第 43 卷 第 4 期 计 算 机 工 程 2017 年 4 月
Vol. 43 No. 4 Computer Engineering April 2017

·人工智能及识别技术· 文章编号：1000-3428(2017)04-0207-05 文献标志码：A 中图分类号：TP18

基于网站层次结构和主题模型 LDA 的网站自动摘要

李舒媛，杨 静，顾君忠

（华东师范大学 计算机科学技术系，上海 200241）

摘 要： 近年来自动摘要方面的研究大多是关于多文档和 Web 网页的，而对网站自动摘要的研究较少。为此，基于主题模型隐含狄利克雷分布（LDA）和网站层次结构提出一个可以自动生成网站摘要的算法。该算法可获取整个网站内的网页信息并进行整合，根据提出的句子权重公式计算句子权重，选取权重最高的句子作为网站摘要。以 20 个商业和学术网站作为实验对象，使用 ROUGE 评测标准，结果表明，与仅使用主题模型 LDA 获取的网站摘要相比，不带停用词的 ROUGE-1 和 ROUGE-L 提高 0.32，带停用词的 ROUGE-1 提高 0.39，ROUGE-L 提高 0.38。与网站首页摘要相比，不带停用词的 ROUGE-1 提高 0.03，ROUGE-L 提高 0.06，带停用词的 ROUGE-1 提高 0.08，ROUGE-L 提高 0.07。

关键词： Web 网页；网站自动摘要；隐含狄利克雷分布；网站层次结构；宽度优先搜索

中文引用格式： 李舒媛，杨 静，顾君忠. 基于网站层次结构和主题模型 LDA 的网站自动摘要[J]. 计算机工程，2017，43(4)：207-211，216.

英文引用格式： Li Shu'ai，Yang Jing，Gu Junzhong. Website Automatic Summarization Based on Website Hierarchy and Latent Dirichlet Allocation[J]. Computer Engineering，2017，43(4)：207-211，216.

Website Automatic Summarization Based on Website Hierarchy and Latent Dirichlet Allocation

LI Shu'ai，YANG Jing，GU Junzhong

（Department of Computer Science and Technology，East China Normal University，Shanghai 200241，China）

【Abstract】 In recent years，the research of automatic summarization is mostly about multi-documents and Web pages，but less about website summarization. A method that summarizes a website automatically based on the hierarchical structure of the website and Latent Dirichlet Allocation is proposed. This method gets the information from web pages in the given website and fuses it，and calculates the weight of sentences according to the proposed sentence weighting formula，and selects the highest weight sentences as the website summarization. An experiment is done based on 20 commercial websites and academic websites，and using ROUGE evaluation. Results show that compared with the summaries only using LDA，ROUGE-1 and ROUGE-L are increased by 0.32 with no stop words；ROUGE-1 is increased by 0.39 and ROUGE-L is increased by 0.38 with stop words. Compared with the summaries only from homepage，ROUGE-1 is increased by 0.03 and ROUGE-L is increased by 0.06 with no stop words；ROUGE-1 is increased by 0.08 and ROUGE-L is increased by 0.07 with stop words.

【Key words】 Web pages；website automatic summarization；Latent Dirichlet Allocation（LDA）；website hierarchy；breadth-first search

DOI： 10.3969/j.issn.1000-3428.2017.04.035

0 概述

随着互联网络规模的不断扩大，大量的网络数据正在以飞快的速度增长，导致了信息过载[1]等问题。网站是网络信息的一个主要来源，然而网站复杂度的不断增长增加了用户寻找信息的难度，网站自动摘要可以帮助人们解决这个问题。目前，已经有了由志愿者人工总结的网站自动摘要，如开放式分类目录 DMOZ（Open Directory Project）[2]。但是人工生成的网站摘要需要花费大量的人力和时间维

