基于潜在语义分析的Deep Web查询接口聚类研究

强保华[1,2] 李 巍[1] 邹显春[2] 汪天天[1] 吴春明[2]
(桂林电子科技大学计算机科学与工程学院 广西 541000)[1]
(西南大学计算机与信息科学学院 重庆 400715)[2]

摘 要 集成查询接口的生成是Deep Web数据集成的重要组成环节。如何对不同领域的查询接口进行有效的聚类是生成集成查询接口时需要解决的核心问题之一。针对传统的向量空间模型在Deep Web查询接口聚类时单纯依赖关键词匹配的缺点,引入潜在语义分析(LSA)的方法来发掘查询接口之间的语义关系,并给出了基于潜在语义分析的Deep Web查询接口聚类算法,最后采用UIUC的Web集成资源库提供的数据进行了实验。结果表明,潜在语义分析的方法提高了同一领域查询接口之间的相似度,明显改善了Deep Web查询接口聚类的质量。

关键词 潜在语义分析,奇异值分解,Deep Web,查询接口聚类
中图法分类号 TP391.3 **文献标识码** A

Research on Deep Web Query Interface Clustering Based on Latent Semantic Analysis

QIANG Bao-hua[1,2] LI Wei[1] ZOU Xian-chun[2] WANG Tian-tian[1] WU Chun-ming[2]
(School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541000, China)[1]
(College of Computer and Information Science, Southwest University, Chongqing 400715, China)[2]

Abstract Generation of integrated query interfaces is the important issue of Deep Web data integration. How to cluster different query interfaces effectively is one of the most core issues when generating integrated query interface. Due to the traditional vector space model can't solve the shortage of relying on keyword maching in the Deep Web query interface clustering, the Latent Semantic Analysis (LSA) method was introduced and then the algorithm of Deep Web query interface clustering based on Latent Semantic Analysis was proposed. The experimental results on UIUC Web integration repository show that LSA method can significantly improve the performance of Deep Web query interface clustering.

Keywords Latent semantic analysis, Singular value decomposition, Deep Web, Query interface clustering

1 引言

随着Web技术的日益成熟和Web服务趋于多样化,Deep Web所蕴含的信息也在飞速增长,同时越来越多的用户选择通过访问Web数据库来获取信息,在此背景下,Deep Web相关研究也受到人们越来越多的关注。对Deep Web研究的根本目的是为用户提供一个统一的访问途径来自动地获取和利用自由分布在整个Web上隐藏于Deep Web中的丰富信息[1]。刘伟、孟小峰等人在文献[1]中提出了Deep Web数据集成框架,并将该框架分为3个主要模块:查询接口集成模块、查询处理模块和查询结果处理模块。而目前国内外的研究主要集中在查询接口集成模块这部分,查询接口集成分为4个步骤:查询接口的发现、查询接口模式抽取、查询接口的分类(聚类)和生成统一的查询接口。因此,如何对不同领域的查询接口进行有效的基于领域的聚类是生成集成查询接口时需要解决的核心问题之一。

Deep Web查询接口是以HTML页面呈现的,但是不同的网页往往是由不同的人或群体开发的,每个人对同一事物的认知也是千差万别,导致网页的形式和内容有着巨大的差异,而同一领域Deep Web查询接口之间的差异性最突出地表现在"一词多义"和"多词一意"的这一现象上。图1和图2所示为两个航班的查询接口(部分截图)。

对于出发地和目的地这两个属性,不同的查询接口使用了不同的语汇,那么在缺乏外部语料库的前提下,单纯地使用基于文本关键字匹配的向量空间模型(VSM),很难将语义上属于同一领域的查询接口聚为一类。

潜在语义分析(LSA)是Scott Deerwester, Susan T. Dumais等人在1990年提出来的一种新的索引和检索方法。其基本思想是假设文档之间存在着某种联系,然后从词语之间的相关性出发,通过分析大量的文本中词语的使用关系,提取

