

基于内容过滤 PageRank 的 Top-k 学习资源匹配推荐

梁婷婷¹, 李春青², 李海生²

(1. 广西师范大学 计算机科学与信息工程学院, 广西 桂林 541004;
2. 广西民族师范学院 数学与计算机科学系, 广西 崇左 532200)

摘 要: 针对在线教育支持技术中关于文本处理的多义词和同义词问题, 提出基于内容过滤 PageRank 语义相似替换的 Top-k 学习资源推荐算法。基于内容的向量空间滤波建立学习资源过滤推荐模型, 该模型采用资源间匹配方式以取代语义相似性, 从而避免多义词或同义词的漏检问题。基于谷歌 PageRank 算法结合前述资源间匹配模型构建考虑资源间关系连接的权重矩阵, 取代传统 PageRank 算法网页间的超链接方式, 进行资源类型划分, 得到特征的马尔可夫收敛矩阵, 并利用 Top-k 算法实现推荐结果细化。实验结果表明, 在公共学习资源数据集中, 所提算法对计算时间的覆盖率是可行的。

关键词: 内容过滤; PageRank 算法; Top-k 排序; 马尔可夫收敛矩阵; 资源匹配

中文引用格式: 梁婷婷, 李春青, 李海生. 基于内容过滤 PageRank 的 Top-k 学习资源匹配推荐[J]. 计算机工程, 2017, 43(2): 220-226.

英文引用格式: Liang Tingting, Li Chunqing, Li Haisheng. Top-k Learning Resource Matching Recommendation Based on Content Filtering PageRank[J]. Computer Engineering, 2017, 43(2): 220-226.

Top-k Learning Resource Matching Recommendation Based on Content Filtering PageRank

LIANG Tingting¹, LI Chunqing², LI Haisheng²

(1. School of Computer Science and Information Engineering, Guangxi Normal University, Guilin, Guangxi 541004, China;
2. Department of Mathematics and Computer Science, Guangxi Normal University for Nationalities, Chongzuo, Guangxi 532200, China)

[Abstract] Aiming at the problem of the polysemous words and synonyms of text processing in the online education support technology, a Top-k learning resource recommendation algorithm based on content filtering PageRank is proposed. A learning resource filtering recommendation model is constructed based on content vector space filtering. The model pays attention to resource matching mode to replace the semantic similarity, so as to avoid missing detection of polysemous words or synonyms. Google PageRank algorithm is combined with the aforementioned resource matching model to construct weight matrix considering the relationship between resources. This is used to replace the hyperlink mode between Web pages of the traditional PageRank algorithm for resource type dividing. The Markov convergence matrix of characteristics is constructed, and the Top-k algorithm is used to refine the recommended results. Experimental results show that the proposed algorithm is feasible for the computation time cover rate in the public learning resource dataset.

[Key words] content filtering; PageRank algorithm; Top-k sorting; Markov convergent matrix; resource matching
DOI: 10.3969/j.issn.1000-3428.2017.02.037

0 概述

近年来, 随着在线教育支持技术的发展, 大量数字化学习资源库已经建立, 例如 OER Commons^[1], MERLOT^[2] 以及 LRE^[3], 其提供开放学习资源学科

(如艺术、人文、科技等), 层次(如小学、中学、高中、高等教育等)和类型(如实验室、课堂讲稿、习题、辅导等)。这些资源允许用户在不同领域对其知识进行自学或巩固。然而, 这些资源的多样性很容易阻碍用户深入学习。例如, 每当用户想要学习某个主

基金项目: 广西高校科学技术研究项目(YB2014417)。

作者简介: 梁婷婷(1983—), 女, 讲师、硕士, 主研方向为信息检索; 李春青、李海生, 讲师、硕士。

收稿日期: 2016-01-11 **修回日期:** 2016-02-28 **E-mail:** lntngtng@tom.com

题时,就搜索或浏览与那个主题相关的资源,并使用试误法进行预览,需要花很多时间查找期望资源^[4]。此外,在学习资源之后,如果想找到其他相关资源,需重新执行搜索/浏览的过程^[5]。

为了鼓励在线学习资源的使用,推荐系统被认为是一个关键的解决方案,特别是在技术增强学习(Technology Enhanced Learning, TEL)方面^[6]。当前已有许多应用推荐方法,以支持资源推荐过程,例如通过协同过滤^[7]、内容过滤^[8]、用户等级检查^[9]、关联规则^[10]或用户反馈分析^[11]等方式进行资源推荐。此外,还有贝叶斯模型、马尔可夫链、资源本体论以及混合模型等方式^[12]。然而,大多数系统仍停留在原型设计阶段,只有极少数系统经过人类用户评估。