基金项目： 国家科技支撑计划项目（2015BAH01F02）；上海张江国家自主创新示范区专项发展资金计划项目（201411-JA-B108-002）。

作者简介： 李舒媛（1993－），女，硕士研究生，主研方向为数据挖掘；杨 静，副教授；顾君忠，教授。

收稿日期： 2016-04-18 **修回日期：** 2016-05-25 **E-mail：** lshlsylshy@126.com

护,并且具有主观性。本文提出一种可以自动生成网站摘要的算法,将传统的句子统计特征结合网站的层次结构,得到句子的网站结构特征,再联合主题模型隐含狄利克雷分布(Latent Dirichlet Allocation,LDA),从统计特征和语义理解两方面获得网站摘要。首先对给定网站运用宽度优先搜索的爬虫程序抓取网站内的网页。然后利用 htmlparser 从下载的网页中抽取出纯文本并进行处理,得到符合 LDA 输入格式的句子集合。在句子集合上进行 LDA 建模得到句子的 LDA 特征权重。再结合网站的层次结构特征计算出句子的统计特征权重,两者相加得到最终的句子权重。最后选取 5 个权重最高的句子作为摘要结果。算法执行的过程中无需人工干预。

1 相关工作

1.1 文档自动摘要

根据算法不同,文档自动摘要中的"摘要"通常被分为两类:文摘(extractive summarization)和摘要(abstractive summarization)[3],extractive summarization 由从原始文档中抽取的句子组成[4],而 abstractive summarization 类似人工生成摘要,不再是仅使用原文当中的语句进行组合,而是对原始文档中的信息进行提炼并且融合,最后形成新的句子组成摘要。尽管 abstractive summarization 更加接近人工生成的摘要,更加简洁流畅,但由于涉及复杂的自然语言处理技术,限制了其适用的领域。相反,extractive summarization 在现有的技术支撑下更具实用性,因此,本文的研究聚焦于 extractive summarization 的生成。

早期的 extractive summarization 生成方法有文献[5-7]提出的根据句子特征获取摘要的方法。这些特征包括句子所包含词的频率、句子在文章中的位置以及句子中的线索词、标题词等。文献[8]提出的 LexPageRank 和文献[9]提出的 G-FLOW 是基于图的方法。这种方法是把文章的句子或段落作为图的顶点,句子或段落之间的关系作为图的边,最后通过图排序的算法计算出各顶点的得分,进而得到文本摘要。文献[10]提出了一种利用最大边缘相关性(Maximal Marginal Relevance,MMR)的方法选择句子,这个方法可以获得与文档主题相关性最大的句子,同时也可以把冗余性降到最低。近些年来潜在语义分析(Latent Semantic Analysis,LSA)也被应用到生成摘要当中[11]。

abstractive summarization 的生成方法有基于词汇链的方法[12]和基于文章修辞结构[13-15]的方法等。近期文献[16]提出在词图(word graph)中利用最短路径的方法。

1.2 Web 文档自动摘要

相较于传统文档规范的形式,Web 文档的结构多变、内容复杂等特性给 Web 文档自动摘要带来了新的挑战。网页中的额外的信息,例如用户评论[17]、标签[18]等成为生成 Web 文档摘要需要考虑的新元素。利用网页中的超链接[19-21]生成摘要适合包含较少文字内容的网页,这种方法假设网页中超链接的上下文一般是对其指向的网页的描述或总结,可以利用这些信息生成目标网页的摘要。文献[22]提出了一个自动获取网页主旨的原型系统 OCELOT。这个系统用自动翻译的方法产生网页的 abstractive summarization。Yves Petinot 等人在 2011 年和 2013 年分别提出了 2 种 Web 网页自动摘要方法:一种是基于 URL 相似的网页具有相同的内容和结构的聚类方法[23];另一个是基于 hierarchical LDA 的方法[24]。

