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.

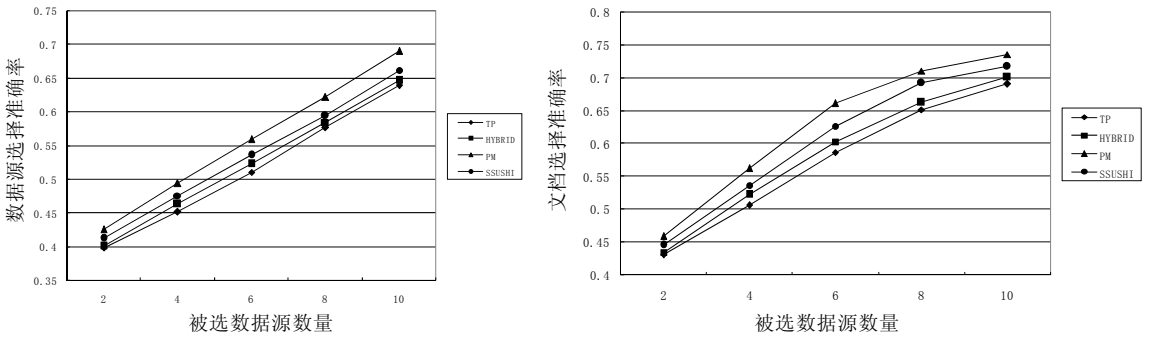

Fig.3 Comparison of different data source selection methods based on correlation in the field of automobile

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

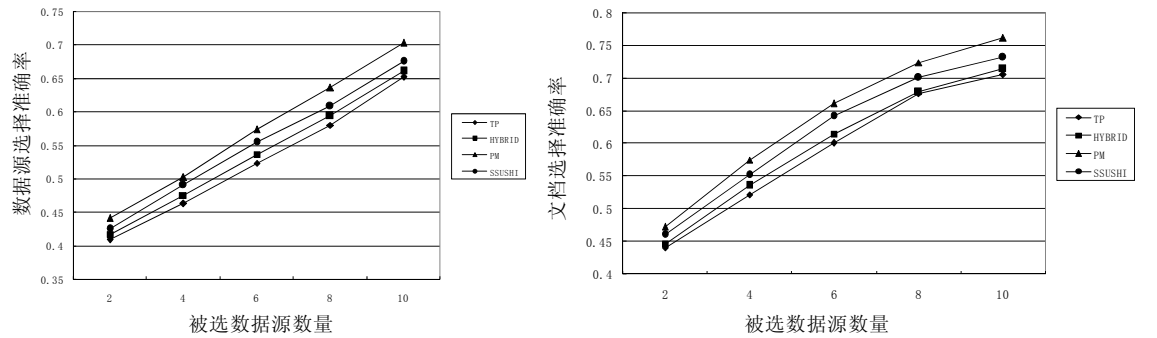

Fig.4 Comparison of different data source selection methods based on correlation in the field of books

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.

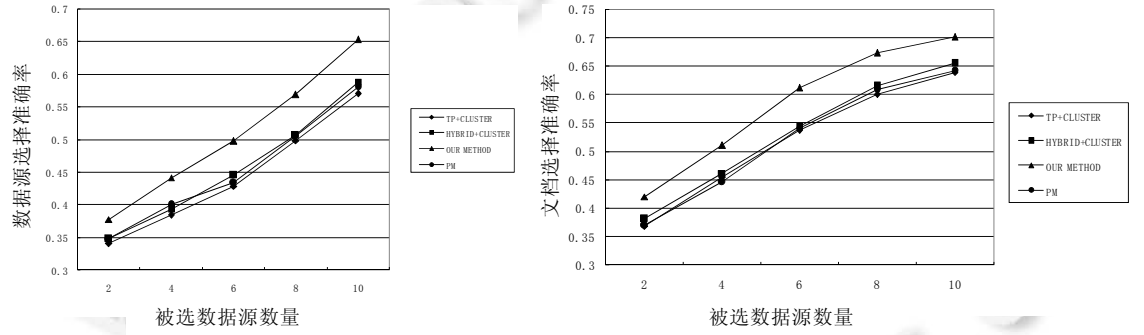

Fig.5 Comparison of different data source selection methods in the field of automobile

