

一种用于文章推荐系统中的用户模型表示方法

赵 鹏^{1,2}, 蔡庆生², 王清毅²

(1. 安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039;

2. 中国科学技术大学 计算机系, 安徽 合肥 230027)

摘 要: 分析了现有文章推荐系统中基于关键词向量的用户模型表示方法存在的不足, 提出了基于聚类兴趣点的用户模型表示方法。该方法可通过文章聚类形成兴趣点。由于传统的基于划分的聚类算法存在的不足, 提出了基于复杂网络特征的文章聚类算法。实验结果表明该用户模型的表示方法较好地反映了用户多方面的兴趣, 提高了文章推荐系统的性能。

关键词: 聚类; 复杂网络; 推荐系统; 用户模型

中图分类号: TP18

文献标识码: A

文章编号: 1673-629X(2007)01-0004-02

A Novel Representation of User Profile in Document Recommendation System

ZHAO Peng^{1,2}, CAI Qing-sheng², WANG Qing-yi²

(1. Ministry of Education Key Lab. of Intelligent Computing & Signal

Processing, Anhui University, Hefei 230039, China;

2. Dept. of Computer Sci. and Tech., Univ. of Sci. and Tech. of China, Hefei 230027, China)

Abstract: After analyzing the disadvantages of the user profile based on keywords vector in the existing document recommendation system, a novel representation of user profile based on clustering was proposed. The representation firstly clustered the documents into clusters. Because of the disadvantage of the traditional partitioned clustering algorithm, a novel document clustering algorithm based on complex networks feature was presented. Experimental results show the representation of user profile proposed can represent user multi-interests better and improves the performance of document recommendation system greatly.

Key words: clustering; complex networks; recommendation system; user profile

0 引言

Internet 的发展使得越来越多的用户习惯于网上阅读和收藏文章。文章推荐系统通过收集和分析用户信息来了解用户的兴趣, 为不同的用户推荐其感兴趣的文章, 从而使其信息需求得到最大程度和最高效的满足。

用户模型的表示、建立和更新是影响文章推荐系统性能的一个主要因素, 即用户模型能否较好地反映用户的兴趣偏好, 随着用户兴趣的转移, 用户模型能否作适应性的改变, 较好地反映用户兴趣的变化^[1]。

用户模型表示方法主要有布尔或加权关键词向量、语义网、n-grams 以及本体论向量^[2]等。由于基于关键词的向量表示方法简单有效, 因而被广泛地应用于隐式获取用户模型的系统中。但是用户收集的文章数量不断增加, 采用关键词向量表示用户模型, 向量空间维数超大, 而且

增长速度很快, 如果限制向量空间维数, 则又不能全面反映用户多方面的兴趣, 尤其是一个新的兴趣点的关键词往往被老的兴趣点的关键词所掩盖, 并且基于关键词的用户模型无法体现用户在各个兴趣点上的差异。

针对现有推荐系统中用户模型存在的上述问题, 文中提出了基于聚类的用户模型表示方法。该方法首先对文章进行聚类, 每一个聚类视为一个兴趣点, 然后将用户模型表示为基于兴趣点的向量。基于划分的 K-means 聚类方法是文章聚类中使用最为广泛的聚类方法之一, 它具有算法简单且收敛速度快的特点。但是算法的性能依赖于聚类中心的初始位置, 即对于随机的初始值选取可能导致不同的聚类结果, 甚至存在无解的情况, 而且算法需要事先确定聚类个数 k ^[3]。为了克服上述缺点, 文中提出了基于复杂网络特征的文章聚类算法。实验结果表明该用户模型表示方法可以显著地提高文章推荐系统的性能。

1 基于复杂网络特征的文章聚类算法

复杂网络自 20 世纪末逐渐兴起以来, 复杂网络理论与实证的研究在国际科学界蓬勃发展^[4]。目前对复杂网

收稿日期: 2006-04-24

基金项目: 安徽省自然科学基金项目(2004kj011); 安徽省高校青年教师基金项目(2006jq1040)

作者简介: 赵 鹏(1976-), 女, 博士研究生, 讲师, 研究方向为人工智能、机器学习。

络的研究主要针对非加权复杂网络,即只考虑所有边的权值都一样的情况。而现实的网络中,边的权值是不一样的,并且会影响整个网络的性能。加权复杂网络能够比较完整地表达复杂网络的结构。

文章之间根据其相关性建立连边,同样构成一个加权复杂网络。文中将文章聚类问题转换为加权复杂网络上节点聚类的问题,利用节点的加权重与加权聚集系数这两个重要特征,提出基于复杂网络特征的文章聚类算法。

定义1: 设 $V = \{v_1, v_2, \dots, v_N\}$ 为一节点集合,令无序偶对 (v_i, v_j) 表示节点 $v_i \in V$ 与 $v_j \in V$ 之间的边, w_{ij} 为边 (v_i, v_j) 的权值。设 $WG(V, E, W)$ 是以 V 为节点集合,以 $E \subset \{(v_i, v_j): v_i, v_j \in V\}$ 为边集合,以 $W = \{w_{ij}: (v_i, v_j) \in E\}$ 为权值集合的图,则