关于网站的自动摘要的研究较少,大部分都是对单个网页摘要的研究。文献[2]以 DMOZ 中的网站作为数据,利用 C5.0 决策树的方法抽取关键词和关键句作为网站的摘要。他们提出的方法需要人工训练集,而本文提出的算法无需人工数据集,并从句子的网站结构特征和语义理解两个方面计算句子的权重,实验结果证明,该算法比通过 LDA 获取的网站首页摘要效果更好。

1.3 主题模型 LDA

LDA 是由 Blei[25]等人在 2003 年提出的一种文档主题生成模型,也被称为三层贝叶斯概率模型。它是由词、主题和文档三层结构构成的一种非监督机器学习技术,采用词袋(bag of words)的方式,目的是以无指导学习的方法从大规模文档集(document collection)或语料库(corpus)中发现隐含的语义维度,即"Topic"或者"Concept"[26]。

LDA 的生成方法用数字语言描述如下[26]:

```
// topic plate
for all topics k ∈ [1,K] do
  sample mixture components φ_k ~ Dir(β)
// document plate:
for all documents m ∈ [1,M] do
  sample mixture proportion ϑ_m ~ Dir(α)
  sample document length N_m ~ Poiss(ξ)
  // word plate
  for all words n ∈ [1,N_m] in document m do
    sample topic index z_{m,n} ~ Mult(ϑ_m)
    sample term for word w_{m,n} ~ Mult(φ_{z_{m,n}})
```

假设文档集 D,M 是文档总数,N_m 是第 m 个文档的单词总数,K 为主题个数。$z_{m,n}$是第 m 个文档中第 n 个词的主题,$w_{m,n}$是第 m 个文档中的第 n 个词。$\boldsymbol{\varphi}_k$ 是服从以 $\boldsymbol{\beta}$ 为参数的 Dirichlet 分布,它表示第 k 个主题下的词分布。$\boldsymbol{\vartheta}_m$ 是服从以 $\boldsymbol{\alpha}$ 为参数的 Dirichlet 分布,它表示第 m 个文档下的主题分布。给定一个文档集,$w_{m,n}$是可以观察到的已知变量,$\boldsymbol{\beta}$ 和 $\boldsymbol{\alpha}$ 是根据经验给定的先验参数,$z_{m,n}$,$\boldsymbol{\varphi}_k$,$\boldsymbol{\vartheta}_m$ 都是未知的隐含变量,本文利用 Gibbs Sampling 间接求得 $\boldsymbol{\varphi}_k$ 和 $\boldsymbol{\vartheta}_m$ 的值,所有变量的联合分布为[26]:

$$p(\boldsymbol{w}_m,\boldsymbol{z}_m,\boldsymbol{\vartheta}_m,\boldsymbol{\Phi}\mid\boldsymbol{\alpha},\boldsymbol{\beta})$$
$$=\prod_{n=1}^{N_m}p(w_{m,n}\mid\boldsymbol{\varphi}_{z_{m,n}})p(z_{m,n}\mid\boldsymbol{\vartheta}_m)\cdot p(\boldsymbol{\vartheta}_m\mid\boldsymbol{\alpha})$$
$$\cdot p(\boldsymbol{\Phi}\mid\boldsymbol{\beta})$$

2　基于网站层次结构和 LDA 的网站自动摘要

在传统的文档自动摘要算法中,有以句子的统计特征作为句子权重的计算依据,统计的范围包括词频、位置等信息。本文将传统的句子位置特征结合网站层次结构得到句子的网站结构特征,再联合 LDA 主题模型,达到从统计特征和语义理解两方面进行句子权重的计算。网站摘要的生成过程如图 1 所示。

