HWANG M, YEOM H N, HWANG M N, et al. Construction of scholarly n-Gram from huge text data [C]//2014 Eighth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing. Birmingham, UK:IEEE,2014: 31-35.

(上接第 60 页)

变化影响,得出算法的高效率参数配置。并进行对照实验,比较传统推荐算法与 KMMD 算法在 MovieLens 公开数据集上的实验效果,得出结论,融合 K-means 聚类后的矩阵分解算法确实有助于推荐准确度的提高,且在引入用户属性数据的条件下,有效改善了用户冷启动问题。但该算法对项目冷启动问题还难以处理。后续工作是将深度学习技术与推荐算法相融合,构建多种推荐算法的组合排序,力求进一步提升推荐算法的准确度,并有效解决项目冷启动问题。

参考文献

[1] 陈彬,张荣梅. 智能推荐系统研究综述[J]. 河北省科学院学报,2018,35(3):82-92.
[2] GOLDBERG D,NICOLS D, OKI B M, et al. Using collaborative filtering to weave an infomation tapestry[J]. Communications of the ACM,1992,35(12):61-70.
[3] 张清,于博,王辉,等. 一种有效缓解数据稀疏问题的协同过滤推荐算法[J]. 合肥工业大学学报(自然科学版),2019,42(4): 473-478.
[4] 刘璐. 推荐算法中冷启动问题的研究与实现[D]. 北京:北京邮电大学,2019.
[5] 于洪,李俊华. 一种解决新项目冷启动问题的推荐算法[J]. 软件学报,2015,26(6):1395-1408.
[6] 万家山,陈蕾,吴锦华,等. 基于 KD-Tree 聚类的社交用户画像建模[J]. 计算机科学,2019,46(S1):442-445,467.
[7] 王建芳,苗艳玲,韩鹏飞,等. 一种基于信任机制的概率矩阵分解协同过滤推荐算法[J]. 小型微型计算机系统,2019,40(1): 31-35.
[8] 张建军,陆国生,刘征宇. 基于奇异值分解和项目属性的推荐算法[J]. 合肥工业大学学报(自然科学版),2018,41(6):761-765,858.
[9] 李艳娟, 牛梦婷, 李林辉. 基于蜂群 K-means 聚类模型的协同过滤推荐算法[J]. 计算机工程与科学,2019,41(6):1101-1109.
[10] QIAO Yuchen, YANG Xu, WU Enhong. The research of BP Neural Network based on one-hot encoding and principle component analysis in determining the therapeutic effect of diabetes mellitus[J]. IOP Conference Series: Earth and Environmental Science,2019,267(4).
[11] HAN Jiawei, KAMBER M, PEI Jian, 等. 数据挖掘: 概念与技术[M]. 范明, 孟小峰, 译. 3 版. 北京:机械工业出版社, 2012.
[12] 许冰晗, 尚鸿运, 马灿, 等. 基于 Movielens 电影数据的可视分析[J]. 计算机工程与科学,2017,39(11):2086-2094.
[13] DÜNTSCH I, GEDIGA G. Confusion matrices and rough set data analysis[J]. arXiv preprint arXiv:1902.01487v1,2019.