节点 v_i 的度 D_i 为:

$$D_i = |\{(v_i, v_j): (v_i, v_j) \in E, v_i, v_j \in V\}| \quad (1)$$

节点 v_i 的加权重 WD_i 为:

$$WD_i = \sum_{(v_i, v_j) \in E} w_{ij} \quad (2)$$

节点 v_i 的加权聚集度 WK_i 为:

$$WK_i = \sum_{(v_i, v_j) \in R} w_{jk} \quad (3)$$

其中 $R = \{(v_j, v_k): (v_i, v_j) \in E, (v_i, v_k) \in E, v_i, v_j, v_k \in V\}$ 。

节点 v_i 的加权聚集系数 WC_i 为:

$$WC_i = WK_i / \binom{D_i}{2} = \frac{2WK_i}{D_i(D_i - 1)} \quad (4)$$

节点的加权重反映了该节点与其它节点的连接强度。节点的加权聚集系数则体现了此节点局部范围内的相互连接密度和强度。

基于复杂网络特征的文章聚类首先根据文章的相关性,建立加权复杂网络,然后依次选取加权重和加权聚集系数大的节点作为聚类中心,与该聚类中心相关的节点聚为一类,并从复杂网络上删除,重复直到所有的节点都被删除,最后将特别小的类合并形成杂类。该算法的具体步骤如下:

STEP1: 计算文章之间的相关度,以文章作为节点,相关文章间建立连边,相关度作为权重,建立加权复杂网络。

STEP2: 然后计算网络上各个节点的加权重 WD_i 、加权聚集系数 WC_i 和综合特征值 WCF_i (公式5)。

$$WCF_i = \alpha WC_i + (1 - \alpha) WD_i \quad (5)$$

其中 α 为可调节的参数, $0 < \alpha < 1$ 。

STEP3: 对各节点的综合特征值进行排序,形成由大到小的队列 queue。

STEP4: 从队列 queue 中依次选取尚未标注类别的文章作为聚类中心,并标注类号,然后将与该聚类中心文章相关的文章标注相同的类号。

STEP5: 重复 STEP4,直到取完队列 queue 中所有的节点。

STEP6: 统计所有的聚类,将聚类中文章数小于 N 的

聚类合并,形成杂类。

2 基于聚类兴趣点的用户模型的表示方法

基于聚类兴趣点的用户模型表示方法首先对所有用户收集的文章进行聚类形成多个兴趣点,然后将用户模型表示为加权兴趣点向量。下面介绍构造基于聚类兴趣点的用户模型的具体步骤。

STEP1: 对所有用户收集的文章进行预处理,将文章表示成加权关键词向量。根据经验,文中抽取每篇文章权重最高的前 20 个词作为关键词,收集所有文章的关键词,去重后构成关键词集,然后将每一篇文章表示为加权关键词向量^[5]。

STEP2: 采用基于复杂网络特征的文章聚类算法,对文章进行聚类,形成聚类集合 CL ,每个聚类 CL_i 视为一个兴趣点。

STEP3: 根据每个用户 U_i 收集的文章,统计文章所属的兴趣点,构建加权聚类兴趣点向量 UM_i 来表示用户模型:

$$UM_i = (w_{i1}, w_{i2}, \dots, w_{i|CL|})$$

其中 w_{ij} (公式6) 为用户 U_i 对兴趣点 CL_j 的权重, $1 \leq i \leq |U|, 1 \leq j \leq |CL|$ 。

$$w_{ij} = DC_{ij} / DN_i \quad (6)$$

其中 DC_{ij} 为用户 U_i 收藏中属于兴趣点 CL_j 的文章篇数, DN_i 为用户 U_i 收藏文章的总篇数。

用户收藏的文章不断增加,关键词不断增加,但是兴趣点相对稳定,并且兴趣点向量的维数要远远小于关键词向量维数。

3 实验结果与分析

为了测试文中提出的用户模型表示方法的有效性,在文章推荐系统中分别用基于关键词向量的用户模型和文中提出的基于聚类兴趣点的用户模型作了比较实验。

文中采用某网站提供的部分真实数据集,其中包括 714 个用户,20000 篇文章。实验从中任取 10 篇文章作为待推荐文章。实验采用推全率作为推荐系统的质量指标,其定义如下:

定义2: 对于推荐系统 RS,令 RD 为待推荐文章, RDU 为系统推荐 RD 的用户集, SDU 为收藏 RD 的用户集,则推全率 Recall 定义为:

$$\text{Recall} = \frac{|RDU \cap SDU|}{|SDU|}$$

实验中的文章聚类后形成 98 个兴趣点。实验中的相似度计算采用余弦相似度计算方法。实验结果见表 1,由表 1 可以看出采用文中提出的基于聚类兴趣点的用户模型的系统文章推全率高于采用基于关键词集的用户模型的系统。这是由于用户模型采用关键词向量表示,无法全面地反映用户兴趣,一些属于用户较新兴趣点的文章,由

