# 非负矩阵分解在标签语义分析中的应用

张雷鸣[1] 李秋丹[1] 廖胜才[2]
(中国科学院自动化研究所复杂系统与智能科学重点实验室 北京 100190)[1]
(中国科学院自动化研究所模式识别国家重点实验室 北京 100190)[2]

**摘 要** 随着Web2.0技术的发展,社会标注系统日渐流行起来,使得标签在用户收藏的检索和分类管理等方面得到了广泛的应用。然而,由于用户使用标签的自由、非控制性,导致标签在使用上存在冗余和语义模糊性。为了处理该问题,提出一种基于非负矩阵分解(Non-negative Matrix Factorization,NMF)的标签语义挖掘算法,通过对用户的标注数据进行非负矩阵分解,得到一个包含一系列语义相关标签基的标签子空间,使得同义及相关的标签聚合于同一标签基,且一词多义的标签归类到语义不同的标签基,从而实现标签语义的近义归类和多义辨析。通过大量实验充分展示了提出的算法在标签语义挖掘方面的有效性。

**关键词** 非负矩阵分解,标签,标签语义挖掘
**中图法分类号** TP391 **文献标识码** A

## Application of Non-negative Matrix Factorization in Tag Semantics Analysis

ZHANG Lei-ming[1] LI Qiu-dan[1] LIAO Sheng-cai[2]
(The Key Laboratory of Complex System and Intelligence Science, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)[1]
(National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)[2]

**Abstract** With the development of Web2.0 technologies, social tagging systems are becoming more and more popular, which makes tags widely used to retrieve, categorize, and manage users' collections. However, people are free and uncontrollable to use tags, resulting in a large number of tags that are redundant, unclear in semantics. To deal with this problem, we proposed a tag semantics mining algorithm based on non-negative matrix factorization method. We got a tag subspace containing a series of semantic related tag-bases by factorizing tagged data of users using non-negativity constraints, to make synonymous and related tags into the same tag-basis, and categorize polysemous tags into different semantic tag-bases. Simultaneously, the tasks of grouping synonymous tags and distinguishing polysemous tags were done by the proposed approach. A large number of experiments demonstrate the effectiveness of the proposed algorithm on mining tag semantics.

**Keywords** Non-negative matrix factorization, Tag, Tag semantics mining

## 1 引言

随着互联网技术的迅猛发展,社会标注成为近年来的热点。以 del. icio. us, Flickr, CiteULike, 豆瓣网等为代表的Web2.0网站,允许用户将网络资源和喜欢的产品等保存在自己的网站账户下,并自由为其添加关键词作为标签(tag),用于日后的检索、分类组织和整理。在添加的过程中,用户还可参考其他人针对同一资源和产品的标注,体现了用户之间的广泛协作以及标注的社会性。

然而,由于用户在标注时可选用任何词语,因此标签不像传统的分类学那样拥有严谨的层级结构,更多地体现了用户对标注内容语义信息的高度概括,带有用户的主观认知特点[4]。这种自由、松散和不受控,导致了标签存在冗余性和模糊性等问题。此外,随着用户使用时间的增长和收藏数的增多,网站中常见的标签列表仅仅指示了使用频率,无法体现标签间的联系,降低了标签的分类组织功用。同时,网站用户总人数的增加带来的标签总数目的增长,也加大了传统搜索方法的检索难度,使在标签空间高效地寻找资源变得日益困难。因此,如何克服标签存在的问题,发掘标签的语义信息,为用户提供更精确的信息服务,已逐渐成为网络挖掘中的研究热点。

目前在互联网搜索和导航中,标签语义挖掘已显现出重要性。Gemmell[8]提出了无监督的层次标签聚类方法,并结合用户兴趣模型,用于个性化搜索。徐雁斐[6]提出了基于协同标记的个性化推荐算法,验证了潜在语义分析(LSA)在清理冗余标签和挖掘深层信息上的有效性。Szomszor[9]通过比

到稿日期:2009-05-12 返修日期:2009-08-04 本文受973国家重点基础研究发展计划(2007CB311007),国家自然科学基金(60703085)资助。

张雷鸣(1982—),男,博士生,主要研究方向为信息检索等,E-mail:leiming.zhang@ia.ac.cn;李秋丹(1976—),女,副研究员,主要研究方向为信息检索、数据挖掘、移动电子商务等;廖胜才(1982—),男,博士生,主要研究方向为模式识别、计算机视觉等。

