软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(12):3241−3256 [doi: 10.13328/j.cnki.jos.005285]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于主题与概率模型的非合作深网数据源选择*

邓 松[1,3], 万常选[2,3]

[1](江西财经大学 软件与通信工程学院,江西 南昌 330013)

[2](江西财经大学 信息管理学院,江西 南昌 330013)

[3](数据与知识工程江西省高校重点实验室(江西财经大学),江西 南昌 330013)

通讯作者: 邓松, E-mail: daonicool@sina.com

摘 要: 在深网数据集成过程中,用户希望仅检索少量数据源便能获取高质量的检索结果,因而数据源选择成为其核心技术.为满足基于相关性和多样性的集成检索需求,提出一种适合小规模抽样文档摘要的深网数据源选择方法.该方法在数据源选择过程中首先度量数据源与用户查询的相关性,然后进一步考虑候选数据源提供数据的多样性.为提升数据源相关性判别的准确性,构建了基于层次主题的数据源摘要,并在其中引入了主题内容相关性偏差概率模型,且给出了基于人工反馈的偏差概率模型构建方法以及基于概率分析的数据源相关性度量方法.为提升数据源选择结果的多样性程度,在基于层次主题的数据源摘要中建立了多样性链接有向边,并给出了数据源多样性的评价方法.最后,将基于相关性和多样性的数据源选择问题转化为一个组合优化问题,提出了基于优化函数的数据源选择策略.实验结果表明:在基于少量抽样文档进行数据源选择时,该方法具有较高的选择准确率.

关键词: 深网;数据源选择;主题;概率模型;TextRank

中图法分类号: TP311

中文引用格式: 邓松,万常选.基于主题与概率模型的非合作深网数据源选择.软件学报,2017,28(12):3241−3256. http://www.jos.org.cn/1000-9825/5285.htm

英文引用格式: Deng S, Wan CX. Non-Cooperative deep Web data source selection based on subject and probability model. Ruan Jian Xue Bao/Journal of Software, 2017,28(12):3241−3256 (in Chinese). http://www.jos.org.cn/1000-9825/5285.htm

Non-Cooperative Deep Web Data Source Selection Based on Subject and Probability Model

DENG Song[1,3], WAN Chang-Xuan[2,3]

[1](School of Software & Communication Engineering, Jiangxi University of Finance and Economics, Nanchang 330013, China)

[2](School of Information and Technology, Jiangxi University of Finance and Economics, Nanchang 330013, China)

[3](Jiangxi Key Laboratory of Data and Knowledge Engineering (Jiangxi University of Finance and Economics), Nanchang 330013, China)

Abstract: It is desirable for a user to get high-quality query results from only a few data sources in deep Web data integration systems. Therefore, data source selection becomes one of the core technologies in the integration systems. In this paper, a method based on correlations and diversities is proposed for selecting deep Web data sources suitable for small-scale sampling document summaries. Firstly, considering the correlations between the query and the data sources, a hierarchical subject summary with a probability model of correlation deviation of the data sources is constructed to discriminate the data sources. Furthermore, a method is described for constructing a deviation probability model based on artificial feedbacks and correlation measurement of the data sources. Meanwhile, the diversity-oriented directed edges are built in the hierarchical subject summary of data source in consideration of the diversities of data

* 基金项目: 国家自然科学基金(61462037, 61562032, 61173146, 61363039, 61363010); 江西省自然科学基金(20152ACB20003); 江西省高等学校科技落地计划(KJLD12022, KJLD14035)

Foundation item: National Natural Science Foundation of China (61462037, 61562032, 61173146, 61363039, 61363010); Natural Science Foundation of Jiangxi Province of China (20152ACB20003); Science and Technology Landing Plan of Colleges in Jiangxi Province of China (KJLD12022, KJLD14035)

收稿时间: 2016-10-12; 修改时间: 2016-11-29, 2017-01-24, 2017-03-09, 2017-03-21; 采用时间: 2017-03-28

sources, and an evaluation metric is proposed to measure data source diversities. Taking the data source selection based on correlation and diversity as a combinatorial optimization problem, an optimal result of data source selection is achieved by solving an optimization function. Experimental results show that the proposed method achieves better selection accuracy in selecting data sources with small sampling documents.

Key words: deep Web; data source selection; subject; probability model; TextRank

在当前的互联网环境下,深网数据源占据较大比例,传统爬虫技术难以有效获取其中的信息资源.深网数据源中的信息需要向搜索接口提交查询才可以获取,如果一个用户需要集成检索一定规模的深网数据源中的数据,早期的做法是向每个深网数据源的接口提交相应的查询以获取相关结果.当前,各领域相关深网数据源成百上千,以上工作显然是非常耗时且令人疲惫的,因此,深网数据集成系统应运而生.

为了帮助用户更容易地使用各深网中的资源,深网数据集成系统建立统一的元查询接口.元查询接口可以使得用户提交的一个查询自动转换成各数据源接口能够接受的查询语句.如果通过以上方法检索各领域下每个数据源以获取用户想要的结果,效率将十分低下.另外,由于深网数据源质量相差较大,且用户通常只对排名较前的检索结果感兴趣,因此人们希望能够在真正执行检索之前获知最佳结果在各数据源中的分布情况,由此产生了数据源选择技术.

数据源选择技术可以使得用户只检索少量几个数据源便可以获取较理想的结果.在有数据源选择部件的集成检索框架中,元搜索接口保存着各数据源的摘要.当一个查询到来时,依据数据源摘要满足用户查询的程度,就可以判定出将用于真实提交查询的 Top-k 数据源.为了提升数据源选择的准确性,面向用户查询的数据源摘要的构建以及基于摘要的数据源评价方法就成为一个关键问题.

由于通常情况下数据源是非合作的,即,不会向使用者自动提供其全部数据,为了构建深网数据源摘要,需要通过抽样技术获取深网中的相关数据分布情况.深网有结构化和非结构化两种类型,其中,非结构化深网数量较多,本文主要针对非结构化深网数据源选择展开相关研究.用户集成检索时,通常会特别关注检索结果与查询的相关性、检索结果的非重复度(即多样性).为了便于说明,本文把数据源返回的检索结果与查询的相关程度称为相关性,数据源返回的检索结果的非重复程度定义为多样性.目前,已有的非结构化深网数据源选择方法较多地考虑了数据源返回的检索结果与查询的相关程度,即仅考虑相关性,少量研究成果考虑了合作环境下基于相关性和多样性的数据源选择问题.非合作环境下,为保证数据源选择的效率,数据源摘要通常仅保留少量词项或文档数据,在此基础上进行基于相关性和多样性的深网数据源选择,这是本文的主要着眼点.

非合作环境下,基于相关性和多样性进行数据源选择需要建立相应的数据源摘要.与基于词项构建数据源摘要相比,基于抽样文档构建数据源摘要,数据源选择的效果会更好[1].针对一个领域,数据源中的文本内容通常涉及多个相对固定的主题,且每个主题下的文档内容关联性较强,如汽车领域数据源包含发动机、轮胎、离合器等主题,每个主题又含有各自的子主题.

因此,本文基于层次主题构建数据源摘要,出发点如下:(1) 基于一个数据源中相同主题下抽样文档内容相关的特点,可以提升数据源相关性判别的准确度;(2) 基于不同数据源相同主题下抽样文档的多样性程度,可以有助于估算不同数据源提供检索结果的多样性.

数据源摘要中,与用户查询相关的抽样文档的代表性是有限的,因此,数据源与查询的相关性一般是通过用户查询相对于数据源摘要各层次主题内容的相关性估算得分来判别.如果能够事先获知该相关性估算得分与用户查询相对于真实数据源的相关性得分(称为真实相关性得分)的偏差概率分布,则可以基于偏差概率选用合适方法调整相关性估算得分,这样就可以进一步提升相关性判别的准确率.

由于一个文档可能包含很多不同方面的内容,因此,即使是同一主题下的两个抽样文档,也可能包含不同方面的内容.由于每篇文档不同方面的内容可由不同的特征词来表征,因此,本文把每篇文档中用特征词表征的不同方面的内容称为文档特征面.一个数据源给定主题下的抽样文档内容可以包含多个文档特征面,文档特征面越多,则表示文档内容的多样性越好.因此,本文依据不同数据源摘要中相同主题下抽样文档所包含特征词的多样性程度来判别该数据源检索结果的多样性.

综上所述,本文采用两阶段法选择 Top-K 深网:首先,基于数据源摘要计算各候选数据源与用户查询的相关性估算得分;然后,综合考虑数据源与用户查询的相关性和数据源的多样性,选择相关性较大、多样性较好(即综合性能达到最优)的 Top-K 个数据源.实验结果表明:本文方法可以较好地满足基于相关性和多样性的数据源选择需求,有着较好的应用前景.

本文第 1 节对已有非结构化深网数据源选择方法进行综述,并分析和总结本文研究工作的创新性.第 2 节阐述基于主题的数据源抽样与摘要构建技术.第 3 节提出改进的数据源相关性判别方法,并构建相关性偏差概率分布模型,给出使用方法.第 4 节介绍多样性计算策略.第 5 节提出基于相关性和多样性的数据源选择策略.第 6 节分析相关实验结果.第 7 节总结全文.

1 相关工作

为了提升实时数据源选择效率,非合作环境下非结构化深网数据源选择使用的源摘要通常较为精炼,一般仅保留少量的深网抽样文档.尽管存在较多的数据源抽样方法,但是基于少量抽样文档构建摘要并实现数据源选择的相关方法较少.一般都是直接使有 RS-Ord 或 RS-Lrd 抽样技术[1]进行深网抽样,研究重点主要放在摘要构建和数据源评价上.RS-Ord 抽样法是从字典中随机选词,而 RS-Lrd 抽样法则是先随机获取少量初始抽样文档,然后基于抽样文档中各词项的词频选择下一个抽样查询词.RS-Ord 抽样法构建的摘要质量稳定性弱于 RS-Lrd.个别数据源选择方法中使用了人工抽样法[2]或实时采样法[3],其中,人工方法效果较好但需要较多的人工参与.对于实时采样法,首先向真实数据源提交查询获取前 N 篇文档,然后通过下载计算文档得分进行数据源选择,该方法需要较大的网络花销.