(下转第 48 页)


```

tionProxyFactoryBean">
    <property name="transactionManager"><ref bean="
"transactionManager"/></property>
    <property name="transactionAttributes">
    <props>
    <prop key="save *">PROPAGATION_REQUIRED</
prop>
    <prop key="remove *">PROPAGATION_REQUIRED
</prop>
    <prop key="*">PROPAGATION_REQUIRED</prop
>
    </props>
    </property>
    </bean>
    <bean id="supplierManage" parent="txProxyTemplate">
    <property name="target">
    <bean class="com.erp.service.implement.Supplier-
ManageImpl">
    <property name="supplierDAO"><ref bean="supplier-
DAO"/></property>
    </bean>
    </property>
    </bean>

```

com.erp.service.implement.SupplierManageImpl 就是实现供应商管理的 JavaBean。通过 parent 元素声明为其提供事务支持的父 bean。此外在管理类 bean 中可以覆盖或添加新的事务处理属性以达到个性化管理的目的。

2)持久层(Hibernate 框架):将数据库中的每张表通过 Hibernate 工具产生相应的持久层对象,然后通过建立 DAO 来使用这些对象。

3)表现层(Struts 框架):通过 Struts MVC 模式来处理

(上接第 5 页)

于该用户收藏这类文章还比较少,其中关键词在该用户模型中无法体现,而基于聚类兴趣点的用户模型表示方法可以较为全面地反映用户兴趣,即便是该用户较新兴趣点,收藏文章较少,也可以在用户模型中得到反映,因而采用基于聚类兴趣点的用户模型的系统推全率较高。

表 1 采用两种用户模型的文章推荐系统的推全率比较

待推荐文章	1	2	3	4	5	6	7	8	9	10
基于关键词集的用户模型	0.71	0.54	0.75	0.82	0.41	0.32	0.28	0.35	0.62	0.65
基于聚类兴趣点的用户模型	0.92	0.82	0.95	0.97	0.65	0.60	0.40	0.70	0.89	0.91

4 结 论

文章自动推荐是个性化服务的重要应用之一。随着系统用户和文章数目日益增加,基于关键词向量的用户模型维数增长过快,且难以全面反映用户兴趣。针对上述问题,文中提出了基于聚类兴趣点的用户模型表示方法。理

客户请求和返回结果。

4 结束语

文中讨论了 J2EE 开发中的轻量级框架组合 Struts + Spring + Hibernate,并对这种组合应用于 Web - MIS 开发进行了研究与设计。三种框架的组合使得 J2EE 的开发有更好的扩展性、可维护性,能充分发挥三者的优势,实现 Web - MIS 系统开发的松耦合,因此具有很好的应用前景。

参考文献:

- [1] Turner J, Bedell K. Struts kick start 中文版[M]. 孙 勇译. 北京:电子工业出版社, 2004.
- [2] Johnson R. Spring Framework reference documentation[EB/OL]. 2004. <http://www.springframework.org/documentation>.
- [3] Johnson R, Hoeller J. Expert One-on-One J2EE Development without EJB[M]. America: Wrox, 2004.
- [4] Fowler M. Inversion of Control Containers and Dependency Injection pattern[EB/OL]. 2004. <http://martinfowler.com/articles/injection.html>.
- [5] 夏 昕. 深入浅出 Hibernate[M]. 北京:电子工业出版社, 2005.
- [6] Eagle M. Wiring Your Web Application with Open Source Java[EB/OL]. 2004. <http://www.onjava.com/pub/a/onjava/2004/04/07/wiringwebapps.html>.
- [7] 曹广鑫, 王谢华, 王建凤, 等. Struts 数据库项目开发宝典[M]. 北京:电子工业出版社, 2006.
- [8] 李伟镰, 卢建朱. 基于 Struts 和 Hibernate 的电子申购系统[J]. 计算机工程, 2005, 31(19): 220 - 222.

论分析和实验结果表明了该方法能够较全面地反映用户多方面的兴趣,大大降低向量空间的维数,使得文章自动推荐系统的性能得到显著提高。

参考文献:

- [1] 曾 春, 邢春晓, 周立柱. 个性化服务技术综述[J]. 软件学报, 2002, 13(10): 1952 - 1961.
- [2] 宋丽哲, 牛振东, 宋瀚涛, 等. 数字图书馆个性化服务用户模型研究[J]. 北京理工大学学报, 2005, 25(1): 58 - 62.
- [3] Ordonez C, Omiecinski E. Efficient disk-based K-means clustering for relational databases[J]. IEEE Trans Knowledge and Data Engineering, 2004, 16(8): 909 - 921.
- [4] 朱孟潇. 复杂网络系统的智能涌现及其应用研究[D]. 合肥:中国科学技术大学, 2004.
- [5] Lee D L, Chuang H, Seamons K. Document ranking and the vector-space model[J]. IEEE Software, 1997, 14(2): 67 - 75.