较多个支持协作标注网站的用户标签云(tag cloud)来发现用户兴趣模型间的相关性和挖掘用户兴趣。Michlmayr[10]从标注数据中学习用户兴趣模型,并利用它们为用户提供个性化导航。

不过,由于标签固有的特点,在应用中日益暴露出冗余性和语义模糊性。针对这些问题,Wu[5]将用户、标签和资源统一于一个多维向量空间中,采用三重概率模型获取隐藏在上述三者共现(co-occurrence)特征背后的语义。Yeung[7]通过网络视觉化工具 Pajek 呈现了用户、Web 文档、标签两两之间的关系,从宏观上展现了标签的语义分布。Begelman[20]将标签的共现数目作为相似性度量依据,构造了一个无向图,采用递归贪心算法寻找语义相关的标签,帮助用户在标注、搜索、浏览、订阅时获得感兴趣的标签推荐。

本文提出一种新的基于非负矩阵分解(Non-negative Matrix Factorization,NMF)的语义相关标签挖掘算法,从产品和标签构成的信息空间出发,利用非负矩阵分解算法得到相关标签的基矩阵,并从中分析标签间的语义相关性,如图 1 所示(图例中全白权值为 0,颜色越黑权值越大)。

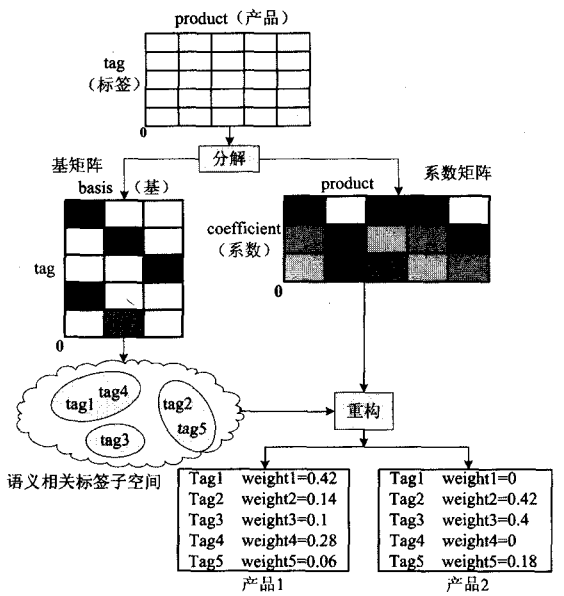


图 1 基于非负矩阵分解的语义相关标签挖掘算法

非负矩阵分解(NMF)算法由 Lee 和 Seung 发表于 1999 年的自然杂志(Nature)[1],在 2001 年神经信息处理系统会议(NIPS)上,两人给出了该算法是严格单调非增,并保证收敛于局部最优值的证明[2]。NMF 与主成分分析(PCA)、K-L 变换、潜在语义索引(LSI)一样,都是多变量分析方法,但在矩阵分解过程中 NMF 受到非负性约束,其他方法则可以存在负值。在实际应用中,负值元素往往缺乏物理意义,NMF 的非负性约束使分解结果具有可解释性。此外,NMF 与矢量量化(VQ)和 PCA 在特征表示方法上也存在差异[1]。NMF 是基于局部(parts-based)的特征表示,分解得到的基矩阵包含了所描述对象集合的局部特征信息,系数矩阵则描述了某个具体对象用到的局部特征以及这些特征所占的权重。非负矩阵分解方法在文档聚类[3]、文档摘要[14,16-19]、文档理解[15]、模式识别[11]、主题检测[13]等多个领域中得到了广泛的应用。

本文应用非负矩阵分解进行语义相关标签的挖掘,本算法具有以下优点:一是能快速有效地实现矩阵降维和分解,得到产品和标签的稀疏的和局部的表达;二是从基矩阵既能直观地得到同义及相关标签的归类,归并冗余的标签,又能分析模糊标签的辨义。将算法在某一点评类网站上进行了实验测试,结果显示,即使在样本数较少的情况下,提出的算法仍能有效地选择出相关度高、能够最大程度覆盖中心语义的标签子空间。