由于给定领域下的一个数据源中的内容涉及多个相对固定的主题,且每个主题下的文档内容关联性较强,因此针对一个特定领域,可以事先获取该领域的主题词层次模型,然后基于层次主题词构建相关查询进行深网数据采样,这样可以提升摘要中少量抽样文档的主题代表性.

非合作环境下非结构化深网数据源选择技术主要可以分为面向单一相关性检索需求、面向个性化检索需求两种类型.

1.1 面向单一相关性检索需求

面向单一相关性检索需求进行数据源选择的研究开始较早,取得了较多研究成果.研究人员主要从挖掘词项与文档相关信息、扩展摘要信息、分析查询日志隐含信息这 3 个角度进行数据源选择.

(1) 基于词项和文档选择数据源

CORI[4]是较为经典的非结构化深网数据源选择方法,该方法依据抽样词频和词项逆文档频率信息构建数据源摘要,并基于摘要计算数据源与用户查询的相关度.为进一步提升数据源选择准确率,文献[5]提出了集中排序数据源选择方法,设计了一种非线性的抽样文档排名与得分转换策略,该方法对应模型中的参数值对数据源选择的影响较大,且没有科学的参数设置方法.SUSHI 算法[6]进一步考虑了文档抽样比例问题,对数据源排序后的结果进行拟合,基于内插值选择最佳数据源.以上方法中,估计参量较多,各方法面对不同的测试数据具有各自的优势[7].为减少数据源选择效果的不确定性,文献[8]在选择数据源时使用以上多种方法,基于投票策略使用综合得分确定数据源的排名.

(2) 基于主题内容特征选择数据源

由于 RS-Ord 或 RS-Lrd 随机抽样方法获取的数据有限,当一个数据源数据量较大时,会丢失很多低频信息.基于同主题数据源倾向有相似内容摘要的假设,Ipeirotis 提出了一种基于数据源分层分类的数据源选择算法[9].同样针对小样本信息缺失问题,文献[10]引入 LDA 分别描述数据源主题内容和用户查询主题的概率分布,基于两个概率分布的相近性衡量数据源和查询的相关性,在某些数据集上一定程度地提升了数据源选择的准确率.

(3) 基于查询日志选择数据源

查询日志与用户对该数据源的检索需求紧密相关,因此,部分数据源选择算法基于已使用查询与用户给定查询的相似度评价数据源的相关性得分[11,12].基于查询词隐藏的丰富模式信息,文献[13]提出了一种基于查询

日志获取 KA(关键词-领域属性)关联构建数据源摘要的数据源选择方法,对模型信息较丰富的数据源,选择的准确率较高.

基于日志进行数据源选择,需要搜索引擎公司提供相关商业数据.基于词项和文档选择数据源的方法为提升小抽样文档摘要下数据源选择准确率,对抽样文档信息进行了深入挖掘,但小规模抽样文档信息代表性有限.文献[9]基于同主题数据源有相似内容的假设,丰富抽样获取的数据源摘要内容以提升数据源选择的准确率,局限性较强,因为现实深网中部分同主题数据源中内容相差较大.文献[10]基于少量抽样文档构建的数据源 LDA 主题模型同样存在内容代表性有限的问题.

通常情况下,一个数据源中具体主题下的文档内容关联程度较强.如果能在小抽样文档摘要的基础上增加一些信息,如进一步建立用户查询基于数据源摘要主题内容的相关性估算得分与真实相关性得分的偏差概率模型,以便更好地确定与用户查询最相关的候选数据源排序,将可以进一步提升数据源相关性判别的准确率.

1.2 面向个性化检索需求

数据源选择最主要的依据是数据源内容与查询的相关性,在结果相关的基础上加入其他考量,可以满足不同的用户需求.面对平衡数据源选择算法效果与花费方面的需求,文献[14]受经济学中边际原理启发,基于边际收益整合数据源,具有较好的可扩展性.面对用户提出的数据时效性需求,文献[15]提出了一组基于时效性的评价指标进行数据源选择,把数据源选择问题转化为 NP 难问题,并给出了近似解决方法.为了解决 P2P 领域数据源中数据重复情况较为严重的问题,Bender[16]基于布隆过滤器技术存储索引数据,并据此计算数据源之间的数据重复度,再综合考虑相关度得分和数据源重复度进行数据源选择.用户可能提出不同数据质量或预算的需求,为此,文献[17]提出了基于多对象优化的数据源选择方法,其中,针对数据重复性问题,在不考虑相关性基础上,基于簇内容相似性进行判别,且假设簇中内容可以全部获取.

数据源选择时,同时考虑相关性和多样性是较多用户可能提出的需求.文献[16]基于 P2P 领域的数据特点设计相应的数据源选择方法;文献[17]尽管初步考虑了数据源多样性问题,但未同时考虑相关性.

因不能事先在摘要构建时获知具体用户查询,因此,本文利用不同数据源摘要中相同主题下抽样文档的特征面来估算数据源返回检索结果的多样性.不同数据源摘要中,相同主题下抽样文档的特征面越多,意味着数据源返回检索结果的多样性越好.而文档特征面又可以通过文档中所包含的特征词来表征,因此可以进一步依据不同数据源摘要中相同主题下抽样文档所包含的特征词的多样性程度来判别数据源返回检索结果的多样性.

综上所述,为满足基于相关性和多样性的文本型深网数据源选择需求,提出了基于主题与概率模型的非合作深网数据源选择策略,创新性主要体现在如下几个方面:

(1) 综合考虑相关性和多样性检索需求,提出了两阶段数据源选择策略.首先是相关性判别阶段,基于数据源摘要中用户查询所对应的主题内容的相关性估算得分与相关性偏差概率分布模型,确定候选数据源与用户查询的相关性估算得分;然后是数据源选择阶段,综合考虑数据源与用户查询的相关性和数据源的多样性,选择相关性较大、多样性较好(即综合性能达到最优)的 Top-K 个数据源;

(2) 在相关性判别阶段,为每个层次化主题构建了相关性偏差概率分布模型,并给出了不同情况下概率模型的使用策略,为提升模型的代表性设计了基于用户反馈的模拟查询选择方法;

(3) 在数据源选择阶段,首先基于层次主题的数据源摘要建立了多样性链接有向边,并提出了基于文档特征面的多样性判别方法;然后,将基于相关性和多样性的数据源选择问题转化为一个组合优化问题,并提出了基于优化函数的数据源选择策略.

实验结果表明:本文方法能较好地满足用户基于相关性和多样性的数据源选择需求,具有较大的应用价值.

2 数据源抽样与摘要构建

为构建深网数据源摘要,首先需要获取用于抽样的某领域数据源对应的主题词.数据源中文档内容的主题由主题词表征,本文把用于数据源文档主题分类的词汇称为主题词.通过观察发现,主题词具有 3 个特征:(1) 指向主题词的不同词语的数量越多,主题性越强;(2) 主题词出现在标题、摘要、句首、句尾的可能性较非主题词

大;(3) 主题词在领域文档集中词频较高.

TextRank 源于 PageRank 思想,采用投票机制对文档重要成分进行排序,不需要事先训练文档.基于主题词的 3 个特征,结合 TextRank 词汇关联思想,可以将主题词抽取问题转换成某领域数据源对应文档集合词语重要性排序问题.

相同领域浅网[1]和深网中,内容涉及的主题一般是一致的,所以本文基于同领域浅网以及 TextRank 技术获取用于同领域深网抽样的主题词.首先,基于 TextRank 获取一个数据源中每篇文档的候选主题词;然后,依据每个候选主题词累计得分计算其重要性.在以上策略中,如何获取每篇文档的候选主题词,成为一个关键问题.

基于 TextRank 思想,首先把文本 T 分割成句子集合;然后,对于每个句子进行分词与词性标注,仅保留名词、动词作为候选主题词;构建候选主题词关联图 $G=(V,E)$,其中,V 为节点集合,E 为节点构建的边集合.如果两个候选主题词包含在同一个句子中,则两个词之间存在一条连接边.

对于任意一个节点 v_i,$In(v_i)$为指向节点 v_i 的节点集合,$Out(v_i)$为节点 v_i 所指向的节点集合.令 w_{ij} 为 v_i 指向 v_j 边的权重,通过以下公式计算 v_i 的得分[18]:

$$Score(v_i)=(1-h)+h\times\sum_{v_j\in In(v_i)}\frac{w_{ji}}{\sum_{v_k\in Out(v_j)}w_{jk}}Score(v_j) \tag{1}$$

其中,h 为阻尼系数,一般取值为 0.85.对于节点 v_i,v_j,基于有向边$\langle v_i,v_j\rangle$计算 v_i 对 v_j 的影响,边的权重 w_{ij} 表征了 v_j 从 v_i 部分获得的分值,即,边的权重值 w_{ij} 代表了转移概率.

基于主题词的 3 个特征可计算边的权重值 w_{ij},具体方法如下.