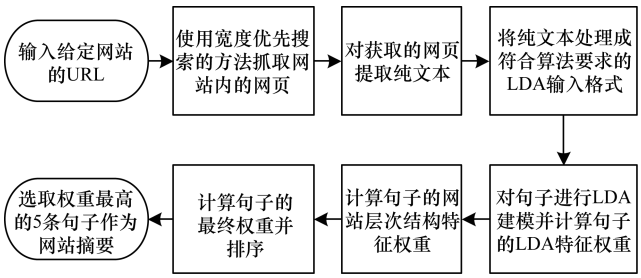

图 1　网站摘要生成过程

2.1　句子的 LDA 主题特征

LDA 的输入文档是给定文档集 $D=\{D_1,D_2,\cdots,D_M\}$,M 是文档个数,D_j 代表第 j 个文档的句子集合。本文以句子作为输入文档,所有文档的句子集合作为输入文档集,即 $D=\{s_1,s_2,\cdots,s_N\}$,其中,$s_i\in D$当且仅当 $s_i\in D_j\in D$,N 是 D 的句子总数。使用 LDA 在句子集合 D 上进行建模,并使用 Gibbs Sampling 学习 LDA 参数,得到句子在主题上的分布 $\boldsymbol{\vartheta}_{z=j}^{(s)}$和主题在词汇上的分布 $\boldsymbol{\varphi}_w^{(z=j)}$,基于这 2 种分布,本文提出了句子的 LDA 特征权重的计算方法。

2.1.1　主题重要度计算方法

由 Gibbs Sampling 学习得到的句子的主题分布 $\boldsymbol{\vartheta}_{z=j}^{(s)}$指的是每个主题在各个句子中所占的权重,每个主题的重要度可以由这个主题在每个句子上的权重加和来计算,然后在所有主题上进行归一化处理以得到合适的概率值:

$$P(z_i\mid D)=\frac{\sum_{n=1}^{N}\boldsymbol{\vartheta}_z^{(s)}}{\sum_{n=1}^{N}\sum_{n=1}^{N}\boldsymbol{\vartheta}_z^{(s)}}\tag{1}$$

其中,$i=1,2,\cdots,K$;$n=1,2,\cdots,N$,N 为句子文档个数;K 为句子集合中主题个数。

2.1.2　句子的 LDA 特征权重

句子的权重可由句子所包含的词汇的权重得到,词汇的权重即词汇的概率,由词汇在主题上的分布和主题的重要程度共同决定:

$$P(w\mid D)=\sum_{i=1}^{K}P(w\mid z_i)\times P(z_i\mid D)\tag{2}$$

其中,$i=1,2,\cdots,K$;$w_j\in s_k\in D$,K 为主题个数;$P(z_i\mid D)$是由式(1)得到的主题重要度;$P(w\mid z_i)$为使用 Gibbs Sampling 学习 LDA 模型得到的参数 $\boldsymbol{\varphi}_w^{(z=j)}$,它是主题 z_i 在词汇 w 上的概率。把通过上面的式子得到的词汇权重带入到词汇所属的句子中,就可以得到句子的 LDA 特征权重:

$$P(s_j\mid D)=\sum_{i=1}^{n}P(w_i\mid D)\tag{3}$$

其中,$i=1,2,\cdots,n$;$w_j\in s_k\in D$,n 为句子 s_j 中的单词个数。一般来说,长句子比短句子包含的信息量要多[27],选择词汇概率的加和进行计算可以忽略句子的长短限制。在这种情况下,句子的权重完全由句子所包含词的权重决定。

2.2　句子的网站结构特征

从网站的层次结构来看,首页信息一般是这个网站的简短介绍,网站内容分几个模块依次展现。以宽度优先搜索的方法抽取网站网页,可以按照网站的层次结构依次获取每层的网页内容,这样获得的网站综合文档是按照网站层次结构排列的,可明确区分处于最高层次的首页内容和低层网页的内容。有研究表明,人工摘要中的句子为段首句的比例为 85%,段尾句的比例为 7%[5]。文献[28]根据上述研究把文档中的段首句、段尾句和其他句子分别分配了 0.8,0.2 和 0 的比例权重,本文以句子作为计算单元,可将网站综合文档看成一段文字,把首页内容看成段首句,把除了首页之外的网页都看成是段尾句,即可以将文档集 D 中的句子 s_i 的网站层次结构特征 $SCORE_L(s_i)$定义为:

$$SCORE\_L(s_i)=\begin{cases}0.8, s_i\in homepage\\ 0.2, \text{others}\end{cases}\tag{4}$$

