

基于VSM的OAI-PMH元数据相似度计算研究

赵治军,陈立潮,谢斌红,王秀慧

(太原科技大学 计算机科学与技术学院,山西 太原 030024)

摘 要:针对OAI-PMH存在大量相似的元数据,结合元数据内容的结构特性,提出了一种基于向量空间模型的OAI-PMH元数据相似度的计算方法。为了较好地反映特征项在元数据内容不同层次的重要程度,采用了结构层次权重系数改进的TF-IDF方法来计算特征项权重。实验分析表明,使用该方法对元数据之间相似度的计算是有效的、可行的。为用户在搜索查询时提供了基于元数据的相似资源,方便了用户,提高了信息服务质量。

关键词:元数据相似度;向量空间模型;结构层次权重系数

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2009)09-0119-03

Research of Calculating Metadata Similarity in OAI Framework Based on VSM

ZHAO Zhi-jun, CHEN Li-chao, XIE Bin-hong, WANG Xiu-hui

(Institute of Computer Science and Technology, Taiyuan University
of Science and Technology, Taiyuan 030024, China)

Abstract: A large number of similar metadatas exist in the OAI-PMH. In order to calculate the similarity among metadatas, a method which is based on VSM is proposed while considering the structural characteristic of metadata content. In order to reflect the importance of characteristic term in different layer of metadata content, the improved TF-IDF which is based on structure-layer weight coefficient is used to measure the characteristic term in VSM. The experimental analysis shows that the use of the method for calculating the similarity among metadatas is effective and feasible. It provides metadatas for users which is based on similar resources, facilitates users and improves the service quality of information.

Key words: metadata similarity; vector space model; structure-layer weight coefficient

0 引言

OAI协议(OAI-PMH)是近几年提出来的一个元数据互操作协议,为实现元数据的传播、共享、分发与互操作提供了新的解决办法。由于它具有简单、灵活、低门槛、易实现等特性,该协议得到了广泛的应用^[1-3]。随着其应用范围的不断扩大,数据提供方中的元数据量也在不断增大,出现了大量的相似元数据。为了提高信息服务质量,给用户基于元数据相似的资源信息,需要对元数据间的相似度进行研究。

基于向量空间模型的相似度计算是当前最为成熟且效果较好的一种方法,它通过将文本简化为特征向量表示,从而把文本相似度计算问题简化为空间向量

的运算,使得问题的复杂性在很大程度上降低。另外,基于向量空间的模型可以使用不同的权重计算方法,使得模型的适应性比较广泛^[4]。

目前,数据提供方对元数据间相似性的挖掘还没有涉及,鉴于元数据相似度计算重要性,文中运用向量空间模型计算数据提供方中的元数据相似度。考虑到特征项在元数据内容的不同层次表达的重要程度不同,在向量空间模型中采用结构层次权重系数改进的TF-IDF方法来计算特征项权重。

1 相关研究

1.1 向量空间模型

向量空间模型^[5,6](Vector Space Model, VSM)是由Salton于20世纪70年代中期提出,主要思想是通过使用向量的相似性来解决语义上的相似性。其基本原理是:对文本进行预处理,抽取代表其元数据特征的特征项,组成向量,借助向量之间的距离,来解决文本

收稿日期:2009-01-16;修回日期:2009-03-09

基金项目:太原市科技项目(ZB0701-04);校青年基金(2007133)

作者简介:赵治军(1982-),男,硕士研究生,CCF会员,研究方向为数据挖掘;陈立潮,教授,研究方向为人工智能与模式识别、智能软件。

之间相似的问题。其中有几个关键的常用概念:

文献:泛指各种机器可读的记录,可指一篇文章或一个网页,也称为文本。文中将侧重对元数据记录的讨论。

特征项:指出现在文本中且能够代表该文本内容的基本语言单位,主要是由词或者短语构成。文本 D 用特征项构成的向量表示: $D = (t_1, t_2, t_3, \dots, t_n)$ 。

特征项的权重:对含有 n 个特征项的文本而言,文本 $D = (t_1, t_2, t_3, \dots, t_n)$, 特征项 $t_k (1 \leq k \leq n)$ 常常被赋予一个数值 w_k , 表示它在文本中的重要程度,称为特征项 $t_k (1 \leq k \leq n)$ 的权重。最后一般用 $D = (t_1, w_1; t_2, w_2; t_3, w_3; \dots; t_n, w_n)$ 的形式表示文本。