本文提出基于内容过滤 PageRank 语义相似替换的 Top-k 学习资源推荐算法。所提算法与前述算法不同点在于基于内容的过滤、语义相似性和网页排名。基于内容的过滤主要用于对与用户查询配置文件相似学习资源进行滤波。将基于内容的过滤和语义相似度过滤应用到词汇语义相似度的处理中。网页排名,基于谷歌 PageRank 算法,根据其相互关系进行资源重要性识别。

算法设计过程可概述为:1) 基于内容的向量空间滤波构建学习资源过滤推荐模型,从而避免多义词或同义词的漏检问题;2) 利用谷歌 PageRank 算法,进行资源间匹配,并构建特征的马尔可夫收敛矩阵,然后使用 Top-k 算法实现推荐结果细化。

1 学习资源推荐

在本节中,为活跃的用户提出详细的方法来推荐学习资源。

1.1 基于内容的向量空间滤波

向量空间模型(Vector Space Model, VSM)是信息检索领域普遍使用的文档间相似度的计算方法。通过词频(Term Frequency, TF)和逆文档频率(Inverse Document Frequency, IDF)可对这些权重进行计算。考虑文档 d_i , 可表示为向量^[13]:

$$d_i = \{w_{i1}, w_{i2}, \dots, w_{in}\} \quad (1)$$

其中, n 为文档中的总单词量; w_{ik} ($k = 1, 2, \dots, n$) 为向量中第 k 个元素的权重, w_{ik} 可计算为:

$$w_{i,k} = TF(i, k) \times IDF(k) = \frac{|w_k|}{|d_i|} \times \log_a \frac{n}{|D_k|} \quad (2)$$

其中, $|w_k|$ 是单词 w_k 在文档 d_i 中出现的次数;

$|d_i|$ 为文档中的总单词量; $|D_k|$ 为包含 w_k 的文件的数量。

然后,可通过文档向量的余弦角度对文件之间的相似性进行计算,例如,2 个文件 d_i 和 d_j 之间的相似性可计算为:

$$sim(d_i, d_j) = \cosine(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i| \times |d_j|} \quad (3)$$

在所提方法中,采用向量空间模型计算用户配置文件和资源描述之间的相似性。用户配置文件被定义为最近浏览的资源的一组关键字,而资源描述是所有用于描述资源的文本。

令 $\{k_1, k_2, \dots, k_t\}$ 为用户 h 最近浏览资源 r_a 的 t 个关键词。考虑将这些历史关键字作为一个查询 q_a , 该查询可在同一资源描述空间中表示为向量:

$$q_a = \{w_{a1}, w_{a2}, \dots, w_{an}\} \quad (4)$$

其中, q_a 为在查询中相应的词语权重。

根据式(3),查询 q_a 和文档 d_i 间的相似度计算形式为:

$$sim(q_a, d_i) = \frac{d_i \cdot q_a}{|d_i| \times |q_a|} = \frac{\sum_{k=1}^n w_{ak} w_{ik}}{\sqrt{\sum_{k=1}^n w_{ak}^2} \times \sqrt{\sum_{k=1}^n w_{ik}^2}} \quad (5)$$

对所有的 $d_i \in \{1, 2, \dots, N\}$ 应用式(5)。然后,按照与 q_a 相似性计算值对资源进行降序排序。最后,在相关资源的历史关键词中为活跃用户进行 Top-k 资源选取。

1.2 资源相似匹配查询

多义词和同义词是文本处理面临的共同问题。如果只处理词的句法匹配,没有考虑语义相似性,很容易错过不同词但意义相同的潜在匹配。与以往文献不同,本文主要专注于资源之间的匹配,而不是语义的相似性。

考虑活跃用户 u_a 最近浏览 h 资源的关键词列表为: $q_a = \{k_1, k_2, \dots, k_t\}$, 将该列表作为查询,考虑与查询 q_a 匹配的每个资源 r_i , 建议利用资源描述对每个单词进行替换。根据查询语义与选定的词语相似度计算值,对资源描述的词语权重进行更新。

具体而言,考虑资源描述 d_i 中的单词 v_x , 其出现次数为 o_x 。假定 v_x 为在语义上与查询 q_a 中关键词 k_y 最为相似的单词,其相似度值为 $s(v_x, k_y) \in (0, 1)$ 。用 k_y 替换文档 d_i 中的单词 v_x , 并更新其权重值为 $w_{xy} = o_x s(v_x, k_y)$ 。这意味着,文档 d_i 中单词 v_x 出现 o_x 次,可视为文档 d_i 中单词 k_y 出现 $o_x s(v_x, k_y)$ 次。对文档 d_i 中的所有单词,重复执行上述替换