(1) 假设 v_i 的覆盖影响力可以被均匀地传递到相邻节点,利用以下公式计算 v_i 到 v_j 的权重 $w_f(v_i,v_j)$:

$$w_f(v_i,v_j)=\frac{1}{|Out(v_i)|} \tag{2}$$

(2) 令 $w_z(v_i,v_j)$为节点 v_i 重要性影响力传递到 v_j 的权重,计算公式如下:

$$w_z(v_i,v_j)=\frac{Z(v_j)}{\sum_{v_k\in Out(v_i)}Z(v_k)} \tag{3}$$

其中,$Z(v_j)$为 v_j 重要性得分.主题词中多词词串比例较高,且主题词在标题、摘要、段落首句、结尾句中出现概率更高.综合考虑位置、长度等几个影响主题得分的因素,设计了以下词语主题重要性计分公式:

$$Z(v)=len(v)\times\left(1+\alpha\times\frac{N(v)}{M(v)}\right) \tag{4}$$

其中,$N(v)$为词 v 在文档中标题、摘要、段落句首、结尾句中出现的次数,$M(v)$为词 v 在文档中出现的总的次数,α 为调节参数.

(3) 高频词可以从邻接点获得更高的影响力权重,v_i 频度影响力传递到 v_j 的权重计算公式如下:

$$w_p(v_i,v_j)=\frac{N(v_j)}{\sum_{v_k\in Out(v_i)}N(v_k)} \tag{5}$$

其中,$N(v_j)$为节点 v_j 在文档集合中出现的次数,$N(v_k)$为相邻节点 v_k 在文档集中出现的次数.获取以上数据后,可以构建词语影响力转移矩阵,实现迭代收敛获取各候选主题词得分[19].

主题词是有层次的,通过以上算法可以直接获取第 1 层主题词及其对应的第 2 层的候选主题词.第 1 层主题词只有一个,自动选取重要性得分最高的词作为第 1 层主题词,余下的词作为第 2 层的候选主题词.从第 2 层开始,可以采用主题词作为用户查询获取该领域浅网数据源中相关文档,基于这些文档,再次使用本文方法获取子主题词,从而构建某领域下各深网基于层次主题的数据源摘要,如图 1 所示.以上过程可以自动完成,不需要人工参与.

图 1 中,层次主题词构成了一棵抽样树,可以基于抽样树中的主题词组合构建查询,用于构建深网数据源摘要,具体步骤是:(1) 假定抽样树有 N 层,根节点处于第 1 层,把根节点到 N–1 层节点路径上的主题词进行连接构建查询,提交给各深网数据源;(2) 各叶子节点按照摘要中文档数量的限制,仅保留检索返回的最相关的若干篇

文档(称为抽样文档).本文中文档与查询的相关性得分的计算,参考了文献[6]的方法.

另外,为了提升数据源相关性评价的准确性和检索内容的多样性,各数据源对应叶子主题建立了相关性偏差概率模型和多样性链接有向边,如图 1 所示(为简化图形,每个数据源仅绘制了一个概率模型和一条链接边),具体构建方法在第 4 节、第 5 节中进行了详细的介绍.每一条多样性链接有向边中记录了一个数据源相对于另一个相关数据源而言在一个对应叶子主题上的多样性得分,它会影响候选数据源被选的次序.这是因为在数据源选择过程中,当某些数据源被选后,在继续选择数据源的时候不仅需要考虑候选数据源与用户查询的相关性,还要考虑候选数据源提供数据的多样性.因此,多样性链接有向边的作用就是:在进行数据源选择时,相对于已选数据源而言,基于多样性链接有向边可以估算每一个候选数据源提供信息的多样性.例如,当数据源 A 已被选定,再选另外一个新数据源 B 时,需要计算 B 相对于 A 提供的数据多样性;反之,需要计算 A 相对于 B 提供的数据多样性.以上两个多样性得分值是不同的,这是建立有向边的意义所在.

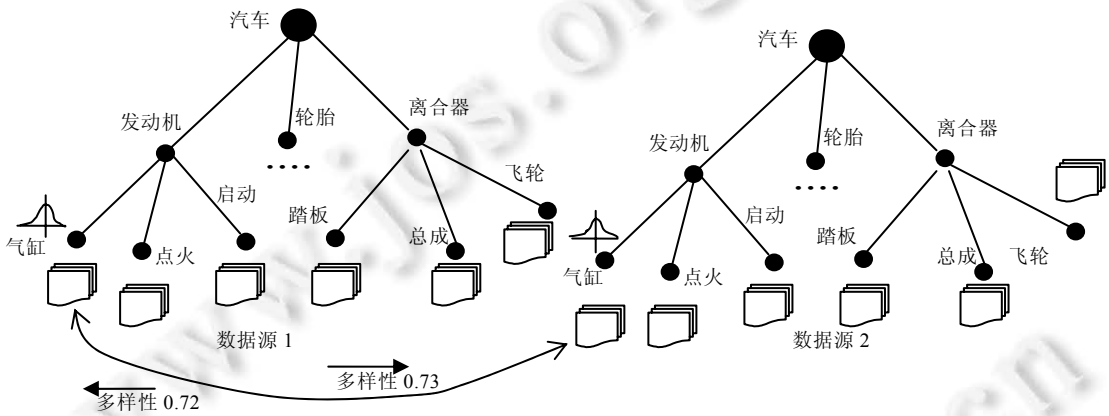


Fig.1 Data source summary graph base on hierarchical subjects

图 1 基于层次主题的数据源摘要图

3 相关性判别

数据源摘要中,叶子主题下的抽样文档具有一定的代表性.首先,使用 SUSHI 算法[6]依据各数据源摘要中查询 q 所对应的叶子主题下的抽样文档的排名情况,度量各数据源与查询 q 的相关性.由于摘要中叶子主题下的抽样文档不能完全反映真实数据源的内容,为进一步提升数据源选择的准确性,我们为各数据源拥有的每一个叶子主题构建一个相关性偏差概率分布模型.

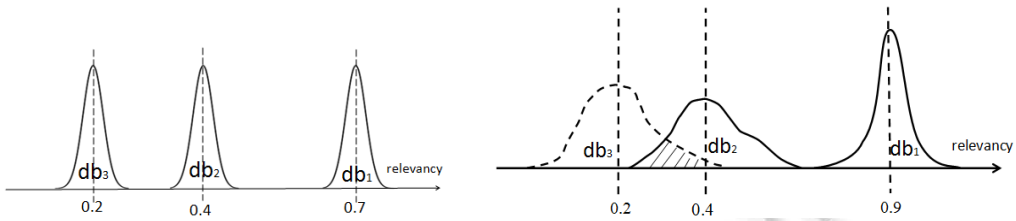
文献[8]使用多种抽样文档排序算法(含 SUSHI 算法)综合计算数据源相关性,而本文仅使用其中的 SUSHI 算法,原因在于:(1) 不同相关性判别方法所采用的证据不同,在此基础上难以建立有价值的相关性偏差概率分布模型;(2) SUSHI 算法在多数测试集中的表现优于文献[8]中使用的其他方法.

3.1 基于相关性偏差概率分布模型的数据源选择

记 $r'(\overline{S}_i,q)$ 为基于数据源 S_i 的摘要 $\overline{S}_i$ 中查询 q 所对应的叶子主题下的抽样文档所计算得到的数据源 S_i 与查询 q 的相关性估算得分,$r(S_i,q)$为真实数据源 S_i 与查询 q 的相关性得分.

要获取 $r(S_i,q)$,可以向数据源 S_i 直接提交查询 q 获取相应的检索结果,但代价太大.对于查询 q, $r'(\overline{S}_i,q)$ 通常与 $r(S_i,q)$有一定偏差.如果能基于离线获取数据源 S_i 的摘要 $\overline{S}_i$ 中查询 q 所对应的叶子主题内容(即叶子主题下的抽样文档)相对于真实数据源 S_i 的相关性偏差概率分布模型,则可以有效提升数据源选择的准确性.

相关性偏差概率分布模型的构建将在第 3.2 节中进行详细说明.现在分析在数据源摘要中查询 q 所对应的叶子主题下已经存在偏差概率分布模型的情况下,如何进行数据源选择的两种情况,如图 2 所示.



(a) 理想情况 (b) 一般情况

Fig.2 Data source selection analysis base on correlation deviation probability distribution model

图 2 基于相关性偏差概率分布模型的数据源选择分析

如图 2(a)所示,假设有 3 个数据源 db1,db2 和 db3,各数据源摘要中,查询 q 所对应的叶子主题下抽样文档的相关性得分分别为 0.7,0.4 和 0.2,此时,查询 q 与各数据源的相关性得分显然有 db1>db2>db3.

如图 2(b)所示,各数据源摘要中查询 q 所对应的叶子主题下抽样文档的相关性得分分别为 0.9,0.4 和 0.2,显然,查询 q 与 db1 的相关性得分最高,但查询 q 与 db2,db3 的相关性排序难以准确判别.每个相关性偏差概率分布图的总面积为 1.0.db3,db2 的相关性偏差概率分布图有重叠,记重叠(即阴影部分)的面积为 *area*.*area* 代表两个数据源与查询 q 的相关性得分难以准确判别的模糊区域;*area* 越小,则 db2 与查询 q 的相关性得分大于 db3 与查询 q 的相关性得分的概率越大.

记两个数据源摘要中查询 q 所对应的叶子主题下的相关性偏差概率分布图的重叠面积为 *area*,如果 *area* 的值超过某个阈值 $\sigma(0<\sigma<1)$,则认为通过这两个数据源提供的叶子主题下的相关性偏差概率分布模型,无法判断哪个数据源与查询 q 的相关性更大.

此时,可以进一步考虑两个数据源摘要中同主题下能够提供的相关文档总数量以及同主题下其他相关查询对应的两数据源排名情况.选择以上两个参数的原因在于:(1) 通常情况下,一个数据源摘要中同主题下能提供的相关文档数量越多,则该数据源提供更多的相关性结果的概率越大;(2) 一个数据源在同主题下其他相关查询能获取的相关性得分越高,则该数据源与给定查询相关性越高的概率越大.