1.2 元数据记录分析

OAI 协议采集请求返回的元数据是以记录的方式进行组织,每条记录都对应一条完整的元数据信息,主要由两部分组成:头部(header)和元数据(metadata)。

(1) 头部。

头部由记录标识符(identifier)、时间戳(date stamp)和集合(set)组成。

(2) 元数据。

按照 OAI-PMH 协议规定,DP 必须以 DC 元数据形式提供元数据。

以下是一条元数据记录的实例:

```
<record> //记录开始
<header> //头部开始
<identifier> oai:oaicat.oclc.org:2002/ocm11992160 </identifier>
<timestamp> 2008-03-31T08:48:29Z </timestamp>
</header> //头部结束
<metadata> //元数据开始
<oai_dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>关于财政支农问题的若干理论思考</dc:title>
//资源标题
  <dc:creator>邓子基</dc:creator> //资源撰写者
  <dc:publisher>福建论坛 人文社会科学版编辑部</dc:publisher> //资源出版者
  <dc:subject>财政支农;三农问题;财政补贴</dc:subject>
//资源关键词
  <dc:language>中文</dc:language> //资源所用语言
  <dc:description>
```

财政投资主要是公共产品领域,由于我国农业的特殊性,农业本身在相当大程度上具有公共产品的性质。因此,我国财政应当对具有公共产品性质的农业承担起保护和扶持责任。

```
</dc:description> //资源摘要
<dc:date> 2004-04-01 </dc:date> //资源日期
<dc:type> 文章 </dc:type> //资源类型
<dc:contributor> 厦门大学 </dc:contributor> //对资源的内容作出贡献的其他实体
<dc:format> pdf </dc:format> //资源格式
</oai_dc:dc>
</metadata> //元数据结束
</record> //记录结束
```

从前面对元数据记录的分析可以看出,描述元数据内容的所有数据项都位于叶子节点,且在同一目录下,因此相似度计算不用考虑路径问题。同时元数据中 title^[2]等项又可以看成一个对元数据描述的文本,所以可以运用向量空间模型计算元数据间的相似度。但如果把元数据所有项表示成为一个文本,会使文本冗长、相似度计算量大,需要选择一些能代表元数据的数据项。元数据中 title、description 和 subject 这三项分别代表了元数据的标题、描述和关键词,是元数据的主要内容和关键部分,能够表示一条元数据,可以用它们描述的数据项来表示一个元数据文本。

2 基于 VSM 的 OAI-PMH 元数据相似度计算

2.1 特征项的选取

文本特征项的选取,有多种选择,可以选择字、词作为特征项。一般情况下,人们普遍认为以单字作为特征项对文本内容的代表性不如以词语作为特征项对文本内容的代表性强。因此,在大多数情况下,词语会被选作特征项。现在对词语的获取方法很多,有人通过使用互信息、信息熵等手段对文本中的词进行抽取,这样就可以抽出对该文本内容最具表现力的若干条词语作为该文本的特征项。鉴于这种方法现在还不能达到人们所期望的程度,特别是对中文的特征项抽取,现在的准确率也只是在 70% 左右。获取文本特征项当前使用较多的方法还是用分词的方法来进行,然后,再利用一些过滤手段对特征项进行筛选,最终得到一组具有较强的代表性的特征项。

采用 Lucene 的分词算法,对元数据文本进行分词处理,过滤掉一些停用字和长度大于 5 的词,为每个文本生成一个向量 $D = (t_1, t_2, \dots, t_n)$, t_i 为特征项词条。

2.2 特征项权重的计算

用 w_{ik} 代表元数据文本 i 中项 k 的权重,权重的计算主要运用 TF-IDF 公式,目前存在 TF-IDF 公式,在系统中采用了一种比较普遍的 TF-IDF 公式,见公式(1)。

$$w_{ik} = tf_{ik} * \log(N/n_k + 0.5) \quad (1)$$

其中: tf_{ik} 代表项 k 在元数据文本 D_i 中出现的频率, N 为文本总数, m_k 代表含项 k 的元数据文本数。

传统的向量空间模型进行统计特征项权重 w_{ik} 时,没有区别不同位置的文本特征项对表达文本内容的不同能力^[7]。考虑到元数据文本由 XML 文档转化而来,而 XML 文档又具有树型结构特点,虽然选取的项都是树型结构的叶子节点,但特征项所在的节点不同,其表示能力不同,根据它们对文本内容表达的重要性可知,出现在 subject 的特征项要比出现在 title 的特征项更能确切代表文本的内容,同样出现在 title 的特征项也要比出现在 description 中的特征项更能代表文本的内容。因此把元数据文本划分为 title、subject 和 description 三个层次。这样在统计每个层次的特征项权重 w_{ik} 后,再乘以一个反映其重要程度的结构层次权重系数^[8]来加以调整,那么特征项权重 $w_{ik}' = \lambda \times w_{ik} \circ \lambda = \sum_{j=1}^3 tf_{ikj} \lambda_j$, tf_{ik1} , tf_{ik2} , tf_{ik3} 分别为特征项在 subject、title、description 中的词频; λ_1 , λ_2 , λ_3 分别为其加权系数,其中 $\lambda_1 > \lambda_2 > \lambda_3$, 且 $\sum_{j=1}^3 \lambda_j = 1$ 。