过程。

考虑 1.1 节所述文档 d_i 和查询 q_a 相似度计算实例,对文档 d_i 利用查询 q_a 中单词进行最相近词语替换,具体如表 1 所示。例如,在文档 d_i 中的单词“domain”与查询 q_a 中的单词“resource”最相似(相似性为 0.015),因此,对文档 d_i 中的单词“domain”用查询 q_a 中单词“resource”进行替换,并更新其权重为 $1 \times 0.015 = 0.015$ 。

表 1 单词替换预处理过程

| v_x | domain | consider | pivot | solution |
|------------|----------|-----------|---------|------------|
| recommend | 0.002 | 0.058 | 0.001 | 0.003 |
| system | 0.007 | 0.000 | 0.003 | 0.015 |
| technology | 0.009 | 0.002 | 0.002 | 0.019 |
| enhance | 0.005 | 0.004 | 0.007 | 0.005 |
| learn | 0.001 | 0.035 | 0.001 | 0.001 |
| resource | 0.016 | 0.003 | 0.003 | 0.014 |
| k_y | resource | recommend | enhance | technology |
| w_{xy} | 0.016 | 0.058 | 0.007 | 0.019 |

如表 1 所示,假定文档 d_i 中共有 n_1 个单词被 k_1 替换,且更新权重为 $\{w_{11}, w_{21}, \dots, w_{n_1 1}\}$,文档 d_i 中共有 n_2 个单词被 k_2 替换,且更新权重为 $\{w_{12}, w_{22}, \dots, w_{n_2 2}\}$,等。文档 d_i 中共有 n_0 个单词没有被查询中的任何单词替换,则资源描述 d_i 变为:

$$d'_i = \{k_1, k_2, \dots, k_t, k_{t+1}, \dots, k_{t+n_0}\} \quad (6)$$

对应权重为:

$$w = \left\{ \sum_{j=1}^{n_1} \omega_{j1}, \sum_{j=2}^{n_2} \omega_{j2}, \dots, \sum_{j=1}^{n_0} \omega_{j0}, 0, \dots, 0 \right\} \quad (7)$$

通过计算 q_a 和 d'_i 间的相似度可获得 l_a 与 d_i 间的相似度。因为在式(7)中, $k_{t+1}, k_{t+2}, \dots, k_{t+n_0}$ 的对应权重为 0,可将这些元素从 d'_i 中予以删除,则 d'_i 的空间向量尺度与 q_a 变为一致。然后,根据式(2)计算文档 d_i 和查询 q_a 中单词的 TF-IDF,应用式(5) VSM 计算其相似度。最后,根据文档 d_i 和查询 q_a 间的相似性进行 Top- k 相似资源选取。

2 语义 PageRank 混合匹配查询

2.1 PageRank 算法

根据资源间相互关系,提出一个资源重要性评估的排序算法。该算法的灵感来自谷歌 PageRank 算法,页面 A 的等级计算形式为^[14-15]:

$$PR(A) = (1 - d) + d \times \sum_{i=1}^n \frac{PR(T_i)}{C(T_i)} \quad (8)$$

其中, $0 \leq d \leq 1$ 是阻尼因子,该因子一般选取为 $d = 0.85$ 。 T_1, T_2, \dots, T_n 为指向页面 A 的页面, $C(T_i)$ 为从页面 T_i 流出的链接数量。页面的初始等级设置

为 $1/N$,其中 N 为页面总数量。根据式(8)对所有页面的等级进行更新,直到其在给定阈值内达到稳定。

但是简单地应用式(8)等级计算形式,对于大数据计算效率不高,为此将网页排名定义为向量 v^* ,满足条件:

$$Gv^* = v^* \quad (9)$$

其中, G 为 Google 矩阵,可定义为:

$$G = \frac{1-d}{N}S + dM \quad (10)$$

其中,在矩阵 G 中,所有矩阵元素等于 1; M 是过渡矩阵。

过渡矩阵 M 表征页面间的链接,其元素 $M_{[i,j]}$ 为从页面 j 到页面 i 的链接权重。元素 $M_{[i,j]}$ 满足:

$$\sum_{i=1}^N M_{[i,j]} = 1, \forall j = 1, 2, \dots, N \quad (11)$$

根据式(3), M 为马尔可夫矩阵,并且如果页面 j 共有 k 个流出链接,则每个链接的权重为 $1/k$ 。根据式(9)可知, v^* 是特征值为 1 的马氏矩阵 M 的特征向量。令 v_0 作为初始页面排序向量,其元素均设置为 $1/N$,则 v^* 可由如下形式迭代计算:

$$v_{i+1} = Gv_i \quad (12)$$

直到终止条件 $|v_{i+1} - v_i| < \varepsilon$ 满足为止,其中 ε 为给定的迭代阈值。

因为 G 为马氏矩阵,则 v_{i+1} 可经过一定迭代次数收敛到 v^* 。 v^* 是根据其超链接的网页等级排序。

2.2 基于资源关系的 PageRank 匹配

受到上述谷歌的 PageRank 算法启发,结合前述语义匹配模型,提出一种计算学习资源的排序算法。与谷歌的 PageRank 算法不同,在所提算法中,考虑到资源间的关系,而不是网页之间的超链接。