基于此,对于按与查询 q 的相关性估算得分降序排列的数据源 S_i^P,设计了以下调整相邻数据源相关性先后次序的判别公式:

$$prior(S_{i+1}^P,S_i^P)=\frac{num(S_{i+1}^P,q)}{0.5\times num(S_i^P,q)+0.5\times num(S_{i+1}^P,q)}+\frac{front(S_{i+1}^P,S_i^P)-back(S_{i+1}^P,S_i^P)}{T} \tag{6}$$

其中,$prior(S_{i+1}^P,S_i^P)$ 为基于数据源摘要计算的调整数据源 S_i,S_{i+1} 相关性先后次序的判别得分,$num(S_i^P,q)$,$num(S_{i+1}^P,q)$ 分别为对应数据源摘要中在查询 q 所对应的叶子主题下能提供的抽样文档的数量,$front(S_{i+1}^P,S_i^P)$,$back(S_{i+1}^P,S_i^P)$ 分别为 T 次相关查询中基于概率分布图所确定的 S_{i+1}^P 位于 S_i^P 之前、之后的次数.

当 $prior(S_{i+1}^P,S_i^P)$ 的值超过某个阈值 ψ 时,则认为数据源 S_{i+1}^P 应该位于数据源 S_i^P 之前,即,数据源 S_{i+1}^P 与查询 q 的相关性应该大于数据源 S_i^P 与查询 q 的相关性.基于实验探测法,ψ 的值设为 1.2.

综上所述,基于公式(6)对相关性估算得分降序排列的两个相邻数据源 S_i^P,S_{i+1}^P 与查询 q 的相关性估算得分 $r'(\overline{S}_i^P,q),r'(\overline{S}_{i+1}^P,q)$ 进行调整,调整的方法为:当 $prior(S_{i+1}^P,S_i^P)\geqslant\psi$,则交换 $r'(\overline{S}_i^P,q)$ 与 $r'(\overline{S}_{i+1}^P,q)$ 的值.

说明:为简化公式(6)的计算,T 次相关查询从构建叶子主题相关性概率分布模型的模拟查询(见第 3.2 节)中选取,依据模拟查询与用户查询的相似度[20]从高到低选择;对于每个模拟查询 q 和任意两个数据源 S_i,S_j,由于基于概率分布图能确定 S_j 位于 S_i 之前、之后的情况,因此可以在构建数据源摘要的过程中用三元组 $q(S_i,S_j,v)$ 事先保存这些判别结果,其中,v=1 表示 S_j 位于 S_i 之前,v=0 表示 S_j 位于 S_i 之后.

3.2 构建相关性偏差概率分布模型

针对数据源 S_i,为数据源摘要 $\overline{S}_i$ 中的每个叶子主题人工构建 F 个查询 q(称为模拟查询),通过分别向数据源

摘要 $\overline{S}_i$、真实数据源 S_i 提交相同的 F 个模拟查询,计算 $r'(\overline{S}_i,q)$、$r(S_i,q)$,并根据 $r(S_i,q)$和 $r'(\overline{S}_i,q)$ 的差值构建数据源摘要 $\overline{S}_i$ 中该叶子主题的相关性偏差概率分布模型,如图 2 所示.

由于真实查询 q 在事先构建数据源摘要时难以获取,因此,本文采用模拟的方式获取较为真实的用户查询.通常情况下,一个查询如果可以较快且被较多用户联想到,那么这个查询在真实环境下被提交的概率也越大.实验中,研究人员模拟用户向真实数据源对应主题提交少量查询,具体策略如下.

(1) 构建常用查询空集合 *Freq-Q* 和非常用查询空集合 *inFreq-Q*;

(2) 如果人工提交的查询在 *Freq-Q* 和 *inFreq-Q* 中均未出现,则将该查询添加到 *inFreq-Q* 中;如果查询仅出现在 *inFreq-Q* 中,则在 *inFreq-Q* 中移除该查询,并将该查询添加到 *Freq-Q* 中,并置其出现次数为 2;

(3) 如果人工提交的查询在 *Freq-Q* 中出现,则其出现次数加 1;

(4) 如果 *inFreq-Q* 中查询已满,删除第 1 个查询,添加新查询.原因在于:第 1 个查询最长时间未遇到相同查询,则表示其为常用查询概率较小.

由于 $P(r(S_i,q)\leqslant\alpha\,|\,r'(\overline{S}_i,q)=\beta)$ 是基于 $r'(\overline{S}_i,q)$ 的条件概率分布,因此,不同的β值对应不同形状的偏差分布模型.为了能够得到准确的偏差模型,需要针对每个β值提交一组抽样查询.由于β是无限的,为了减少构建偏差分布模型的代价,假设相关性概率分布与 $r'(\overline{S}_i,q)$ 是不相关的,即,假设 $(r(S_i,q)-r'(\overline{S}_i,q))/r'(\overline{S}_i,q)$ 与 $r'(\overline{S}_i,q)$ 不存在依赖关系. $(r(S_i,q)-r'(\overline{S}_i,q))/r'(\overline{S}_i,q)$ 比 $r(S_i,q)-r'(\overline{S}_i,q)$ 更能反映该主题下内容对应查询的偏差强度.因此,我们可以通过提交 F 个模拟查询获取 $(r(S_i,q)-r'(\overline{S}_i,q))/r'(\overline{S}_i,q)$ 值的单一分布情况(此值不依赖于具体查询),即,有如下公式成立:

$$P(r(S_i,q)\leqslant\alpha\,|\,r'(\overline{S}_i,q)=\beta)=P\left(\frac{r(S_i,q)-r'(\overline{S}_i,q)}{r'(\overline{S}_i,q)}\leqslant\frac{\alpha-\beta}{\beta}\middle|\,r'(\overline{S}_i,q)=\beta\right)=P\left(\frac{r(S_i,q)-r'(\overline{S}_i,q)}{r'(\overline{S}_i,q)}\leqslant\frac{\alpha-\beta}{\beta}\right) \tag{7}$$

4 多样性计算

在选择 Top-K 数据源的过程中,可能存在如下情况:有些数据源与用户查询的相关性较高,但是与其他数据源提供的数据存在较大的相似性.如果已选定某些数据源,再选择一个查询相关性得分较高但多样性信息贡献较少的数据源,显然不是数据源选择的最好结果.为解决以上问题,我们事先离线在各相关数据源摘要的每一个对应叶子主题之间建立多样性关联(即,多样性链接有向边),在进行数据源选择时,便可以估算每一个候选数据源相对于已选数据源而言提供信息的多样性.在摘要构建过程中,基于叶子主题组织数据源抽样文档,各叶子主题下的抽样文档对应不同的特征面.一个叶子主题下抽样文档对应的特征面越多,则越有可能满足用户多样性的检索需求.特征面通过特征词得以反映,同样,利用本文改进的 TextRank 方法获取数据源摘要各叶子主题下抽样文档对应的特征词,并依据不同数据源摘要中相同主题下抽样文档所包含的特征词的多样性来判别数据源返回检索结果的多样性.

假定数据源 S_i 已被选,相对于数据源摘要 $\overline{S}_i$ 中叶子主题 c 下的抽样文档而言,计算数据源摘要 $\overline{S}_j$ 中相同主题下的抽样文档的多样性,计算步骤如下.

(1) 基于本文改进的 TextRank 方法,分别获取数据源摘要 $\overline{S}_i$,$\overline{S}_j$ 中主题 c 下的每篇抽样文档中的特征词,分别添加到集合 H_i 和 H_j 中.

(2) 计算 H_i 中的第 m 个特征词 $h_{i,m}$ 与 H_j 中的第 n 个特征词 $h_{j,n}$ 之间的关联度 $a(h_{i,m},h_{j,n})$,计算公式如下:

$$a(h_{i,m},h_{j,n})=\frac{\sum_{k=1}^{D}[\min(f(h_{i,m},d_k),f(h_{j,n},d_k))/\max(f(h_{i,m},d_k),f(h_{j,n},d_k))]}{D} \tag{8}$$

其中,D 为数据源摘要 $\overline{S}_i$ 和 $\overline{S}_j$ 中主题 c 下含有 $h_{i,m}$ 或 $h_{j,n}$ 的抽样文档总数量;$f(h_{i,m},d_k)$和 $f(h_{j,n},d_k)$分别为特征词 $h_{i,m}$,$h_{j,n}$ 在抽样文档 d_k 中的词频,如果 $f(h_{i,m},d_k)$=$f(h_{j,n},d_k)$=0,则不参与计算;min,max 分别获取 $f(h_{i,m},d_k)$,$f(h_{j,n},d_k)$中的最小值和最大值.

(3) 计算 H_j 中的第 n 个特征词 $h_{j,n}$ 与数据源摘要 $\overline{S}_i$ 的关联度 $score\_x(h_{j,n},\overline{S}_i)$,见公式(9):

$$score\_x(h_{j,n},\overline{S}_i)=\xi\times\max\{a(h_{j,n},h_{i,1}),...,a(h_{j,n},h_{i,m}),...,a(h_{j,n},h_{i,|H_i|})\} \tag{9}$$

其中,$h_{i,m}$ 为 H_i 中的第 m 个特征词,ξ为调节参数.如果 $h_{j,n}\in H_i$ 或者 $h_{j,n}$ 可以在 H_i 中找到同义词或近义词,则:

$$score\_x(h_{j,n},\overline{S}_i)=\xi\times 1=\xi\ .$$

(4) 计算主题 c 下数据源摘要 $\overline{S}_j$ 中的抽样文档相对于数据源摘要 $\overline{S}_i$ 中的抽样文档的多样性估算得分 $div(\overline{S}_i,\overline{S}_j,c)$,见公式(10):

$$div(\overline{S}_i,\overline{S}_j,c)=\sum_{n=1}^{|H_j|}\gamma_{h_{j,n},H_i}(1-score\_x(h_{j,n},\overline{S}_i)) \tag{10}$$

其中,$score\_x(h_{j,n},\overline{S}_i)$ 表示特征词 $h_{j,n}$ 与数据源摘要 $\overline{S}_i$ 的关联度;相应地,$(1-score\_x(h_{j,n},\overline{S}_i))$ 则可理解为特征词 $h_{j,n}$ 与数据源摘要 $\overline{S}_i$ 的差异度,即多样性.由于关联度 $score\_x(h_{j,n},\overline{S}_i)$ 是取 $h_{j,n}$ 与 H_i 中某个特征词的关联度的最大值(见公式(9),假设 $h_{j,n}$ 与 H_i 中的特征词 $h_{i,m}$ 的关联度最大),因此,在计算多样性估算得分 $div(\overline{S}_i,\overline{S}_j,c)$ 时,还需要考虑将 H_i 中与 $h_{j,n}$ 关联度最大的特征词 $h_{i,m}$ 的得分 $Score(h_{i,m})$作为权重,因为特征词得分不同,则其反映文档特征面的作用也会不同,特征词得分的计算见公式(1).即:$\gamma_{h_{j,n},H_i}$ 是将 $h_{j,n}$ 与 H_i 中关联度最大的特征词 $h_{i,m}$ 的得分 $Score(h_{i,m})$作为权重,并进行归一化处理得到的权重值.