到稿日期:2013-01-13 返修日期:2013-04-22 本文受国家自然科学基金(61163057),广西自然科学基金(2012jjAAG0063),广西可信软件重点实验室开放基金(KX201117),广西研究生科研创新项目(YCSZ2012070)资助。
强保华(1972—),男,博士,教授,主要研究方向为Web数据管理、智能搜索,E-mail:qiangbh@guet.edu.cn;李 巍(1988—),男,硕士生,主要研究方向为Deep Web信息搜索及数据集成;邹显春(1965—),男,硕士,副教授,主要研究方向为网络应用与Web智能;汪天天(1987—),女,硕士生,主要研究方向为网络搜索;吴春明(1972—),男,博士,副教授,主要研究方向为数据库与智能检索技术。

出隐含的语义信息，有效地获取和表示了词汇的语义知识[2]。本文将 LSA 的方法应用于 Deep Web 查询接口之间潜在语义的分析，通过构造“站点-属性”潜在语义空间，在一定程度上消除了同义词和多义词的影响，提高了同一领域查询接口之间的相似度，最后结合适当的聚类算法对查询接口进行聚类，为解决 Deep Web 查询接口基于领域的聚类问题提供了新思路。

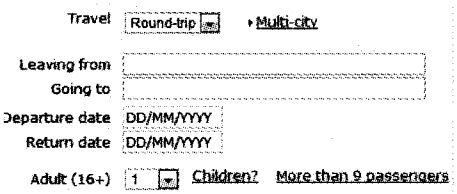


图 1　air Canada

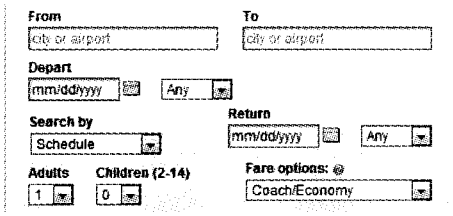


图 2　US airways

2　潜在语义分析基本原理

潜在语义分析(LSA)，也被称为潜在语义索引(LSI)，该方法的核心思想是通过奇异值分解及地址逼近，将高维的“文档-词项”矩阵投影到较低维的潜在语义空间中，达到消除噪声数据、发掘潜在语义关系的目的[3]。

2.1　奇异值分解

奇异值分解(SVD)是 LSA 的数学基础。设 C 为 $m\times n$ 的“词项-文档”矩阵，则存在奇异值分解：

$$C=U\Sigma V^T \tag{1}$$

式中，U 为 $m\times n$ 的正交矩阵，列为 CC^T 的正交特征向量。V 为 $m\times n$ 的正交矩阵，列为 CC^T 的正交特征向量。Σ 为 $m\times n$ 的矩阵，对于 Σ 需要特别说明：

(1) $C_{m\times n}^T C_{m\times n}$ 的特征值是 $\lambda_1,\lambda_2,\cdots,\lambda_r$；

(2) $\forall i\in[1,r]$，有 $\delta_i=\sqrt{\lambda_i}$，且 $\lambda_i\geqslant\lambda_{i+1}$，$\Sigma_{m\times n}$ 满足 $\Sigma_{ii}=\delta_i$，而矩阵的其他位置均为 0。Σ_{ii} 也叫做 $C_{m\times n}$ 的奇异值。

2.2　低秩逼近

在 LSA 中，去除“噪声”数据实质是通过矩阵的低秩逼近来实现的[4]。首先给出低秩逼近的定义：设 C 为 $m\times n$ 的矩阵，其秩为 r，C_k 是一个秩为 k 的 $m\times n$ 矩阵，并且有 $r\geqslant k$。令 $X=C-C_K$，如果 X 的 F 范数最小，即式(2)最小：

$$\|X\|_F=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}X_{ij}^2} \tag{2}$$