基本上,资源可以是其他资源的一部分,包含其他资源或与其他资源相关的 3 种类型,可分别定义为:“is part of”,“contains”以及“associates to”,分别对应 1-1 关系、1- n 关系和 n - n 关系,如图 1 所示。

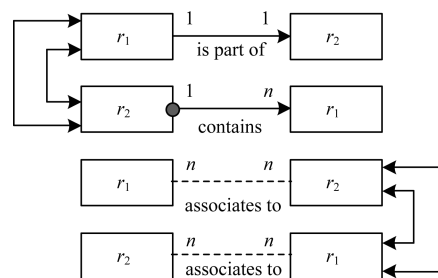


图 1 资源关系描述

这种关系不仅呈现资源的超链接,还揭示了特定语义。例如,“associates to”表示关联资源;

“contains”则列出涉及主题范围内的资源,其中一些可能不关联;“is part of”是另一种资源,其与其他资源的一部分,但不明确提出涉及的相关资源。

根据上述含义,采用为每个关系类型赋予相应关系权重的方式进行处理。具体来说,“associates to”类型的权重要高于“contains”类型,“contains”类型的权重要高于“is part of”类型。令 w_{ra} , w_{rc} 和 w_{rp} 分别对应“associates to”,“contains”以及“is part of”3种类型的权重,并且有 $w_{ra} > w_{rc} > w_{rp}$,为简化,设置如下:

$$w_{ra} = \alpha w_{rc} = \alpha^2 w_{rp}, 0 < \alpha < 1 \quad (13)$$

在谷歌的 PageRank 算法中,所有超链接的权重设置是相等的。而在所提算法中,资源之间的关系根据所属类型设置不同的权重,而不是为所有关系设置平均权重 (PageRank 设置方式,如 1.1 节所述)。具体地说,假设资源 r_i 具有 a 组“associates to”类型, b 组“contains”关系类型以及 c 组“is part of”关系类型,则有:

$$aw_{ra} + bw_{rc} + cw_{rp} = 1 \quad (14)$$

根据式 (13) 和式 (14) 可得:

$$\begin{cases} w_{rp} = \frac{1}{a\alpha^2 + b\alpha + c} \\ w_{rc} = \frac{\alpha}{a\alpha^2 + b\alpha + c} \\ w_{ra} = \frac{\alpha^2}{a\alpha^2 + b\alpha + c} \end{cases} \quad (15)$$

式 (14) 可确保式 (10) 中的矩阵 M 和 G 是马尔可夫矩阵。因此,这些矩阵与初始加权向量的乘积将收敛到一个特征向量。这意味着可以基于新的权重对 PageRank 资源向量进行计算。式 (15) 计算结果通常作为式 (10) 矩阵 M 和 G 的初始化形式。然后,通过式 (9) 对网页排名 v^* 进行计算。

例如,图 2 给出了 4 种资源关系 (r_1, r_2, r_3, r_4) 的语料库。然后,根据式 (15) 可以计算每个关系类型的权重 (w_{ra}, w_{rc}, w_{rp}),并创建关系矩阵 M 。每个矩阵元素 $M_{[i,j]}$ 表述从 r_j 到 r_i 的关系权重,矩阵中每列中的所有元素的总和等于 1。基于矩阵 M ,并通过式 (10) 和式 (12),计算出 G 和 v^* 。

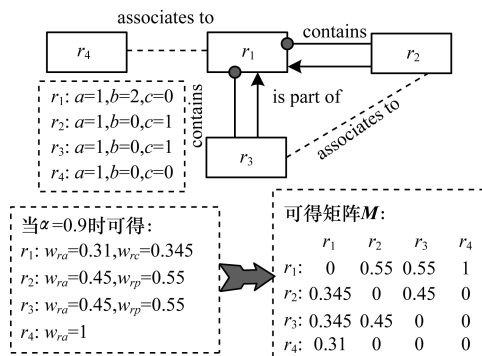


图2 资源关系和相应矩阵实例

PageRank 算法根据资源间的关系定位其重要性。其排名可应用在资源搜索程序中,类似于使用谷歌搜索引擎的页面排名,也可被看作是应用推荐程序中的一个参数,以细化推荐结果。