通过式(3)可以得到句子的 LDA 特征权重,式(4)得到句子的网站层次结构特征权重,这两者分别代表着句子的统计特征和语义特征。本文将两者加和作为最终的句子权重,没有考虑两者的权重比例,最终句子的权重 $SCORE(s_i)$为:

$$SCORE(s_i)=SCORE\_L(s_i)+P(s_j\mid D)\tag{5}$$

2.3　文摘算法

文摘算法步骤如下:

1)使用宽度优先搜索(breadth-first-search)的方法,爬虫抓取同一网站内具有相同域名的网页,然后利用 htmlparser 从网页的内容里提取纯文本文档。网页中的导航条、广告条等非正文文本信息大多数以非句子形式存在,本文用标点符号来辨别并剔除这些信息。

对每个文档进行去重生成一个包含所有网页内容的综合文档。将综合文档以整句的形式划分为多个句子文档。将得到的句子文档集作为 LDA 的输入文档集,去标点和停用词,将其转化为 LDA 的输入格式。

2)为每个文档建立 LDA 模型,并用 Gibbs Sampling 学习模型得到句子在主题上的分布 $\vartheta_{z=j}^{(s)}$ 和主题在词汇上的分布 $\varphi_w^{(z=j)}$。

3)根据式(1)~式(3),计算句子的 LDA 特征权重。

4)根据式(4),为每个句子分配不同的网站层次结构特征权重。

5)利用式(5)得出句子的权重得分,对句子权重进行排序,从前到后抽取 5 条句子形成摘要。

3 实验结果与分析

本文从 DMOZ 开放式分类目录中随机选取 20 个学术网站和商业网站作为实验对象,它们分别来自 Software / Software Engineering,Artificial Intelligence / Academic Departments, Major Companies/Publicly Traded,E-Commerce / Technology Vendors。每个子目录选取 5 个网站,运用本文提出的算法对这些网站进行计算得到网站摘要。其中对网站 http://www. sei. cmu. edu 抽取结果如图 2 所示。这个网站是关于学术机构 SEI 的网站,网站分别从 Work Areas,Engage with Us,Products &Services,Library,Careers,News,About Us 等方面介绍了 SEI 的工作内容。

The SEI helps advance software engineering principles and practices and serves as a national resource in software engineering, computer security,and process improvement.
Therefore, our singular dependency on assured software as the heart of this cyber environment is more prevalent than ever.
The size and complexity of software, as well as the interconnectedness of software-enabled systems, mean possible exposure to disruptive, damaging events.
The best way to assure software quality, security,and resiliency is to design,develop,and integrate software in a way that does not allow defects and vulnerabilities in the first place.
Ourwork is not done unless we do both parts of our job.

图 2 SEI 网站摘要

ROUGE 是一种基于召回率的统计方法,主要评估机器生成摘要与专家摘要之间(如 n-gram、词序、词对等)同现单元的个数。ROUGE-N 统计候选摘要与专家摘要之间 n-gram 的共现个数,ROUGE-L 统计摘要之间的最长公共子序列,ROUGE-W 是加权的 ROUGE-L, ROUGE-SU 是 ROUGE-S 的延伸[29]。实验中使用这 4 个评测标准,分别使用带停用词和去停用词 2 种计算方式对本文提出的摘要生成方法进行评测。人工生成的 DMOZ 摘要和算法生成的摘要都是用 Porter Stemmer 取词干。