## 2 基于非负矩阵分解的标签语义分析方法

标签的冗余性和语义模糊性问题是社会标注系统面临的主要问题之一。此问题的存在是因为人与人之间在知识层面、个人喜好、语言习惯等方面不可避免地存在差异,导致创建词与它们所指对象间语义关系的过程注定是一种不完美的、自然涌现的过程[5]。本文提出的基于非负矩阵分解的语义相关标签挖掘方法,针对标签和用户标注的对象,定义了一种产品资源与标签关系的信息空间,借助非负矩阵分解算法得到一组描述语义相关标签的标签基,由此分析标签间的语义关系。

### 2.1 非负矩阵分解算法

给定一个非负矩阵 $V$,由 $m$ 个 $n$ 维特征向量构成,寻找非负矩阵因子 $B$ 和 $C$,使得:

$$V \approx BC \tag{1}$$

其中,$B_{n\times r}$ 称为基矩阵,$C_{r\times m}$ 称为系数矩阵。通常选择 $r$ 小于 $n$ 和 $m$,使分解得到的 $B$ 和 $C$ 小于原始矩阵 $V$,从而达到维数约简的目的。

为了寻找 $V$ 的近似分解,定义损失函数为两个非负矩阵 $V$ 和 $BC$ 间的距离:

$$J(B,C)=\|V-BC\|^2=\sum_{i=1}^{n}\sum_{j=1}^{m}[V_{ij}-(BC)_{ij}]^2 \tag{2}$$

易知距离越小,分解越接近原始矩阵。当且仅当 $V=BC$ 时,取得距离最小值 0。

非负矩阵分解算法使用了两个简单的乘性迭代公式进行求解:

$$B_{ik}\leftarrow B_{ik}\frac{(VC^T)_{ik}}{(BCC^T)_{ik}} \tag{3}$$

$$C_{kj}\leftarrow C_{kj}\frac{(B^TV)_{kj}}{(B^TBC)_{kj}} \tag{4}$$

利用上述两个公式,目标函数(2)将逐步收敛于局部极小值。

### 2.2 相关术语定义

**定义 1**(产品资源与标签的信息空间 $IS$) $IS=\{P,T\}$ 是表征产品资源与用户标签间关系的集合。$T=\{t_1,\cdots,t_n\}$ 是标签空间,用于描述产品资源。$T$ 包含了 $n$ 个具有一定描述能力的用户标签,一般选取协作标注系统中使用频率最高的前 $n$ 个标签构成。$P=(p_1,\cdots,p_m)$ 表示产品资源在标签空间 $T$ 中的特征矩阵,每一列包含一个产品资源的特征向量,$m$ 为产品资源总数。对于任意一个产品资源的特征向量 $p_i=(p_{i1},p_{i2},\cdots,p_{in})^T$,都有 $\sum_{j=1}^{n}p_{ij}=1$,其中 $p_{ij}$ 是目前所有用户使用标签 $t_j$ 标注产品资源 $i$ 的频率。

**定义 2**(标签基) $u=(t_{k_1},t_{k_2},\cdots,t_{k_s})$ 是标签空间 $T$ 中的一个标签基,由 $k_s$ 个语义相关的标签共同构成。标签基中排名第一的标签往往占有较大的权重,在该标签基中占有主导地位。其他标签围绕第一个标签描述了一种共同的语义关系。

$U=\{u_1,u_2,\cdots,u_r\}$ 是标签空间 $T$ 中 $r$ 个标签基张成的子

空间，其中每个标签基 $u_k=(t_{k_1},t_{k_2},\cdots,t_{k_s})$，$k_s$ 是第 $k$ 个标签基包含的标签数。在该子空间中，同义和相关的标签常聚合于同一标签基中，一词多义的标签常同时出现于多个语义不同的标签基里。

### 2.3 基于非负矩阵分解的标签语义挖掘算法

给定一个信息空间 $IS=\{P,T\}$，特征矩阵 $P$ 描述了用户使用标签空间 $T$ 中的标签标注每一个产品资源的频率，于是 $P$ 的每一列都可以看成与一个产品资源相对应的样本信号。由于所有信号都是非负的，因此可以针对特征矩阵 $P$ 应用非负矩阵分解算法，分解得到一个基矩阵 $B$ 和一个相应的系数矩阵 $C$，使得 $P\approx BC$。