2.3 混合推荐方法

基于内容的过滤算法和词汇语义相似度,使那些在语法和语义上与用户配置近似的资源得到检索。同时,资源排序有助于识别资源关系的重要性。因此,上述算法的组合可同时用户对用户重要和相关的资源进行检索,有助于提高检索质量。

本文提出基于用户配置文件和资源的排序乘积相似性,以推断其最终的匹配分数。考虑活跃用户 u_a ,其具有配置文件 q_a 和描述 d_i 的资源 r_i 。令 d'_i 为根据查询 q_a 相似度替换重新定义的资源 d_i ,则最终的查询 q_a 和资源 r_i 的匹配度计算如下:

$$scr(q_a, r_i) = sim(q_a, d'_i) \times v^*(i) \quad (16)$$

计算查询 q_a 与所有资源间的匹配度,将这些分数降序排列并选择相应的 Top-k 资源推荐。算法 1 描述了混合算法的伪代码。

算法 1 混合内容过滤和资源关系排序

输入 q_a 用户 u_a 查询配置文件, R 学习资源集

输出 $rec(a)$ 为用户 u_a 的推荐结果

1. $v^* \leftarrow \text{PageRank}(R)$;
2. foreach $r_i \in R$ do
3. $d_i \leftarrow \text{Textdescription}(r_i)$;
4. 基于 q_a 单词相似度替换, $d_i \rightarrow d'_i$;
5. 计算 q_a 的 TF-IDF 向量, q'_a ;
6. 计算 d_i 的 TF-IDF 向量, d'_i ;
7. $sim(q_a, d_i) \leftarrow \text{cosine}(q_a, d'_i)$;
8. $scr(q_a, r_i) \leftarrow sim(q_a, d_i) \times v^*(i)$;
9. end
10. 基于 $scr(q_a, r_i)$ 对资源 $r_i \in R$ 进行降序排序;
11. $rec(a) \leftarrow \text{top-k}$

代码的第 1 行,对资源排序进行计算并存储在向量 v^* 中。代码的第 3 行~第 7 行,计算查询配置文件 q_a 和每个资源 r_i 间的相似度,第 8 行计算其最终的匹配度,第 10 行计算其资源最终匹配分数排序,第 11 行利用 Top-k 机制进行资源推荐。

3 实验与结果分析

3.1 实验描述

本节实验基于公共学习资源数据集进行测试。自测指标选取相似性值、推荐覆盖范围、向量排名以及算法计算时间,对算法的可行性进行评估。对比指标选取精度/召回/MAE、RMSE 等指标进行对比,对比算法选取文献[16-17]算法。由于资源描述是

公众的标准格式,可以为实验提取和分析必要信息。

为此,本文开发了一个 Java 程序来抓取和提取公共资源。共收集到 1 294 个资源描述,其中涉及 62 家出版商(大学、工程学院等),14 种教学类型(幻灯片、动画、讲座、教程等),12 种格式(文本/HTML/MPEG、视频、应用程序/PDF 等),10 个不同教育层次(学校、中等教育、培训、BAC + 1、BAC + 2 等)。在 1 294 种资源中,880 种资源与其他资源相关,其中 692 种资源是“is part of”关系,333 种资源是“contains”关系,573 种资源是“associates to”关系。使用 Apache Lucene^[12] 进行词干提取,用 DISCO 库^[13] 对词语语义相似度进行计算。对 1 000 个查询进行模拟,其中每个查询包括 10 个关键字资源。

所提算法实验在 MAC 笔记本上执行,电脑配置为:CPU 2 GHz 酷睿 i7 处理器,内存为 8 GB 的 1 600 MHz,SSD 251g 和 OS X 10.9.2。

3.2 自测指标结果分析

实验 1 平均相似度分析

设置阻尼因子 $d = 0.85$,微分参数 $\alpha = 0.9$,对 1 000 个不同查询进行实验,并对每个查询设置 6 个不同情形进行推荐测试:

情形 1 (SKW) 考虑关键词和资源描述为奇异的关键词设置,并采用内容过滤算法。

情形 2 (CKW) 通过与查询的复合词匹配对资源描述进行预处理,并采用内容过滤算法。

情形 3 (SCB) 在查询中,利用最相似的词来替换资源描述中的单词,并采用内容过滤算法。

情形 4 (SKW-PageRank) 混合内容过滤与奇异词匹配算法,并基于关系进行资源排序查询。

情形 5 (CKW-PageRank) 混合内容过滤与复合词匹配算法,进行资源排序查询。

情形 6 (SCB-PageRank) 混合内容过滤和语义相似资源匹配算法,进行资源排名。

目标是测量查询和资源之间的相似性值。对于每一个查询,计算其 Top- k 资源平均相似值。实验结果如图 3 所示。

由图 3 可知,SCB 和 SCB-PageRank 算法可获得最高的平均相似度值,因为其考虑到句法和语义词的相似性。同时,该 CKW 和 CKW-PageRank 获取的相似度指标最低,因为其只考虑复合词之间的相似度匹配,其相似度指标较小。具有 PageRank 算法情况下,相似度指标较小,主要原因是资源的排序值是非常小的,以满足所有资源排序的总和等于 1, $\max = 9.64 \times 10^{-3}$, $\min = 1.78 \times 10^{-4}$ 。

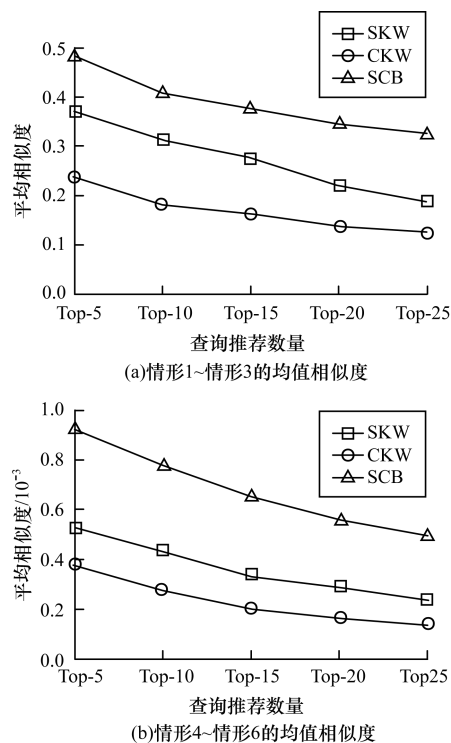


图 3 平均相似度值

实验 2 收敛对比

在本次实验中,对推荐过程的收敛性进行测量,即 Top- k 选择资源占总资源的百分比。收敛性选取 SKW*, CKW* 以及 SCB* 这 3 种情形进行对比,其中,SKW* 表示 SKW 和 SKW-PageRank 的收敛结果;CKW* 和 SCB* 含义与之类似。收敛性指标实验结果如图 4(a) 所示,收敛时间如图 4(b) 所示。

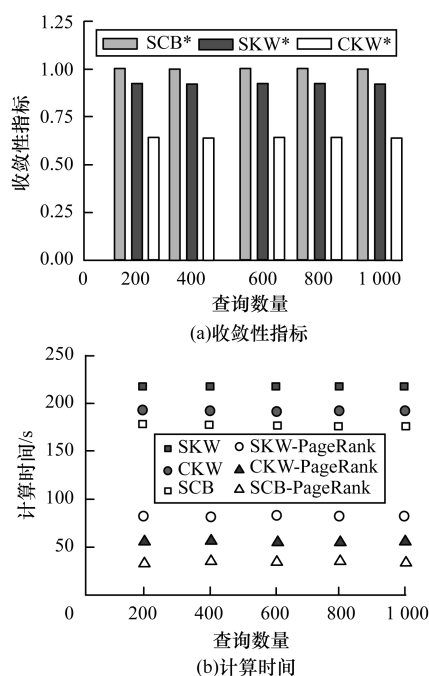


图 4 收敛对比

根据图 4(a)收敛性指标对比数据可知,CKW*(SKW 和 SKW-PageRank)具有最低的收敛指标,这是因为复合词匹配数比奇异词匹配和语义匹配数要小得多。同时,可看出 SCB*(SCB 和 SCB-PageRank)的收敛指标始终为 1,这意味着,对于每一个查询,总能在资源描述中找到至少一个字,其与查询中的字的语义匹配度大于 0。

根据图 4(b)计算时间对比数据可知,SCB 和 SCB-PageRank 算法在同等情况下的计算时间最短,这体现了算法的计算效率提升。同时综合图 4(a)和图 4(b)可知,算法的收敛指标和收敛时间与查询数量多少无关。

实验 3 阈值参数影响

本实验中,目标是衡量不同阈值情形下的资源查询过程收敛性,阈值选取 $10^{-4} \sim 10^{-9}$ 。实验结果如图 5 所示。

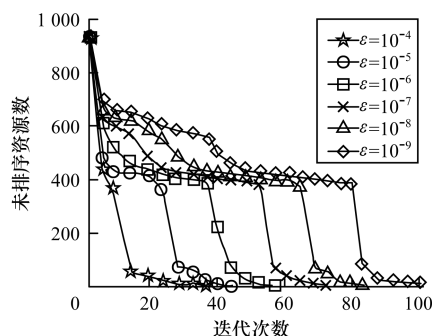


图 5 资源排序实验

根据图 5 资源排序实验可知,随着选取阈值的缩小,算法的收敛速度逐渐降低,例如,在 $\epsilon = 10^{-4}$ 时,算法的收敛步数约为 26,在 $\epsilon = 10^{-6}$ 时,算法的收敛步数增加为 58,而当 $\epsilon = 10^{-9}$ 时,算法的收敛步数增加为 101,可见阈值选取对于算法的收敛速度影响较大。