当 k 比 r 小得多时，我们称 C_k 为 C 的低秩逼近矩阵。SVD 可以解决矩阵低秩逼近的问题，即保留 $\Sigma_{m\times n}$ 中最大的 k 个奇异值，将其他的 $r-k$ 个奇异值置为 0，得到 Σ_k，计算 $C_k=U\Sigma_k V^T$，用 C_k 作为 C 的逼近。

定理 1　$\min\limits_{Z|rank(Z)=k}\|C-Z\|_F=\|C-C_k\|_F=\sqrt{\sum\limits_{i=k+1}^{r}\delta_i^2}$

可以证明上述过程产生了一个秩为 k 的矩阵 C_k，它的范数误差最小。

2.3　LSA 应用示例

选取航空领域的部分属性。假设有 4 个查询接口属性集，如表 1 所列。

表 1

A1	"from","Arrival","Departure","Date","Date","Return"
A2	"Depart","Destination","Departure","Date","Date","Return"
A3	"going","to","Departing","on","Return","on","return","on"
A4	"from","to","leave","on","Return","on"

根据 VSM 得到“查询接口-属性”矩阵 C，如表 2 所列。

表 2

	A1	A2	A3	A4
to	0.0	0.0	1.0	1.0
going	0.0	0.0	1.0	0.0
Departure	1.0	1.0	1.0	0.0
return	0.0	0.0	1.0	1.0
from	1.0	0.0	0.0	0.0
Depart	0.0	1.0	0.0	0.0
leave	0.0	0.0	0.0	1.0
Date	2.0	2.0	0.0	0.0
Return	1.0	1.0	1.0	1.0
Arrival	1.0	0.0	0.0	0.0
on	0.0	0.0	3.0	2.0
Departing	0.0	0.0	1.0	0.0
Destination	0.0	0.0	1.0	0.0

然后对“查询接口-属性”矩阵 C 进行奇异值分解，得到奇异值：4.5203，3.6328，1.6731，1.2528。令 $K=2$，即选取最大的两个奇异值：4.5203，3.6328。进行低秩逼近，得到新的“查询接口-属性”矩阵 C_k，如表 3 所列。

表 3

	A1	A2	A3	A4
to	0.03	−0.00	1.14	0.77
going	−0.01	−0.04	0.68	0.46
Departure	0.99	0.99	−0.06	0.09
return	−0.01	−0.04	0.68	0.46
from	0.55	0.53	0.44	0.37
Depart	0.49	0.49	−0.04	0.03
leave	0.05	0.03	0.46	0.31
Date	1.98	1.98	−0.13	0.19
Return	1.03	0.98	1.08	0.87
Arrival	0.49	0.49	−0.01	0.05
on	0.05	−0.05	2.98	2.02
Departing	−0.01	−0.04	0.68	0.46
Destination	0.49	0.49	−0.04	0.03

观察表 2 和表 3，从表 2 中可以看出，查询接口属性集 A2 中并没有“from”属性，但是表 3 中“A2-from”为 0.53，说明“from”与 A2 是有关系的。事实上，在航空领域的查询接口中，“from”和“Depart”都代表“出发”的意思，而 A2 中存在属性“Depart”，所以在表 3 中“A2-from”的值被提高了。

综上可知，LSA 通过奇异值分解，将高维的向量空间模型映射到低维的潜在语义空间中，在新的潜在语义空间中，凸显了词语文档之间的语义关系，提高了文本表示的准确性[5]。

3　基于 LSA 的 Deep Web 查询接口聚类研究

进行文本聚类分析时，一般选择向量空间模型(VSM)，同样，在 Deep Web 查询接口上使用潜在语义分析的基础也是 VSM。

假设已经获取 n 个 Deep Web 查询接口集，将它们表示

为 $F=\{f_1,f_2,\cdots,f_n\}$，作为矩阵的列索引；$A=\{a_1,a_2,\cdots,a_m\}$表示从所有的查询接口中提取的 m 个属性，作为矩阵的行索引。