本文算法的目的是自动生成网站摘要,并且生成的摘要效果比只浏览网站首页得到的信息更多,所以把只浏览首页得到的网站信息作为对比。对网站首页进行纯文本抽取,运用部分的 LDA 模型计算句子权重,然后选取权重最高的 5 条句子作为只浏览首页得到的网站摘要。同时把只考虑 LDA 而不考虑网站层次结构得到的摘要也作为对比项。表 1 和表 2 分别给出了去停用词和不去停用词的对比实验结果,其中,LDA 表示只进行 LDA 建模而不考虑网站层次结构得到的摘要;HOMEPAGELDA 即对网站首页进行 LDA 建模,计算句子权重后得到的网站首页摘要;NEWPROPOSED 即本文提出的算法。

表 1 去停用词实验结果

算法	ROUGE-1	ROUGE-L	ROUGE-W-1.2	ROUGE-SU *
LDA	0.064 52	0.064 52	0.033 22	0.004 04
HOME PAGELDA	0.354 84	0.322 58	0.143 48	0.117 17
NEW PROPOSED	0.387 10	0.387 10	0.192 04	0.139 39

表 2 带停用词实验结果

算法	ROUGE-1	ROUGE-L	ROUGE-W-1.2	ROUGE-SU *
LDA	0.172 41	0.155 17	0.057 87	0.032 16
HOMEPAGELDA	0.482 76	0.362 07	0.122 56	0.202 34
NEW PROPOSED	0.568 97	0.431 03	0.172 62	0.273 10

对于相同算法而言,带停用词的结果要优于去停用词的结果,因为停用词的加入会导致统计结果的增加。通过表 1 和表 2 的对比结果可以看出,无论是带停用词还是去停用词,本文提出的方法都要优于只对网站首页应用式(4)计算得出的摘要,和对整个网站应用式(4)而没有考虑网站层次结构计算得出的摘要。说明本文提出的算法可以得到更好的网站摘要,比只浏览网站首页能获得更全面的网站信息,同时也可以看出,考虑网站层次结构特征可以有效地提高摘要的质量。

4 结束语

本文基于网站层次结构和主题模型 LDA 提出了一个自动生成 Web 网站摘要的算法。首先利用宽度优先搜索获取网站内的网页,从网页中提取纯文本,然后对纯文本中的句子集合进行 LDA 建模,计算句子的 LDA 特征权重;利用网站层次结构得到

句子的网站结构特征权重,两者加和得到句子最终的权重。实验结果表明,本文方法得到的网站摘要比从首页获得的网站摘要效果要好。本文方法适用于机构类网站和企业类网站,这 2 种类型的网站内容以机构和企业的基本形象和服务内容为主。网站的首页通常是对机构或企业的简单介绍,然后分模块详细介绍研究内容或产品服务等。这些模块在网站层次结构上的位置如同一棵树的不同子树,而首页相当于这棵树的根节点。

下一步可以从以下 3 个方面提升本文方法的性能:

1)本文方法依赖于网站内的文本数据,如果网站包含的图片或者链接过多,将会影响摘要的准确率,可以考虑利用文献[19]的方法处理这些页面。

2)LDA 的主题个数的确定。本文选取 30 个主题个数,实验过程中发现选择主题个数为 30 时结果最好,下一步将考虑如何自动确定主题个数。

3)本文方法仅区分了网站层次结构中的第一层次和其他层次,以后的工作中可以更进一步地细分网站层次,分析并加以利用网站的多层特征。

参考文献

[1] Mani I, Maybury M T. Advances in Automatic Text Summarization[M]. [S.1.]:MIT press,1999.

[2] Zhang Y, Zincir-Heywood N, Milios E. World Wide Web Site Summarization[J]. Web Intelligence and Agent Systems,2004,2(1):39-53.

[3] Knight K, Marcu D. Summarization Beyond Sentence Extraction: A Probabilistic Approach to Sentence Compression[J]. Artificial Intelligence,2002,139(1):91-107.

[4] Jing H, McKeown K R. Cut and Paste Based Text Summarization[C]//Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference. [S.1.]:Association for Computational Linguistics,2000:178-185.

[5] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research and Development,1958,2(2):159-165.