非负矩阵分解由于在非负可解释性、稀疏性和局部性上的优势，分解得到的 $B$ 和 $C$ 都是非负且稀疏的，实现了压缩编码。更重要的是，矩阵 $B$ 的每一列都是局部稀疏的，只有少数非零的元素，这些非零元素描述了标签空间 $T$ 中的一些潜在的语义结构。如果基矩阵 $B$ 的每一列归一化之和为 1，此时 $b_{ik}$ 便代表了标签 $t_i$ 在列 $k$ 中所占的权重。权重越大，表示该标签在该列中的重要性越大。于是对 $B$ 中的每一列 $k$，可以对其权重由大到小排列，选取小于某个阈值之前的所有对应的标签构成一个标签基 $u_k$，由此可以得到一个标签子空间 $U=\{u_1,u_2,\cdots,u_r\}$，其中 $r$ 是标签基的个数，也即 $B$ 的列数。我们发现，由此得到的标签子空间 $U$ 比原始标签空间 $T$ 能更好地描述产品资源，子空间 $U$ 里的每一个标签基所包含的标签都是语义相关的，具有更综合的描述能力。

基于上述思想，我们得到一种基于非负矩阵分解的语义相关标签挖掘算法，详细流程如下所示。

输入：信息空间 $IS=\{P,T\}$，待学习的标签基个数 $r$，最大迭代步数 $M$，迭代停止阈值 $\varepsilon$ 及累加权重阈值 $E$。

输出：$r$ 个标签基张成的标签子空间 $U$。

步骤 1　初始化基矩阵 $B_{n\times r}$ 和系数矩阵 $C_{r\times m}$，保证矩阵所有元素均为正随机数。计算初始损失函数 $J_0=\|P-BC\|^2$。

For $i=1$ to $M$

步骤 2　以 $P$ 取代非负矩阵分解里的数据矩阵 $V$，利用式(3)、式(4)的乘性迭代规则计算新的 $B$ 和 $C$。

步骤 3　计算新的损失函数 $J_i=\|P-BC\|^2$。

步骤 4　若 $|J_i-J_{i-1}|<\varepsilon$，停止迭代。

End for

步骤 5　将基矩阵 $B$ 的每一列 $k$ 归一化，即 $\sum_{i=1}^{n}b_{ik}=1,k=1,2,\cdots,m$，并由大到小排序，选取该列累加权重小于阈值 $E$ 之前的所有对应的标签构成一个标签基 $u_k$，由此得到标签子空间 $U=\{u_1,u_2,\cdots,u_r\}$。

## 3 实验分析

### 3.1 数据描述

在某点评类网站上，以电影作为产品资源验证了本文算法。数据集采用该网站从 2009 年 2 月到 3 月间的电影数据样本，包含 12401 部电影和 15443 个不同的标签。在实验中，选择该网站通过统计得到的所有用户使用最多的前 2000 位标签构成标签空间 $T$，剔除使用这 2000 个标签累计次数小于 20 的电影，最终剩余 11424 部电影。每部电影根据标签空间 $T$ 里相应标签的标注频率构成特征向量，所有电影的特征向量组成特征矩阵 $P$。由此得到一个电影和相关标签组成的信息空间 $IS=\{P,T\}^{2000\times 11424}$。

### 3.2 参数设定

针对以上得到的信息空间 $IS$ 验证本文提出的算法。图 2 展示了对于指定不同的标签基个数 $r$ 随着迭代的进行非负矩阵分解误差的变化。

从图 2 中可发现，标签基个数 $r$ 取 10 到 70 时，矩阵分解误差下降较快，$r=80$ 以后误差下降幅度逐渐减小，此时分解得到的标签子空间已能比较好地近似表征产品资源与标签间的关系。过大的 $r$ 值会使生成的标签基的语义过于分散，也增加了基空间的维数和计算量。因此，选定要学习的标签基个数 $r=120$。从图中可知，对于矩阵分解的迭代步数，大约 80 次迭代以后两次分解产生的损失函数 $J$ 之间的差值已经很小，算法已接近收敛。据此，设定最大迭代步数 $M=100$，迭代停止阈值 $\varepsilon=0.12$。

我们在 Pentium D 3.4GHZ CPU，2G 内存，Windows XP 操作系统环境下验证了本文算法，其中 $m=11424$，$n=2000$，$r=120$，$M=100$，$\varepsilon=0.12$。算法迭代了 79 步收敛，共用时约 70min。