这样可以得到一个 $m\times n$ 的初始的属性-查询接口矩阵 C：

$$C=(c_{ij})_{m\times n} \tag{3}$$

该矩阵的每一行代表一个唯一的属性，每一列代表一个唯一的查询接口。元素 c_{ij} 表示属性 a_i 在查询接口 f_j 中出现的次数。

上一步的处理，只考虑了属性在查询接口中是否出现。接下来为每个属性增加权值，本实验选择 TF-IDF 权重，其指导思想是：如果一个属性在一个查询接口中出现的次数很多，那么该属性在另一个同类查询接口中出现的次数也会很多，反之亦然[6]。权重计算公式如下：

$$w_{ij}=TF_{ij}\times\log(N/DF_i) \tag{4}$$

式中，TF_{ij} 表示属性 a_i 在查询接口 f_j 中出现的次数，N 表示总的查询接口数目，DF_i 表示属性 A_i 在 N 个查询接口中出现的总次数。

接下来对增加权值的属性-查询接口矩阵 C 进行 SVD 的分解，得到式(1)的形式。然后选取合适的 k 得到低秩逼近矩阵 C_k。

最后我们选取合适的聚类算法。其中 K-均值和 K-中心是最著名、最常用的两种划分方法。但是存在噪声和离群点时 K-中心点算法更具“鲁棒性”，因为中心点不像均值那样容易受离群点或其他极端值的影响。所以论文中采用 PAM 算法(K-中心点的一种流行实现)。当计算聚类对象之间距离时，一般都采用欧式距离，但是考虑到不同领域的查询接口属性个数存在差异，我们选择余弦相似度，计算公式如下：

$$\mathrm{sim}(i_1,i_2)=\frac{\vec{V}(i_1)\cdot\vec{V}(i_2)}{|\vec{V}(i_1)||\vec{V}(i_2)|} \tag{5}$$

下面给出基于潜在语义分析的 Deep Web 查询接口聚类算法的步骤。

输入：查询接口集 $F=\{F_1,F_2,\cdots,F_n\}$，低秩逼近矩阵的空间维数 K，结果簇的数目 k。

输出：$FC=\{fc_1,fc_2,\cdots,fc_k\}$查询接口聚类结果集，簇的中心点集 $FO=\{fo_1,fo_2,\cdots,fo_k\}$。

```
1. A=extract(F);//从 F 中抽取属性,并预处理
2. C=init(A,F);//生成初始的属性-查询接口矩阵
3. C=SVD(C);
4. C_K=LSA(C);
5. FO←random(C_K,K);//随机选择 K 个中心点
6. repeat;
7. foreach f_i∈F and f_i∉FO do
8.    index←0; maxSiml←0;
9.    foreach f_i∈FO do
10.       if(siml(f_i,f_j)>maxSiml) then
11.       maxSiml←siml(f_i,f_j);
12.       index←j;
12.   end
13.   FC[index]←f_i
14. end
15. foreach f_i∈FO do
16.   index←0; maxTC←0;
17.   foreach f_j∈F and f_i∉FO do
18.      TC←totalCost(replace(f_j,f_i));
19.      if(TC>maxTC) then
20.         maxTC←TC; index←j;
21.   end
22.   if(maxTC>0) then
23.      substitude(f_index,f_i);
24. end
25. until notChange(FO);
26. return FO, FC;
```

4 实验和分析

4.1 实验数据

本实验的数据来自 UIUC 的 Web 集成资源库，该资源包含多领域的查询接口，并以 XML 文件的形式存储。实验选取 airfares，automobiles，books 和 musicRecords 这 4 个领域的 221 个查询接口。

4.2 实验数据的预处理

由于查询接口属性定义的不规则，不仅实验引入了过多的噪声，而且也使部分查询接口语义信息不足。针对本实验的特点需要对数据进行以下 3 点处理：

(1)去除停用词，但是保留具有领域含义的词，如航空领域中“to”(到达)，“from”(出发)等；

(2)词干还原和词形归一化，将属性单词的不同时态、单复数形态等还原成单词的词干；

(3)查询接口的语义扩充，对于每一个领域建立特征词表，如，对于航空领域，定义特征词表：$a_f=\{'flight','air'\ldots'arrive'\}$。如果某一查询接口中的属性出现在词表中，则将整个词表中的属性单词添加到该查询接口的属性集中。