图 5 汽车领域下不同数据源选择方法的效果比较

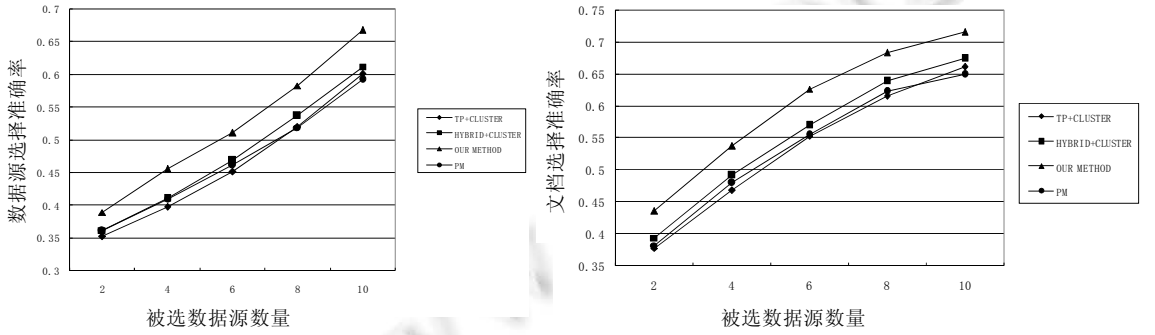

Fig.6 Comparison of different data source selection methods in the field of books

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.

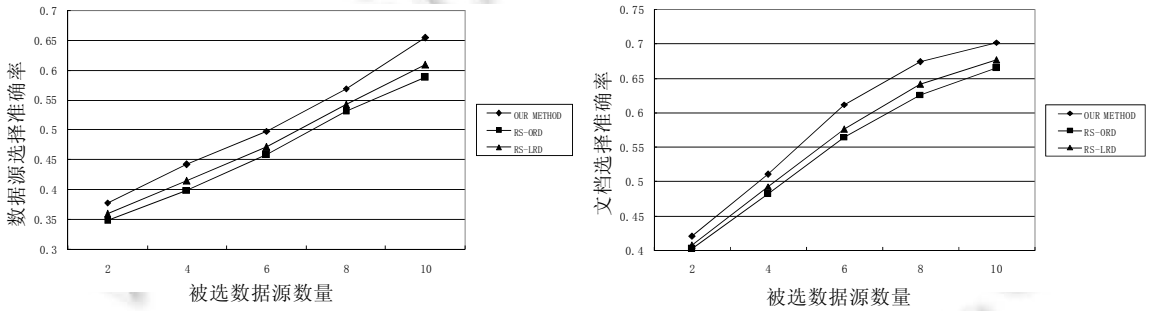

Fig.7 Comparison of the data source selection method under different sampling strategy in the field of automobile

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.

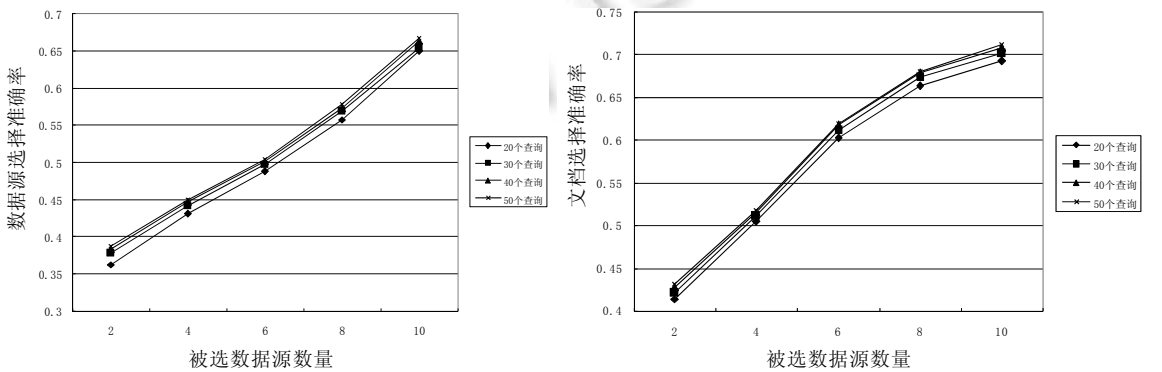

Fig.8 Effect of number of simulated queries on data source selection

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.