综上所述,可以在主题 c 上建立一条由数据源摘要 $\overline{S}_j$ 指向数据源摘要 $\overline{S}_i$ 的多样性链接有向边 $l_{j,i}$,边的权值为 $div(\overline{S}_i,\overline{S}_j,c)$,它反映了数据源 S_j 相对于数据源 S_i 的多样性价值,即:如果数据源 S_i 先被选入 Top-K 数据源,可以基于边的权值估算下一个候选数据源 S_j 的多样性价值.同理,在主题 c 上还可以建立从数据源摘要 $\overline{S}_i$ 指向数据源摘要 $\overline{S}_j$ 的多样性链接有向边 $l_{i,j}$,边的权值为 $div(\overline{S}_j,\overline{S}_i,c)$.

5 数据源选择

在数据源集成检索过程中,用户关心的首先是选取与用户查询最相关的若干个数据源,其次是希望所选取的数据源具有多样性.鉴于以上原因,采用两阶段法选择 Top-K 个数据源:首先,基于数据源摘要计算各候选数据源与用户查询的相关性估算得分;然后,综合考虑数据源与用户查询的相关性以及数据源的多样性选择 K 个数据源.具体如下.

1) 基于数据源摘要计算各候选数据源与用户查询的相关性估算得分

首先,基于数据源摘要 $\overline{S}_i$ 中查询 q 所对应的叶子主题下的抽样文档,计算数据源 S_i 与查询 q 的相关性估算得分 $r'(\overline{S}_i,q)$;然后,基于数据源摘要 $\overline{S}_i$ 中的相关性偏差概率分布模型,对排名相邻的数据源与查询 q 的相关性估算得分进行调整.

2) 基于数据源与用户查询的相关性和数据源的多样性选择数据源

假设给定数据源集合 $DB=\{S_1,S_2,\ldots,S_{|DB|}\}$($|DB|$表示数据源集合 DB 中包含的数据源个数),每一个数据源相当于一个文档集合,记 $S_i=\{d_{i,1},d_{i,2},...,d_{i,|S_i|}\}$ ($|S_i|$表示数据源 S_i 中包含的文档个数),则数据源选择的任务是:从数据源 DB 中选择 K 个数据源 $S_{\max}^K=\{S_{i_1},S_{i_2},...,S_{i_K}\}$ ($K<<|DB|$,$1\leqslant i_1,i_2,...,i_K\leqslant|DB|$,且 $i_1,i_2,...,i_K$ 互不相同),使 $S_{\max}^K$ 与查询 q 的相关性较大且 $S_{\max}^K$ 的多样性较好的综合性能达到最优.

因此,基于数据源与用户查询的相关性和数据源的多样性选择数据源,是一个组合优化问题,本文采用遗传算法 $Genetic_algorithm(DB,K)$获取数据源选择的最优结果,其步骤如下.

(1) 将问题表述成位串形式,随机产生初始群体;

(2) 优化目标为:基于数据源集合 DB 选择 K 个数据源 $S_{\max}^K$,使 $S_{\max}^K$ 与查询 q 的相关性较大且 $S_{\max}^K$ 的多样性较好的综合性能达到最优,并基于该优化目标设计适应度函数;

(3) 基于计算得到的适应度反复对群体进行选择、交叉组合和变异操作,直至得到一个趋于收敛状态的

最优解.

本文采用 MATLAB 2014b 遗传算法工具箱获取 *Genetic_algorithm*(*DB*,*K*)计算结果.

首先,根据第(1)步,基于数据源摘要已经计算了各候选数据源 S_k 与用户查询 q 的相关性估算得分 $r'(\overline{S}_k,q)$;其次,在数据源摘要中,查询 q 所对应的叶子主题 c 下已经建立由数据源摘要 $\overline{S}_j$ 指向数据源摘要 $\overline{S}_i$ 的多样性链接有向边 $l_{j,i}$,边的权值为 $div(\overline{S}_i,\overline{S}_j,c)$,它反映了数据源 S_j 相对于数据源 S_i 的多样性价值.为了简化计算,将数据源选择的优化目标转化为:基于数据源集合 *DB* 选择 *K* 个数据源 $S_{\max}^K$,使如下适应度函数取最大值:

$$Fitness=\left(\sum_{k=1}^{K}r'(\overline{S}_k,q)\right)\times\frac{\sum_{1\leqslant i\neq j\leqslant K}div(\overline{S}_i,\overline{S}_j,c)}{P_K^2} \tag{11}$$

其中,*Fitness* 为个体的适应值; P_K^2 为计算排列,表示该个体中 *K* 个数据源之间的多样性链接有向边的数量.

在数据源选择时,需要将用户查询 q 映射到数据源摘要中的某个对应叶子主题 c,具体的映射方法见文献[21].数据源选择算法见算法 1.

算法 1. 二阶段数据源选择算法.

输入:候选数据源集合 $DB=\{S_1,S_2,\ldots,S_n\}$,$n=|DB|$为候选数据源数量;用户查询 q,返回数据源数量为 *K*;

输出: $S_{\max}^K$.

1. $DB^P\leftarrow\varnothing,DB^D\leftarrow\varnothing$;
2. 计算各数据源 S_i 的 $r'(\overline{S}_i,q)$,按相关性估算得分降序放入 DB^P 中;
3. **for** i=1~n−1 **do** //基于概率分布图调整数据源的相关性估算得分
4. **if** ($area(S_i^P,S_{i+1}^P,q,c)\geqslant\sigma$ **and** $prior(S_{i+1}^P,S_i^P)>\psi$)
5. $S_i^P\leftrightarrow S_{i+1}^P, r'(S_i^P,q)\xleftrightarrow{\geqslant} r'(S_{i+1}^P,q)$; // S_i^P 与 S_{i+1}^P 互换,并按降序调整相关性估算得分
6. **end if**
7. $DB^D=DB^D\cup S_i^P$; //把 DB^P 中第 i 个元素放入 DB^D 中
8. **end for**
9. $DB^D=DB^D\cup S_n^P$; //把 DB^P 中最后一个元素放入 DB^D 中
10. $S_{\max}^K$ =*Genetic_algorithm*(DB^D,*K*); //遗传算法从 DB^D 中选择 *K* 个数据源 $S_{\max}^K$,使适应度函数取最大值
11. return $S_{\max}^K$;

6 实验结果分析

深网数据源选择领域一般采用人为构建的文档集作为测试数据集,鉴于人工构建文档集与真实深网数据内容的差别,为展示数据源选择方法的实际效用,选取了汽车深网与图书深网两个领域的数据进行评测.33 个带有查询接口的真实的商业化汽车深网(如凤凰汽车、易车网、爱卡汽车、太平洋汽车、汽车之家、网易汽车、车讯网、网上车市、腾讯汽车、和讯网、环球汽车网、新浪汽车、搜狐汽车、汽车之友、汽车点评网、汽车口碑网、第一车网、中国汽车消费网、新车评网、车问网、购车网、越野 e 族、无敌汽车网、天涯汽车、汽车在线、万户论坛、中华网汽车、58 车、央广网汽车、车主之家、爱意汽车、车神榜、网通社等深网)作为汽车深网测试数据源.33 个带有查询接口的真实的商业化图书深网(如中国图书网、亚马逊、当当网、互动出版网、文轩网、博库网、中国图书网、云中书城、读览天下、蔚蓝网、蜘蛛网、99 网上书城、淘宝图书、北发图书网、京东商城图书、广州购书中心网上书店、三联韬奋书店、蓝泉图书、中国互动出版网、金书网、社会科学文献网、网上书店、华储网、孔夫子书网、布衣书局、新世界书库、图书网、晨星网路书店、包年优品、天猫书城、人教商城、有路网、广购书城等深网)作为图书深网测试数据源.主题词获取时需用到分词程序,实验采用的是中国科学院开发的 ICTCLAS2015 中文分词系统[22].

在信息检索领域,一般是通过查准率(即准确率)和查全率(即召回率)来评价一个检索系统的性能,其中,查

准率是指检出的相关文献与检出文献总数的比值,而查全率是指检出的相关文献与相关文献总数的比值.但是对于数据源选择而言,其准确率的内涵不再是所选择的 K 个数据源集合 $S_{\max}^{K}$ 中包含的相关数据源数量与 K 的比值,这是因为一般情况下所选择的 K 个数据源都是跟用户需求相关的;其准确率的内涵应该是所选择的 K 个数据源集合 $S_{\max}^{K}$ 中出现在数据源相关性降序排列的前 K 个数据源集合中的数据源数量与 K 的比值;同时,并不关心召回率(从某种意义上说,此时的准确率就是基于数据源相关性降序排列的前 K 个最相关数据源集合的召回率).为了更好地评价数据源选择的性能,不仅考虑数据源选择的准确率,还进一步考虑数据源选择的文档准确率,它们的定义分别如下:

定义 1(基于相关性的数据源选择准确率). 所选择的 K 个数据源集合 $S_{\max}^{K}$ 中,出现在候选数据源集合 DB 按相关性降序排列的前 K 个数据源集合中的数据源数量与 K 的比值(要求 $K<<|DB|$),称为基于相关性的数据源选择准确率,即,基于相关性的数据源选择 Top-K 准确率.