4.3 低秩逼近矩阵维度 k 的选取

k 表示语义空间的维度，选取的 k 值过小则会使有用的语义信息丢失，反之如果 k 值选取太大，不仅达不到降维去除噪声信息的目的，而且还增大了计算的复杂度[8]。因此 k 值的大小直接影响到最终实验的效果。

目前常用的确定 k 值的方法是参考因子分析中的贡献率不等式法[9]。假设原始矩阵 C 经过 SVD 分解得到 r 个奇异值，选取合适的 ε，使得 $\|C-C_k\|_F=\sqrt{\lambda_{k+1}+\lambda_{k+2}+\cdots+\lambda_r}<\varepsilon$，则此时 C_k 和 C 达到最优近似。经计算，本实验 k 取值为 2。

4.4 实验结果评价指标和结果分析

由于 Deep Web 查询接口聚类结果为无序集合，我们选择查准率 P、查全率 R 和 F_1 值作为评价指标。令 tp(真正例)表示正确分配到相应簇中的情况；fp(伪正例)表示簇中错误分配的情况；fn(伪反例)表示本该分配到簇而没有分配到相应簇的情况。则有：

$$P=\frac{tp}{tp+fp},R=\frac{tp}{tp+fn},F_1=\frac{2PR}{P+R} \tag{6}$$

图 3 和图 4 示出分别为仅使用 VSM 和使用 LSA+VSM 后聚类的效果。

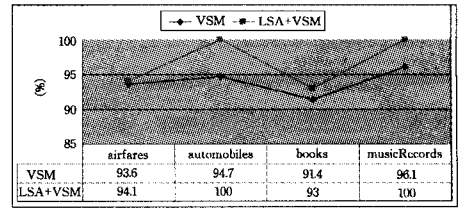


图 3 查准率 P 对比

(下转第 247 页)

tive query expansion[C]//Proceedings of the 14th ACM international conference on information and knowledge management. New York, NY, USA: ACM, 2005: 696-703

[4] Cao G, Nie J Y, Gao J, et al. Selecting good expansion terms for pseudo-relevance feedback[C]//Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM, 2008: 243-250

[5] Gao J, Nie J Y, Xun E, et al. Improving query translation for cross-language information retrieval using statistical models[C]// Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: ACM, 2001: 96-104

[6] Callan J P, Croft W B, Broglio J. Trec and tipster experiments with inquery [C]//Information Processing and Management: an International Journal-Special issue: the second text retrieval conference(TREC-2). 1995: 327-343

[7] Allan J, Callan J, Croft W B, et al. Inquery at trec-5 [C]// TREC. 1997: 119-132

[8] Bendersky M, Croft W B. Discovering key concepts in verbose queries[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2008: 491-498

[9] Kumaran G, Allan J. Effective and efficient user interaction for long queries[C]//Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2008: 11-18

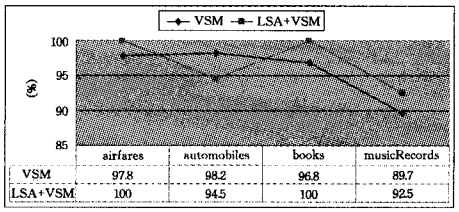
[10] Kumaran G, Carvalho V R. Reducing long queries using query quality predictors[C]//Proceedings of the 32nd annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM, 2009: 564-571

[11] Lease M, Allan J, Croft W B. Regression Rank: Learning to Meet the Opportunity of Descriptive Queries[C]//Proceedings of the 31st European Conference on IR Research on Advances in Information Retrieval. Toulouse, France, 2009: 99-101

[12] Nivre J, Hall J, Nilsson J. MaltParser: A data-driven parser-generator for dependency parsing [C]//Proc. of LREC. 2006