[6] Baxendale P B. Machine-made Index for Technical Literature:An Experiment[J]. IBM Journal of Research and Development,1958,2(4):354-361.

[7] Edmundson H P. New Methods in Automatic Extracting[J]. Journal of the ACM,1969,16(2):264-285.

[8] Erkan G, Radev D R. LexPageRank: Prestige in Multi-document Text Summarization[C]//Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing. Washington D. C., USA: IEEE Press,2004:365-371.

[9] Christensen J, Mausam S S, Soderland S, et al. Towards Coherent Multi-document Summarization[C]//Proceedings of HLT-NAACL'13. Washington D. C., USA: IEEE Press,2013:1163-1173.

[10] Carbonell J, Goldstein J. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press,1998:335-336.

[11] Gong Y, Liu X. Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis[C]//Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2001:19-25.

[12] Barzilay R, Elhadad M. Using Lexical Chains for Text Summarization[C]//Proceedings of Workshop on Advances in Automatic Text Summarization. New York, USA:ACM Press,1999:111-121.

[13] Ono K, Sumita K, Miike S. Abstract Generation Based on Rhetorical Structure Extraction[C]//Proceedings of the 15th Conference on Computational Linguistics-Volume 1. [S. 1.]: Association for Computational Linguistics,1994:344-348.

[14] Marcu D. Improving Summarization Through Rhetorical Parsing Tuning[C]//Proceedings of the 6th Workshop on Very Large. Washington D. C., USA: IEEE Press, 1998:206-215.

[15] Marcu D. The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts[D]. Toronto, Canada:University of Toronto,1998.

[16] Filippova K. Multi-sentence Compression: Finding Shortest Paths in Word Graphs[C]//Proceedings of the 23rd International Conference on Computational Linguistics. [S. 1.]: Association for Computational Linguistics,2010:322-330.

[17] Hu M, Sun A, Lim E P. Comments-oriented Document Summarization: Understanding Documents with Readers' Feedback[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA:ACM Press,2008:291-298.

[18] Park J, Fukuhara T, Ohmukai I, et al. Web Content Summarization Using Social Bookmarks: A New Approach for Social Summarization[C]//Proceedings of the 10th ACM Workshop on Web Information and Data Management. New York, USA:ACM Press,2008:103-110.

[19] Sun J T, Shen D, Zeng H J, et al. Web-page Summarization Using Clickthrough Data[C]//Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2005:194-201.

[20] Delort J Y, Bouchon-Meunier B, Rifqi M. Enhanced Web Document Summarization Using Hyperlinks[C]//Proceedings of the 14th ACM Conference on Hypertext and Hypermedia. New York, USA:ACM Press,2003:208-215.

[21] Amitay E, Paris C. Automatically Summarizing Web Sites:Is There a Way Around It?[C]//Proceedings of the 9th International Conference on Information and Knowledge Management. New York, USA:ACM Press, 2000:173-179.

(下转第 216 页)

于商品历史价格曲线的评分、标签对商品基于商品历史价格曲线的重要程度进行二部图物质扩展,并考虑用户、标签、商品的度以及用户对标签的兴趣习惯进行加权推荐。实验结果表明,该方法的多样性、召回率优势显著。下一步研究的重点是利用社会化标签进行大数据的个性化推荐,降低大数据下推荐的复杂度。

参考文献

[1] 郭 磊,马 军,陈竹敏,等. 一种结合推荐对象间关联关系的社会化推荐算法[J]. 计算机学报,2014,37(1):219-228.

[2] 邹本友,李翠平,谭力文,等. 基于用户信任和张量分解的社会网络推荐[J]. 软件学报,2014,25(12):2852-2864.

[3] Cantador I, Bellogín A, Vallet D. Content-based Recommendation in Social Tagging Systems[C]// Proceedings of the 4th ACM Conference on Recommender Systems. New York, USA: ACM Press, 2010: 237-240.