定义 2(基于相关性和多样性的数据源选择准确率). 记从候选数据源集合 DB 中任意选择 K 个数据源形成的集合为 DB^{K},如果 DB^{K} 中去除文档重复度大于某个阈值ρ的文档后与用户查询的相关性取最大,则称 DB^{K} 为相关性和多样性综合性能最优的 K 个数据源集合,记为 $DB_{\max}^{K}$.所选择的 K 个数据源集合 $S_{\max}^{K}$ 中,出现在相关性和多样性综合性能最优的 K 个数据源集合 $DB_{\max}^{K}$ 中的数据源数量与 K 的比值(要求 $K<<|DB|$),称为基于相关性和多样性的数据源选择的准确率,即,基于相关性和多样性的数据源选择 Top-K 准确率.

定义 3(基于相关性的数据源选择的文档准确率). 记候选数据源集合 DB 中所包含的相关文档集合为 DOC,DOC 中的文档按相关性降序排列的前$|DOC|\times K/|DB|$个文档所构成的集合称为 Top-K 数据源选择最相关文档集合,记为 $DOC_{\max}^{K}$.所选择的 K 个数据源集合 $S_{\max}^{K}$ 所包含的相关文档集合中,出现在 Top-K 数据源选择最相关文档集合 $DOC_{\max}^{K}$ 中的文档数量与 $|DOC_{\max}^{K}|$ 的比值,称为基于相关性的数据源选择的文档准确率,简称为基于相关性的文档选择准确率.

定义 4(基于相关性和多样性的数据源选择的文档准确率). 记候选数据源集合 DB 中去除文档重复度大于某个阈值ρ的文档后所包含的相关文档集合为 DOC,DOC 中的文档按相关性降序排列的前$|DOC|\times K/|DB|$个文档所构成的集合称为 Top-K 数据源选择按相关性和多样性综合性能最优的文档集合,记为 $DOC_{\max}^{K}$.所选择的 K 个数据源集合 $S_{\max}^{K}$ 去除文档重复度大于某个阈值ρ的文档后所包含的相关文档集合中,出现在 Top-K 数据源选择按相关性和多样性综合性能最优的文档集合 $DOC_{\max}^{K}$ 中的文档数量与 $|DOC_{\max}^{K}|$ 的比值,称为基于相关性和多样性的数据源选择的文档准确率,简称为基于相关性和多样性的文档选择准确率.

在构建数据源摘要时,保留抽样获取的文档数量越多,则摘要质量越高,但必然降低数据源选择的效率.本文研究着眼于基于少量抽样文档进行数据源选择,因而在实验过程中,针对一个数据源分别选择了 3 000,4 000,5 000 篇抽样文档构建数据源摘要,数据源选择的结果取以上 3 种数据源摘要下的平均值.

在构建基于层次主题的数据源摘要时,基于领域知识以及计算量考虑,每个非叶节点的主题词对应的子主题词数量设置为 15 个.实验中,依据指定抽样文档数量设定摘要中每个叶节点需要保留的抽样文档的相关性阈值.在构建常用查询集合时,让 10 位研究生分别针对两个领域中每个叶子主题提出 30 个模拟查询(假设同领域中各数据源摘要主题词完全相同),基于第 3.2 节提出的常用查询集合构建方法获取常用查询,并依据出现次数降序排列各查询.在离线构建每个叶子主题对应的相关性偏差概率模型时,取出现次数位于前 30 的常用查询获取的相关数据(模拟查询数量对数据源选择的影响将在后面进行分析).

为了评测本文数据源选择方法的有效性,针对每个领域让 5 位大一本科生提交了 150 个查询,每个查询包含 2~4 个查询关键词,包含 2 个、3 个、4 个关键词的查询均为 50 个,数据源选择准确率和文档选择准确率取所有查询结果的平均值.依据实验探测法,在进行相关性判别时参数 T 值设为 7,在叶子主题内容多样性计算时仅保留每篇抽样文档中得分较高的 7 个特征词用于表征该抽样文档特征面,参数ξ设为 0.7.

本文方法先计算数据源与用户查询的相关性,然后进一步考虑提供信息的多样性.因此,实验中先对比本文提出的相关性判别方法与已有方法进行深网选择的效果,然后展示本文提出的基于相关性和多样性相结合的

方法进行数据源选择的效果,最后分析重要参数对本文方法的影响.

6.1 基于相关性的数据源选择

由于本文针对的是非合作环境下非结构化深网数据源选择问题,因此选取了文献[8]提出的 HYBRID(混合)方法和文献[10]提出的 TP(主题模型)方法作为对比方法,评价本文提出的数据源相关性判别方法(PM)的效果.选取以上两个对比方法的原因在于:(1) 文献[8]所提出的方法准确性较高,且在不同数据集下有稳定的表现;(2) 文献[10]可以基于小抽样样本自动挖掘主题,采用 LDA 模型描述主题内容,算法较新且在某些数据集上数据源选择准确率较高.另外,为观察主题内容相关性偏差概率模型所起的作用,在实验中还展示了 PM 方法中去掉主题内容相关性偏差概率模型的 SSUSHI 方法(即,仅使用本文的抽样主题摘要结合 SUSHI 算法)进行数据源选择的效果.以上数据源选择方法仅以检索结果相关性为目标,因此在评测过程仅考虑相关性选择最相关数据源.实验结果如图 3、图 4 所示.

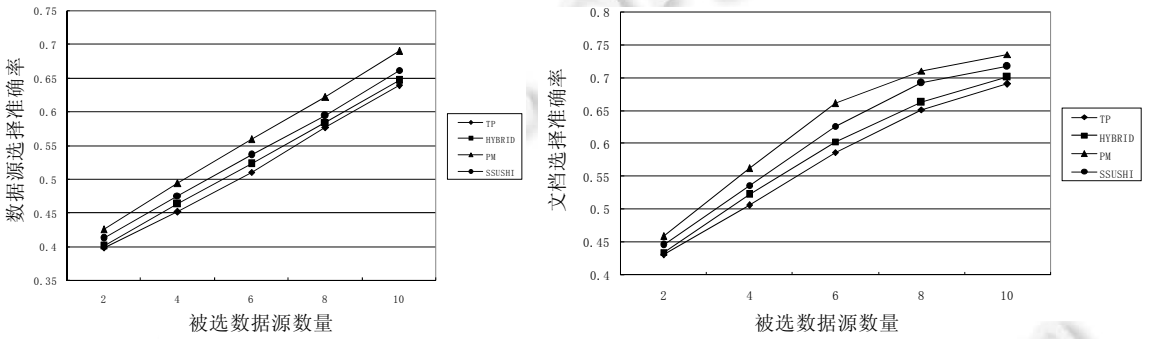


Fig.3 Comparison of different data source selection methods based on correlation in the field of automobile

图 3 汽车领域下基于相关性的不同数据源选择方法的效果比较

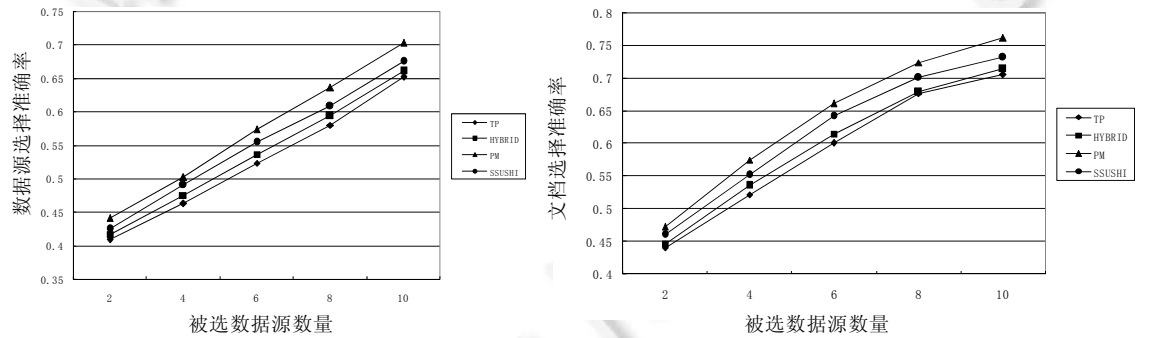


Fig.4 Comparison of different data source selection methods based on correlation in the field of books

图 4 图书领域下基于相关性的不同数据源选择方法的效果比较

从图 3、图 4 可以看出:对于数据源相关性判别的准确率,PM 方法较 HYBRID,TP 两种方法有明显优势.当选择 Top-2 数据源时,PM 较 HYBRID,TP 数据源选择准确率提升 2.5 个百分点以上;当选择 Top-10 数据源时,数据源选择准确率提升 4.3 个百分点以上.原因是:我们在抽样文档的基础上引入了额外的信息,即主题相关性偏差概率模型,并考虑了主题内数据的关联特性.PM 方法中,去掉主题内容相关性偏差概率模型进行数据源选择时,数据源选择准确率随被选数据源数量增加与 PM 方法差距拉大,当选择 Top-10 数据源时,数据源选择准确率下降了 2.8 个百分点以上.出现以上情况可能的原因在于:被选数据源数量增多时,排名靠后的数据源所提供的数据质量差别减小,导致 SSUSHI 难以准确判别数据源排序.