2.3 相似度计算

在向量空间模型中,两个元数据文本分别用两个向量表示。采用夹角余弦的方法来计算这两个向量间的相似度 $\text{Sim}(D_i, D_j)$ 见公式(2)。夹角余弦是一种归一化的相似度计算方法,两个向量之间的夹角越小,其 \cos 值越大,对应元数据文本间的相似程度就越高。两个向量的夹角余弦等价于把它们标准化为单位长度后的向量内积,它反映了两个向量的词条分量相对分布的相似性。

$$\text{Sim}(D_i, D_j) = \frac{\sum_{k=1}^n W_{ik} W_{jk}}{\sqrt{\sum_{k=1}^n W_{ik}^2} \sqrt{\sum_{k=1}^n W_{jk}^2}} \quad (2)$$

3 实验结果

该实验中采用了 OAISter Search 网站中的元数据,并将其用文本格式表示,构成了一个元数据文本集。

在实验中,从元数据文本集中选取一些文本作为基准元数据文本。分别在元数据文本集中查找与之相似的文本,并且按照相似度从大到小排序。由于篇幅

限制,表 1 中只列出了一个基准元数据文本的实验结果。

表 1 元数据文本相似度计算的部分结果

	实验元数据文本	相似度
基准元数据文本	title: 基于 GIS 和模型的流域非点源污染控制区划; subject: 农业非点源污染; description: 以模型和 GIS 的定量结果为依据,对九龙江流域农业非点源污染控制进行了区划,利用 GIS 和经验模型回答了流域农业非点源污染氮磷来源与贡献,标识了农业非点源污染氮磷等污染物的关键源区,发挥了经验模型所需模型参数少、研究尺度较大、效率较高的优点	
第一条	title: 基于 GIS 和模型的流域非点源污染控制区划; subject: 农业非点源污染; description: 以模型和 GIS 的定量结果为依据,对九龙江流域农业非点源污染控制进行了区划,利用 GIS 和经验模型回答了流域农业非点源污染氮磷来源与贡献,标识了农业非点源污染氮磷等污染物的关键源区,发挥了经验模型所需模型参数少、研究尺度较大、效率较高的优点	1.0
第二条	title: AnnAGNPS 模型在九龙江流域农业非点源污染模拟应用; subject: 农业非点源污染; description: 运用连续-分布式参数模型进行中国南方山区中等尺度流域——九龙江流域农业非点源污染负荷估算和对流域过程和管理措施的模拟。利用 4 个典型汇水区校正模型参数,并进一步在九龙江的北溪和西溪两大支流流域验证模型的适宜性	0.758
第三条	title: 南方丘陵地区农业小流域最佳管理措施模拟评价; subject: 农业非点源污染; description: 农业非点源污染严重影响流域水质,对营养盐氮磷流失的模型模拟与控制措施方案的科学制定是当前国内外研究重点。利用多年的降雨-径流、水质实测数据对农业非点源污染模型 (AGNPS) 进行了模拟校验,通过校验后的模型用于最佳管理措施的模拟评价	0.649
第四条	title: 构建流域农业非点源污染控制的环境经济手段研究; subject: 农业学报; description: 阐述了环境经济手段在我国农业非点源污染控制中应用的必要性和可行性。并结合福建省九龙江流域农业非点源污染的主要问题,提出排污权交易,化肥和农药施用税、补贴、保险,退耕补贴、水保押金等环境经济手段及其应用形式,并分析论证了其实施后的费用效果和可行性	0.593
第五条	title: 如何评价乡镇企业对中国农业增长和农村发展的影响; subject: 农业增长; description: 对福建省乡镇企业发展和农业增长情况实地调查,集中讨论自经济改革以来乡镇企业迅速扩张对农业增长的影响,同时运用生产函数研究对这一分析做了进一步的补充	0.223
第六条	title: 关于财政支农问题的若干理论思考; subject: 三农问题; description: 由于我国农业的特殊性,农业本身在相当大程度上具有公共产品的性质。因此,我国财政应当对具有公共产品性质的农业承担起保护和扶持责任。鉴于目前三农问题的严重性,应当在经济发展和财政增收的基础上,逐步加大财政支农补贴的力度	0.201

对于这个基准元数据文本,从实验结果中选出 6 个最相似的。从实验结果可知,相似度的值介于 0 与 1 之间,值越接近于 1 表明两元数据文本越相似,等于 1 达到极值,表明两元数据文本相同。