3.3 算法横向对比

对比指标选取查询精度指标、召回指标、平均绝对误差 (Mean Absolute Error, MAE)、均方根误差 (Root Mean Squared Error, RMSE)、多样性指标以及算法运行时间,其中,多样性指标主要衡量推荐系统对用户推荐内容的多样性。算法参数:选取查询数量为 200,Top-k=20,阈值参数选取 $\epsilon = 10^{-4}$ 。上述指标定义参见文献[18]算法,对比算法选取文献[16-17]算法。实验对比数据如表 2 所示。

表 2 算法对比实验结果

| 算法 | 精度/ % | 召回率/ % | MAE | RMSE | 多样性 | 计算 时间/s |
|--------------|----------|-----------|--------|-------|-------|------------|
| 文献[16]算法 | 86.7 | 85.4 | 15.637 | 3.874 | 0.534 | 64.25 |
| 文献[17]算法 | 91.4 | 78.3 | 13.296 | 2.915 | 0.689 | 68.69 |
| SCB-PageRank | 95.3 | 92.4 | 6.764 | 1.874 | 0.863 | 44.38 |

由表 2 数据可知,在精度指标上,SCB-PageRank 算法可达到 95.3%,高于文献[16]算法的 86.7% 和文献[17]算法的 91.4%,体现了 SCB-PageRank 算法的查询准确性;在召回率指标上,SCB-PageRank 算法可达到 92.4%,高于文献[16]算法的 85.4% 和文献[17]算法的 78.3%,体现了 SCB-PageRank 算法较高的查全率。在 MAE 和 RMSE 指标上,SCB-PageRank 算法指标数据分别为 6.764 和 1.874,小于对比算法,体现了 SCB-PageRank 算法较高的查询质量;在多样性指标方面,SCB-PageRank 算法为 0.863,高于文献[16]算法的 0.534 和文献[17]算法的 0.689,体现了 SCB-PageRank 算法查询分布的均匀性,可扩展性能较高;在计算时间上,SCB-PageRank 算法为 44.38 s,少于文献[16]算法的 64.25 s 和文献[17]算法的 68.69 s,体现了 SCB-PageRank 算法较高的计算效率。

4 结束语

本文提出基于内容过滤 PageRank 语义相似替换的 Top-k 学习资源推荐算法,建立基于资源间相互关系的权重值设定方式,与统一方式的权重值设定方式相比,可有效解决在线教育支持技术中关于文本处理的多义词和同义词问题,提高了算法的计算性能。与文献[16]算法和文献[17]算法相比,本文算法在精度、召回率、MAE、RMSE、多样性指标以及运行时间等指标上均优于对比算法,体现了算法良好的查询计算能力。在下一步的工作中,将继续探索在真实数据集上的算法性能测试;考虑融合资源等级,进行算法过滤,提高学习质量;并整合协同过滤和聚类技术,以提高学习推荐质量。

参考文献

- [1] Klačnja-Milicevic A, Ivanovic M, Nanopoulos A. Recommender Systems in e-learning Environments: A Survey of the State-of-the-Art and Possible Extensions[J]. Artificial Intelligence Review, 2015, 44(4): 571-604.
- [2] Keller M, Shrestha P M. Solute Accumulation Differs in the Vacuoles and Apoplast of Ripening Grape Berries[J]. Planta, 2014, 239(3): 633-642.
- [3] Liu Huanlin, Chen Gaoxiang, Chen Yong, et al. A Trust-based P2P Resource Search Method Integrating with Q-learning for Future Internet[J]. Peer-to-Peer Networking and Applications, 2015, 8(3): 532-542.
- [4] Arif M, Illahi M, Karim A, et al. An Architecture of Agent-based Multi-layer Interactive e-learning and e-testing Platform[J]. Quality & Quantity, 2015, 49(6): 2435-2458.