PM 方法在文档选择准确率上同样优于 HYBRIDmTP 对比方法.当选择 Top-2 数据源时,文档选择准确率高过 HYBRIDmTP 这两种方法 2.4 个百分点以上;当选择 Top-10 数据源时,文档选择准确率高过 HYBRID,TP 这

两种方法 3.5 个百分点以上.PM 方法中,去掉主题内容相关性偏差概率模型进行数据源选择时,当选择 Top-10 数据源时,文档选择准确率下降了 1.9 个百分点以上.PM 在文档准确率上的优势不如数据源选择准确率明显,可能的原因在于:排序前后接近的有些数据源提供的检索结果数量与得分相差不大.另外还可以发现,图书领域下各数据源选择方法的效果在一定程度上优于汽车领域.原因可能在于,图书领域中的文本信息编辑更为规范.

6.2 基于相关性和多样性的数据源选择

在定义 2、定义 4 中,“去除文档重复度大于某个阈值ρ的文档”的内涵是:基于文献[23]中的方法发现给定文档集中相似度大于等于阈值ρ=0.75 的所有文档子集,且每一个文档子集仅保留与用户查询相关性得分最高的一篇文档.

文献[16]考虑了非合作环境下基于相关性和多样性进行数据源选择的问题,但是其主要面向 P2P 领域的特殊存储数据,难以与本文方法进行直接对比.文献[17]提出了基于簇的多样性数据源选择方法,但未同时考虑查询相关性.为了更有说服力,分别把文献[8]的 TP 方法、文献[10]的 HYBRID 方法和文献[17]的 CLUSTER 方法结合起来,与本文同时考虑相关性与多样性的数据源选择方法(OUR METHOD)进行综合比较;同时,为了评判多样性的作用,还对比了只考虑相关性(PM 方法)的数据源选择效果.实验结果如图 5、图 6 所示.

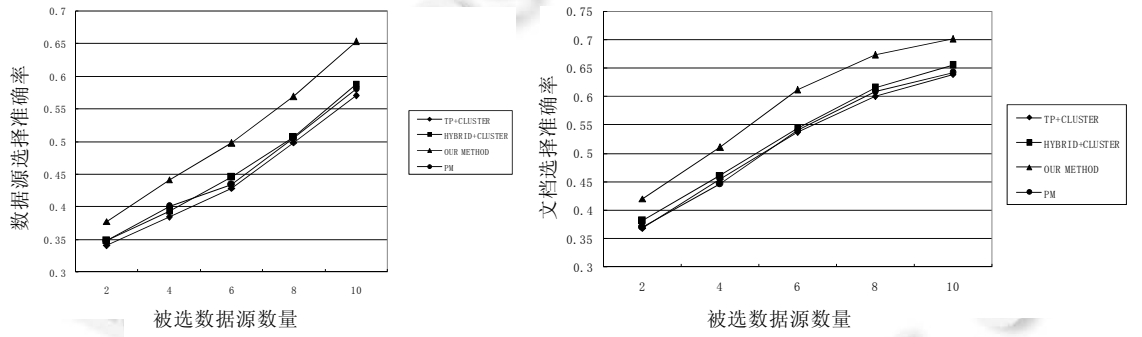


Fig.5 Comparison of different data source selection methods in the field of automobile

图 5 汽车领域下不同数据源选择方法的效果比较

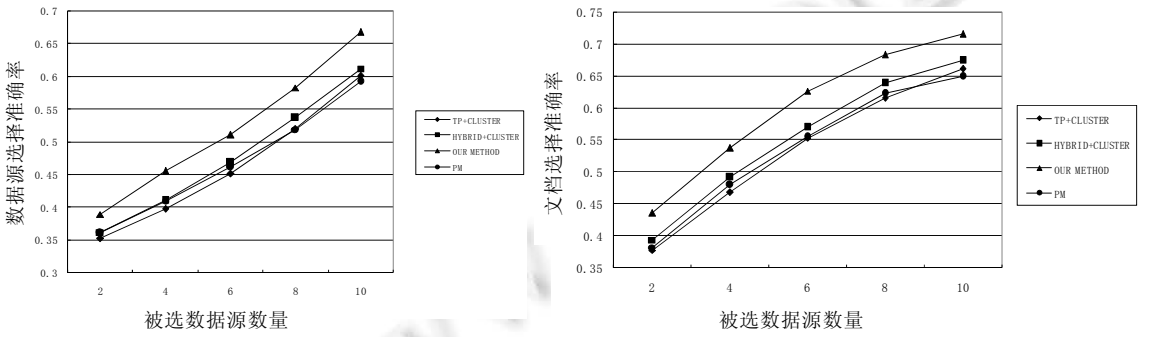


Fig.6 Comparison of different data source selection methods in the field of books

图 6 图书领域下不同数据源选择方法的效果比较

从图 5、图 6 可以看出:在两个领域中,各数据源选择方法的效果均较图 4、图 5 中有所下降.原因在于,它们选择最佳数据源的标准不同:图 4、图 5 仅考虑相关性进行数据源选择,而图 5、图 6 需要综合考虑相关性和多样性进行数据源选择,难度较大.

从图 5 和图 6 可以看出:当被选数据源数量增多时,各方法对应的数据源选择准确率都是上升的.原因在于:被选 Top-K 数据源数量越多,其严格排序的要求被降低.在两个领域中 OUR METHOD 方法较 TP+CLUSTER 和 HYBRID+CLUSTER 方法有明显优势,数据源选择准确率超过对比方法 3 个百分点以上,文档选择准确率高于

对比方法 3.9 个百分点以上;且当被选数据源数量增多时,优势有所加强.原因在于:一是采用了基于层次主题的数据源摘要和主题相关性偏差概率模型;二是综合考虑了相关性与多样性,并使用基于优化函数的数据源选择算法.如果只基于本文的相关性(PM 方法)进行数据源选择,数据源选择准确率下降 2.9 个百分点以上,文档选择准确率下降 5.0 个百分点以上.这充分说明了基于相关性基础上综合考虑多样性对数据源选择的重要性.

6.3 抽样技术对数据源选择的影响

已有的基于少量抽样文档进行数据源选择的方法大多采用 RS-Ord,RS-Lrd 抽样技术.实验中,把本文的抽样方法(OUR METHOD)分别用 RS-ORD,RS-LRD 抽样方法进行了替换,因此可以观察不同抽样算法带来的影响.鉴于不同领域下抽样技术对数据源选择的影响趋势是大体相同的,因此仅列出汽车领域的相关评测结果,如图 7 所示.从图 7 可以看出:若采用 RS-ORD 或 RS-LRD 抽样方法,本文提出的数据源选择策略的准确率会有所降低.原因在于:两种对比抽样方法均为随机抽样,抽样数据主题代表性不强.另外,对比图 5 和图 7 可以发现:尽管其他抽样策略会导致数据源选择的准确率下降,但是仍然略优于对比的数据源选择方法.原因在于:尽管随机抽样方法会导致文档主题代表性下降,但通过相关性偏差概率模型、基于优化函数综合考虑相关性与多样性的数据源选择策略等因素,会抵消随机抽样方法导致抽样数据主题代表性不强的影响.同样,本文抽样方法对应的文档选择准确率也优于其他抽样方法,并随着被选数据源数量的增加优势更为明显.

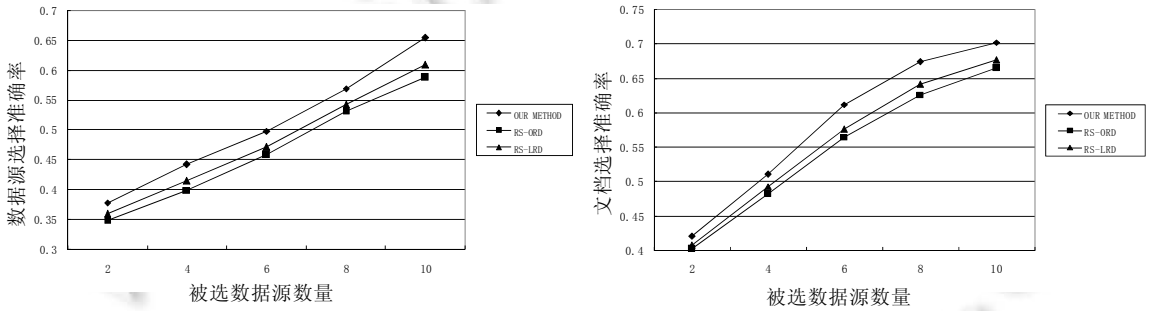


Fig.7 Comparison of the data source selection method under different sampling strategy in the field of automobile

图 7 不同抽样策略下汽车领域数据源选择的效果比较

6.4 模拟查询数量对数据源选择的影响

在构建一个数据源摘要中某主题下内容对应于用户查询的相关性偏差概率模型时,分别采用了 20,30,40,50 个模拟查询,观测其对数据源选择的影响.鉴于不同领域下模拟查询数量对数据源选择的影响趋势是大体相同的,因此仅列出汽车领域的相关评测结果,如图 8 所示.