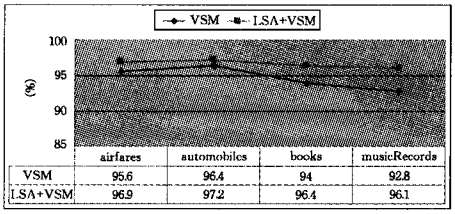
[13] 李珏伶. 搜索引擎网页相关性评估方法设计及其在 rank 模型上的应用[D]. 北京: 北京交通大学, 2011

(上接第 230 页)



	airfares	automobiles	books	musicRecords
VSM	97.8	98.2	96.8	89.7
LSA+VSM	100	94.5	100	92.5

图 4 查全率 R 对比



	airfares	automobiles	books	musicRecords
VSM	95.6	96.4	94	92.8
LSA+VSM	96.9	97.2	96.4	96.1

图 5 F1 值对比

对比实验结果可以得出：(1)使用 LSA 后，查准率较单纯使用 VSM 有所提高。(2)对于存在领域属性重叠的查询接口，采用 LSA 后，反而使部分并不属于同一领域的查询接口之间的相似度增大，本实验中，由于 Automobiles 领域和其他的 3 个领域存在属性重叠现象，由图 4 可以看出，使用 LSA 后查全率反而降低。(3)综合考虑查准率 P 和查全率 R，从 F_1 值可以看到使用了 LSA 后，整体聚类质量有了明显提升。

结束语 本文提出一种基于潜在语义分析的 Deep Web 查询接口聚类方法，新方法通过引入 LSA 理论，发现了 Deep Web 查询接口之间的隐含的语义关系，提升了领域内查询接口之间的相似度。通过在 UIUC 的 Web 集成资源库提供的查询接口集上进行实验，验证了本文提出方法的有效性，相较于传统的基于向量空间模型的聚类方法，新方法明显地提升了聚类质量。

本文提出的新方法仍有值得讨论的问题，如在数据预处理阶段，本文是通过人为地定义特征词表来进行语义拓展，因此在下一步研究工作中，我们将考虑如何进行本体学习，将领域本体引入进来，尝试在本体的指导下进行查询接口的语义拓展，进一步提高 Deep Web 查询接口聚类的效果。

参 考 文 献

[1] 刘伟，孟小峰，孟卫一. Deep Web 数据集成研究综述[J]. 计算机学报，2007，30(9)：1475-1489

[2] Olney, Andrew M. Generalizing Latent Semantic Analisis[C]// 2009 IEEE International Conference on Semantic Computing. 2009: 40-46

[3] Liu Yun-feng, Qi Huan. Latent Semantic Analysis of Chinese Information[J]. Journal of South China University of Technology (Natural Science), 2004(32): 107-111

[4] Li Ya-xiong, Zhang Jian-qiang, Dan Hu. Text Clustering Based on Domain Ontology and Latent Semantic Analysis[C]//2010 International Conference on Asian Language Processing. 2010: 219-222

[5] Thomas H. Unsupervised Learning by Probabilistic Latent Semantic Analysis[J]. Machine Learning, 2001, 42(2): 177-196

[6] 黄承慧，印鉴，侯昉. 一种结合词项目 TF-IDF 方法的文本相似度量方法[J]. 计算机学报，2011，34(5)：857-864

[7] Mao Qin-jiao, Feng Bao-qin, Pan Shan-lang. Latent Semantic Analysis for Query Ierfaces of Deep Web Site[J]. Journal of SouthEast University (English Edition), 2008, 24(3): 312-314

[8] 盖杰，王怡，武港山. 基于潜在语义分析的信息检索[J]. 计算机工程，2004，30(2)：58-60

[9] Wu Chen, Vidyasagar P, Chang E. Latent Semantic analysis-The Dynamics of Semantics Web Services Discovery[J]. Lecture Notes in Computer Science, 2008, 4891: 346-373