- [5] Udumyan N, Rouchier J, Ami D. Integration of Path-dependency in a Simple Learning Model; The Case of Marine Resources[J]. Computational Economics, 2014, 43(2):199-231.
- [6] Zervas P, Sergis S, Sampson D G, et al. Towards Competence-based Learning Design Driven Remote and Virtual Labs Recommendations for Science Teachers[J]. Technology, Knowledge and Learning, 2015, 20(2):185-199.
- [7] Anderson O R, Love B C, Tsai Meng-jung. Neuroscience Perspectives for Science and Mathematics Learning in Technology-enhanced Learning Environments [J]. International Journal of Science and Mathematics Education, 2014, 12(3):467-474.
- [8] 孙光福, 吴 乐, 刘 淇, 等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013, 24(11):2721-2733.
- [9] 李忠俊, 周启海, 帅青红. 一种基于内容和协同过滤同构化整合的推荐系统模型[J]. 计算机科学, 2009, 36(12):142-146.
- [10] Saparudin F A, Faisal N, Ghafar A S A. Distributed Resource Allocation for Femtocell Networks; Regret Learning with Proportional Self-belief [J]. Wireless Personal Communications, 2014, 79(1):453-471.
- [11] Campbell T, Longhurst M, Duffy A M. Science Teaching Orientations and Technology-enhanced Tools for Student Learning [J]. Research in Science Education, 2013, 43(5):2035-2057.
- [12] 彭泽环, 孙 乐, 韩先培. 基于排序学习的微博用户推荐[J]. 中文信息学报, 2013, 27(4):96-99.
- [13] Cober R, Tan E, Slotta J, et al. Teachers as Participatory Designers; Two Case Studies with Technology-enhanced Learning Environments[J]. Instructional Science, 2015, 43(2):203-228.
- [14] Parreira J X, Castillo C, Donato D, et al. The Juxtaposed Approximate PageRank Method for Robust PageRank Approximation in a Peer-to-Peer Web Search Network[J]. The VLDB Journal, 2008, 17(2):291-313.
- [15] Montefinese M, Zannino G D, Ambrosini E. Semantic Similarity Between Old and New Items Produces False Alarms in Recognition Memory [J]. Psychological Research, 2015, 79(5):785-794.
- [16] 刘志勇, 刘 磊, 刘萍萍. 一种基于语义网的个性化学习资源推荐算法[J]. 吉林大学学报(工学版), 2009, 39(2):391-398.
- [17] Lops P, Gemmis M, Semeraro G. Content-based and Collaborative Techniques for Tag Recommendation; An Empirical Evaluation [J]. Journal of Intelligent Information Systems, 2013, 40(1):41-61.
- [18] 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):163-172.

编辑 顾逸斐

(上接第 219 页)

- [7] Zhang Shaowu, Liu Huali, Yang Liang, et al. A Cross-domain Sentiment Classification Method Based on Extraction of Key Sentiment Sentence[C]//Proceedings of the 4th Conference on Natural Language Processing and Chinese Computing. Berlin, Germany: Springer, 2015: 90-101.
- [8] Zhou Guangyou, He Tingting, Wu Wensheng, et al. Linking Heterogeneous Input Features with Pivots for Domain Adaptation [C]//Proceedings of the 24th International Conference on Artificial Intelligence. Georgia, USA: AAAI Press, 2015: 1419-1425.
- [9] Bollegala D, Weir D, Carroll J. Cross-domain Sentiment Classification Using a Sentiment Sensitive Thesaurus[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(8):1719-1731.
- [10] Xia Rui, Zong Chengqing, Hu Xuelei, et al. Feature Ensemble Plus Sample Selection; Domain Adaptation for Sentiment Classification [J]. IEEE Intelligent Systems, 2013, 28(3):10-18.
- [11] Zhang Yuhong, Xu Xu, Hu Xuegang. A Common Subspace Construction Method in Cross-domain Sentiment Classification [C]//Proceedings of International Conference on Electronic Science and Automation Control. [S. l.]: Atlantis Press, 2015: 48-52.
- [12] Zhou Guanyou, Zhou Yin, Guo Xiyue, et al. Cross-domain Sentiment Classification via Topical Correspondence Transfer [J]. Neurocomputing, 2015, 159(1):298-305.
- [13] Bengio Y. Deep Learning of Representations for Unsupervised and Transfer Learning [J]. Unsupervised and Transfer Learning Challenges in Machine Learning, 2012, 7:19-41.
- [14] Glorot X, Bordes A, Bengio Y. Domain Adaptation for Large-scale Sentiment Classification; A Deep Learning Approach [C]//Proceedings of the 28th International Conference on Machine Learning. New York, USA: ACM Press, 2011: 513-520.
- [15] Collobert R, Weston J. A Unified Architecture for Natural Language Processing; Deep Neural Networks with Multitask Learning [C]//Proceedings of the 25th International Conference on Machine Learning. New York, USA: ACM Press, 2008: 160-167.

编辑 刘 冰