在表 1 中,第一条元数据文本与基准元数据文本完全相同,计算的结果为 1.0;第二、三条与基准元数据文本的 subject 内容相同,根据文中所采用的方法 subject 有较高的权重,得到的相似度值较大,但第二条 title 内容又比第三条更为接近基准元数据文本,计算的相似度的值较大;第五、六条中 subject 内容与基准元数据文本不完全相同,计算的相似度值较第一、二、

(下转第 199 页)

传感器	类型	端口号	是否报警	操作	
1号烟感	烟感	1#	是	修改	删除
机房东时间	温度	2#	是	修改	删除
红外探测器	红外	3#	是	修改	删除
2号烟感	烟感	4#	是	修改	删除

图5 传感器管理界面

4 结束语

DAM 机房监控设备的 Web 服务系统划分为运行状态模块、采集机信息模块和系统设置模块。系统采用 CGI 技术设计 Web 服务。DAM 采集机满足了机房环境设备的网络化管理要求,通过网络管理各监控分站,实现远程集中监控。在厦门和福建省环保领域的应用来看,DAM 采集机产品满足了设计要求和系统可靠性要求。依据用户的需求,今后还需要进一步开发和改进远程中心控制系统^[10],拓展产品的应用领域,实现环境监控的大集中管理和维护的目标。

参考文献:

[1] 方刚,于晓宝.计算机机房管理[M].北京:清华大学出

(上接第121页)

三条小,虽然第四条元数据文本的 title 内容与基准元数据文本比较相似,但它的 subject 没有第三条与基准元数据文本接近,而 subject 的权重系数又比 title 的大,所以它比第三条与基准元数据文本的相似度值小。通过试验,可以得出文中使用方法能够较好地计算元数据之间的相似度,实现相似元数据的提取。

4 结束语

OAI-PMH 数据提供方中元数据量的不断增加,出现了大量的相似元数据;而且相似性度计算被广泛用于基于内容相似资源的推荐,对服务质量有着重要的影响。文中给出了一种计算元数据间相似度的方法,从实验结果可知,采用该方法对元数据间相似度的计算是可行的。

将来研究工作打算改进分词算法,提高特征项选取的精度;在文中权重计算方法的基础上采用语义相似度改进。

参考文献:

[1] Suleman H, Virginia Tech, Blacksburg. Introduction to the

版社,2001.

[2] 王存健,张建正.嵌入式 Linux 下 Qt/Embedded 的应用[J].计算机技术与发展,2006,16(11):185-187.

[3] 马溪骏,陈宜义,杨善林.基于实时嵌入式 Linux 的金属液智能检测仪设计[J].计算机技术与发展,2007,17(1):212-215.

[4] 石勇,王凡.基于低速数据采集的虚拟综合测试系统设计与实现

[J].电子测试,2008(1):46-49.

[5] Shibave S, Counsell G. MAST Data Acquisition System[J]. Fusion Engineering and Design,2006,81(15):1789-1793.

[6] 田军营,韩建海,马志荣.μclinux 源代码中 Make 文件完全分析——基于 ARM 开发平台[M].北京:机械工业出版社,2005.

[7] Nishimura M. Web 应用程序:CGI 到 Web 三层系统[M].高敬译.北京:科学出版社,2004.

[8] 张移山.CGI 程序设计指南[M].北京:中国水利水电出版社,1998.

[9] 刘英,罗家融.EAST 数据采集控制系统[J].计算机工程,2008,34(14):228-230.

[10] 余立建.水位远程测量与数据传输技术[J].测试技术学报,2008(4):17-21.

Open Archives Initiative Protocol for Metadata Harvesting [C]//Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries. Portland, Oregon, USA: [s. n.], 2002.

[2] Liu X. Federating Heterogeneous Digital Libraries by Metadata Harvesting[D]. [s. l.]: Department of Computer science, Old Dominion University, 2002.

[3] Lagoze C, Van de Sompel H. The making of the Open Archives Initiative Protocol for Metadata Harvesting[J]. Library Hi Tech,2003,21(2):18-20.

[4] 沈斌.基于分词的中文文本相似度计算研究[D].天津:天津财经大学,2006.

[5] Pang Jianfeng, Bu Dongbo, Bai Shuo. Research and implementation of text categorization based on VSM[J]. Application Research of Computers,2001(9):23-26.

[6] 张冉.基于 XML 和 N 层 VSM 的 web 信息检索[J].计算机技术与发展,2006,16(5):56-58.

[7] 陈治纲,何丕廉,孙越恒,等.基于向量空间模型的文本分类方法的研究与实现[J].计算机应用,2004(6):277-279.

[8] 刘海峰,王元元,王倩.基于分类的 VSM 模式下文本检索若干问题研究[J].情报科学,2006,24(11):1700-1703.