[4] Rafailidis D, Daras P. The TFC Model: Tensor Factorization and Tag Clustering for Item Recommendation in Social Tagging Systems[J]. IEEE Transactions on Systems, Man, and Cybernetics Systems, 2013, 43(3):673-688.

[5] Liu Rujuan, Niu Zhendong. A Collaborative Filtering Recommendation Algorithm Based on Tag Clustering[M]//Park J J, Stojmenovic I, Choi M, et al. Future Information Technology. Berlin, Germany: Springer, 2014:177-183.

[6] 马费成,张 斌. 图书标注环境下用户的认知特征[J]. 中国图书馆学报,2014(1):4-14.

[7] 王 军,张子柯. 基于社会化标签信息熵的个性化推荐算法[J]. 图书情报工作,2013(23):31-35.

[8] 琚春华,鲍福光,刘中军. 基于社会化评分和标签的个性化推荐方法[J]. 情报学报,2014(12):1302-1310.

[9] Durao F, Dolog P. A Personalized Tag-based Recommendation in Social Web Systems[C]//Proceedings of International Workshop on Adaptation & Personalization for Web. Washington D. C., USA: IEEE Press, 2012:40-49.

[10] 孔欣欣,苏本昌,王宏志,等. 基于标签权重评分的推荐模型及算法研究[J]. 计算机学报,2015,38(23):1-13.

[11] 房小可,纪春光. 基于标签主题和概念空间的个性化推荐研究[J]. 情报理论与实践,2015,38(5):105-111.

[12] Zhang Zike, Zhou Tao, Zhang Yicheng. Personalized Recommendation via Integrated Diffusion on User-Item-Tag Tripartite Graphs[J]. Statistical Mechanics and Its Applications, 2010, 389(1):179-186.

[13] Zhang Yin, Zhang Bin, Gao Kening, et al. Combining Content and Relation Analysis for Recommendation in Social Tagging Systems[J]. Statistical Mechanics and Its Applications, 2012, 391(22):5759-5768.

[14] Lian Jie, Liu Yun, Zhang Zhenjiang, et al. Personalized Recommendation via an Improved NBI Algorithm and User Influence Model in a Microblog Network[J]. Statistical Mechanics and Its Applications, 2013, 392(19):4594-4605.

[15] Zhang Zike, Yu Lu, Fang Kuan, et al. Website-oriented Recommendation Based on Heat Spreading and Tag-aware Collaborative Filtering[J]. Statistical Mechanics and Its Applications, 2014, 393(4):82-88.

[16] Mao Jin, Lu Kun, Li Gang, et al. Profiling Users with Tag Networks in Diffusion-based Personalized Recommendation[J]. Journal of Information Science, 2016, 42(5):75-89.

编辑 顾逸斐

(上接第 211 页)

[22] Berger A L, Mittal V O. OCELOT: A System for Summarizing Web Pages[C]//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2000:144-151.

[23] Petinot Y, McKeown K, Thadani K. Cluster-based Web Summarization[C]//Proceedings of IJCNLP'13. New York, USA: ACM Press, 2013:1124-1128.

[24] Petinot Y, McKeown K, Thadani K. A Hierarchical Model of Web Summaries[C]//Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers-Volume 2. [S. l.]: Association for Computational Linguistics, 2011:670-675.

[25] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3(3):993-1022.

[26] Heinrich G. Parameter Estimation for Text Analysis[D]. Leipzig, Germany: University of Leipzig, 2008.

[27] 杨 潇,马 军,杨同峰,等. 主题模型 LDA 的多文档自动文摘[J]. 智能系统学报,2010,5(2):169-176.

[28] 刘 娜,路 莹,唐晓君,等. 基于 LDA 重要主题的多文档自动摘要算法[J]. 计算机科学与探索,2015,9(2):242-248.

[29] Lin C Y. Rouge: A Package for Automatic Evaluation of Summaries[C]//Proceedings of ACL'04. New York, USA: ACM Press, 2004:8-16.

编辑 索书志