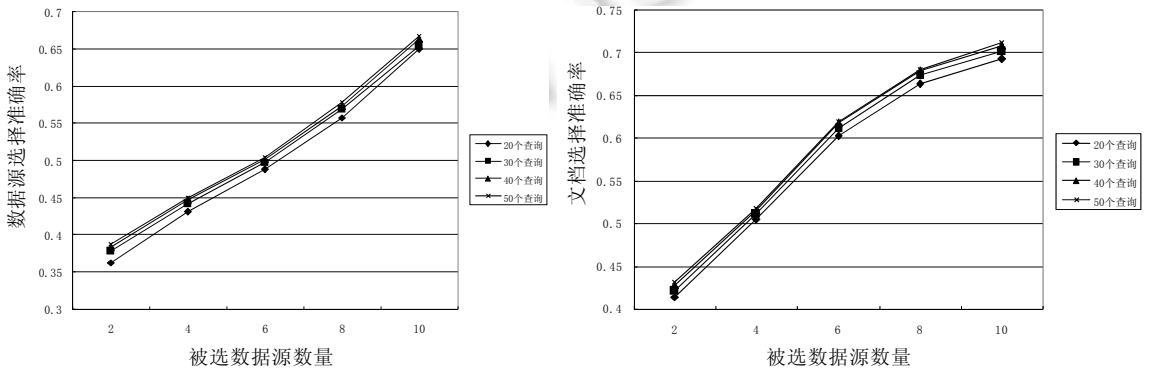


Fig.8 Effect of number of simulated queries on data source selection

图 8 模拟查询数量对数据源选择的影响

从图 8 可以看出:模拟查询数量为 20 的时候,其数据源选择准确率与文档选择准确率明显低于模拟查询数量在 30 以上的时候.以上原因可能在于:(1) 模拟查询越多,其包含用户提交相似查询的可能性越大;(2) 模拟查询越多,相关性偏差概率模型越为准确.模拟查询数量分别为 30,40,50 时,数据源选择准确率差距缩小.原因在于:当模拟查询数量到达一定程度后,查询数量对数据源选择准确率的影响性减弱.综合考虑模拟查询数量增加所带来的效用与代价,实验中,模拟查询数量取值为 30.

7 总结与展望

为解决既考虑相关性又考虑多样性的数据源选择问题,提出了一种基于主题与概率模型的非结构化、非合作深网数据源选择方法.为增强小规模抽样文档的主题代表性,采用 TextRank 算法获取层次化的抽样主题词,用于构建基于层次主题的深网数据源摘要;为提升数据源选择的相关性判别的准确性,在数据源摘要中引入了相关性偏差概率模型,给出了基于数据源摘要中叶子主题下的抽样文档与概率分析的数据源与用户查询相关性估算策略;为提升数据源选择结果的多样性程度,在数据源摘要中建立了多样性链接有向边,边的权值反映了数据源的多样性价值,并给出了多样性权值的估算方法.

接下来,将基于相关性和多样性的数据源选择问题转化为一个组合优化问题,优化目标为:基于数据源集合 DB 选择 K 个数据源 $S_{\max}^{K}$,使 $S_{\max}^{K}$ 与查询 q 的相关性较大且 $S_{\max}^{K}$ 的多样性较好的综合性能达到最优.最后,基于优化目标设计了适应度函数,提出了基于优化函数的数据源选择策略.

实验结果表明:本文方法在数据源选择准确率和文档选择准确率上都较已有方法有较大优势,可以较好地满足基于相关性和多样性的数据源选择需求.在未来的工作中,将进一步研究数据源选择过程中的检索结果语义关联问题,以更好地满足用户的检索需求.

References:

[1] Ipeirotis PG, Gravano L. Classification-aware hidden-Web text database selection. ACM Trans. on Information Systems (TOIS), 2008,26(2):1−66. [doi: 10.1145/1344411.1344412]

[2] Crestani F, Markov I. Distributed information retrieval and applications. In: Proc. of the European Conf. on Advances in Information Retrieval. Heidelberg: Springer-Verlag, 2013. 865−868. [doi: 10.1007/978-3-642-36973-5_104]

[3] Thomas P. To what problem is distributed information retrieval the solution? Journal of the American Society for Information Science and Technology, 2012,63(7):1471−1476. [doi: 10.1002/asi.22684]

[4] D'Souza D, Zobel J, Thom J. Is CORI effective for collection selection? An exploration of parameters, queries, and data. In: Proc. of the 9th Australasian Document Computing Symp. Melbourne: ADCS, 2004. 41−46.

[5] Milad S. Central-Rank-Based collection selection in uncooperative distributed information retrieval. In: Proc. of the 29th European Conf. on IR Research. Heidelberg: Springer-Verlag, 2007. 160−172. [doi: 10.1007/978-3-540-71496-5_17]

[6] Thomas P, Shokouhi M. SUSHI: Scoring scaled samples for server selection. In: Proc. of the 32nd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2009). New York: ACM Press, 2009. 419−426. [doi: 10.1145/1571941.1572014]

[7] Markov I, Crestani F. Theoretical, qualitative, and quantitative analyses of small-document approaches to resource selection. ACM Trans. on Information Systems (TOIS), 2014,32(2):1−37. [doi: 10.1145/2590975]

[8] Markov I, Azzopardi L, Crestani F. Reducing the uncertainty in resource selection. In: Proc. of the 35th European Conf. on IR Research (ECIR 2013). Heidelberg: Springer-Verlag, 2013. 507−519. [doi: 10.1007/978-3-642-36973-5_43]

[9] Hong D, Si L, Bracke P, Witt M, Juchcinski T. A joint probabilistic classification model for resource selection. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2010). New York: ACM Press, 2010. 98−105. [doi: 10.1145/1835449.1835468]

[10] Wang QY, Cao W, Shi SC. Deep Web resource selection using topic models. Journal of Computer Applications, 2015,35(9): 2553−2559, 2595 (in Chinese with English abstract). [doi: 10.11772/j.issn.1001-9081.2015.09.2553]

[11] Cetintas S, Si L, Yuan H. Learning from past queries for resource selection. In: Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM 2009). New York: ACM Press, 2009. 1867−1870. [doi: 10.1145/1645953.1646251]

[12] Gutiérrez-Soto C, Hubert G. Probabilistic reuse of past search results. In: Proc. of the Database and Expert Systems Applications. Heidelberg: Springer-Verlag, 2014. 265−274. [doi: 10.1007/978-3-319-10073-9_21]

[13] Fan J, Zhou LZ. Keyword-Based deep Web database selection. Chinese Journal of Computer, 2011,34(10):1797−1804 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2011.01797]

[14] Dong XL, Saha B, Srivastava D. Less is more: Selecting sources wisely for integration. In: Proc. of the 39th Int'l Conf. on Very Large Data Bases (VLDB 2013). San Francisco: Morgan Kaufmann Publishers, 2013. 37−48. [doi: 10.14778/2535568.2448938]

[15] Rekatsinas T, Dong XL. Characterizing and selecting fresh data sources. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD 2014). New York: ACM Press, 2014. 919−930. [doi: 10.1145/2588555.2610504]

[16] Bender M, Michel S, Triantafillou P, Weikum G, Zimmer C. Improving collection selection with overlap awareness in P2P search engines. In: Proc. of the 28th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR 2005). New York: ACM Press, 2005. 15−19. [doi: 10.1145/1076034.1076049]

[17] Rekatsinas T, Dong XL. Finding quality in quantity: The challenge of discovering valuable sources for integration. In: Proc. of the 7th Biennial Conf. on Innovative Data Systems Research (CIDR 2015). New York: ACM Press, 2015. 1−7.

[18] Mihalcea R, Tarau P. TextRank: Bringing order into texts. In: Proc. of the Empirical Methods in Natural Language Processing. Barcelona: ACL, 2004. 404−411.

[19] Li P, Wang B, Shi ZW, Cui YC, Li HX. Tag-TextRank: A webpage keyword extraction method based on tags. Journal of Computer Research and Development, 2012,49(11):2344−2352 (in Chinese with English abstract).

[20] Chen GL, He L, Hu QM, Yang J. Improve dialogue short text clustering by fusion form and semantic similarity. Journal of Chinese Computer Systems, 2015,36(9):1963−1967 (in Chinese with English abstract).

[21] Zhang Y, Song W, Liu T, Li S. Query classification based on URL topic. Journal of Computer Research and Development, 2012, 49(6):1298−1305 (in Chinese with English abstract).

[22] Du LP, Li XG, Yu G, Liu CL, Liu R. New word detection based on an improved PMI algorithm for enhancing segmentation system. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016,52(1):35−40 (in Chinese with English abstract). [doi: 10.13209/j.0479-8023.2016.024]

[23] Huang CH, Yin J, Hou F. A text similarity measurement combining word semantic information with TF-IDF method. Chinese Journal of Computers, 2011,34(5):856−864 (in Chinese with English abstract).

附中文参考文献:

[10] 王秋月,曹巍,史少晨.基于主题模型的深层网数据源选择算法.计算机应用,2015,35(9):2553−2559, 2595. [doi: 10.11772/j.issn.1001-9081.2015.09.2553]

[13] 范举,周立柱.基于关键词的深度万维网数据库的选择.计算机学报,2011,34(10):1797−1804. [doi: 10.3724/SP.J.1016.2011.01797]

[19] 李鹏,王斌,石志伟,崔雅超,李恒训.Tag-TextRank:一种基于 Tag 的网页关键词抽取方法.计算机研究与发展,2012,49(11):2344−2352.

[20] 陈国梁,贺樑,胡琴敏,杨静.融合形态和语义相似度的对话短文本聚类.小型微型计算机系统,2015,36(9):1963−1967.

[21] 张宇,宋巍,刘挺,李生.基于 URL 主题的查询方法分类.计算机研究与发展,2012,49(6):1298−1305.

[22] 杜丽萍,李晓戈,于根,刘春丽,刘睿.基于互信息改进算法的新词发现对中文分词系统改进.北京大学学报:自然科学版,2016,52(1):35−40. [doi: 10.13209/j.0479-8023.2016.024]

[23] 黄承慧,印鉴,侯昉.一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法.计算机学报,2011,34(5):856−864.

邓松(1982－),男,江西南昌人,博士,讲师,CCF 专业会员,主要研究领域为 Web 数据管理,情感分析,虚假舆情识别,大数据分析.

万常选(1962－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为 Web 数据管理,情感分析,信息检索,数据挖掘.