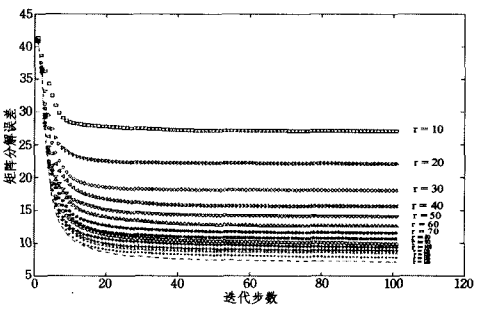


图 2　不同的标签基个数下的矩阵分解误差

### 3.3 标签语义挖掘

通过分析非负矩阵分解得到的基矩阵 $B_{n\times r}$，取累加权重阈值 $E=0.9$，得到 120 组标签基，构成了语义相关的电影标签子空间 $U$。表 1 展示了其中一组例子。

表 1　语义相关的电影标签子空间

| 标签 | 所占权重比例 | 标签 | 所占权重比例 |
|---|---|---|---|
| 惊悚 | 0.287225 | 犯罪 | 0.029533 |
| 恐怖 | 0.136919 | 死神来了 | 0.023679 |
| 美国 | 0.106925 | 血腥 | 0.019366 |
| 沉默的羔羊 | 0.086278 | 恐怖片 | 0.016237 |
| 悬疑 | 0.055189 | 电锯惊魂 | 0.015799 |
| 美国电影 | 0.054057 | 寂静岭 | 0.010121 |
| 心理 | 0.053623 | | |

表 1 中第一个标签“惊悚”构成了核心语义，占有最大权重。“恐怖”、“悬疑”、“心理”等多个标签共同阐述了“惊悚”的语义。从核心标签“惊悚”，用户可以推断出“沉默的羔羊”、“电锯惊魂”、“寂静岭”等影片的类型，进而从这些影片中获得更多的感性认识。

通过非负矩阵分解产生的标签基，还可以推断某些标签的确切含义。如图 3 所示，用户可以通过观察 120 组标签基的权重分布，推测标签空间 $T$ 中“后天”一词在电影背景下的意思以及其与电影产品的相关性。

由图 3 可知，标签“后天”在第 23 和 107 标签基中占有高权重。第 23 标签基含有的标签是“灾难、美国、titanic、美国电影、电影、爱情”，第 107 标签基是“蝴蝶效应、科幻、心理、电影、悬疑、惊悚”。从这两个标签基中，首先可以判断“后天”一词很可能指的是一部美国电影；其次，这部电影的类型属于灾

难、科幻、悬疑和惊悚;第三,电影中还可能涉及到主人公之间的爱情。实际的电影《后天》与两个标签基中所描述语义非常吻合。

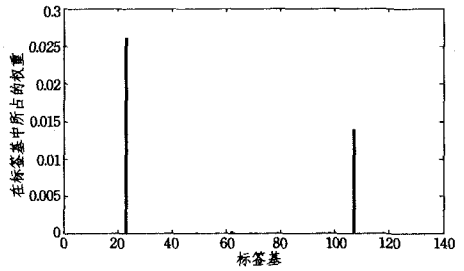


图 3 标签"后天"所属的标签基

### 3.4 系数矩阵分析与电影数据矩阵的重构

非负矩阵分解产生的系数矩阵 $C_{r\times m}$ 表达了每部电影在学习得到的标签子空间 $U$ 上投影的重构系数,反映了每部电影用到的标签基及它们的权重。如图 4 所示,日本著名动画片导演宫崎骏的代表作品《天空之城》,在第 95 标签基中占有高权重。

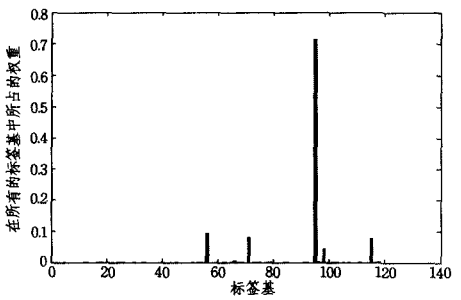


图 4 宫崎骏的动漫电影《天空之城》

表 2 是第 95 标签基中包含标签的列表。该标签基围绕"宫崎骏"构建了核心语义,描述了"天空之城"与其它标签间的一种语义关系。在该标签基中,"吉卜力"是宫崎骏成立的动画片工作室的名称,"高畑勋"是宫崎骏的主要合作者,"悬崖上的金鱼公主、龙猫、魔女宅急便"则是宫崎骏的其他 3 部作品。

表 2 《天空之城》第 95 标签基包含的部分标签

| 宫崎骏 | 动画 | 吉卜力 | 高畑勋 | 动画片 |
|---|---|---|---|---|
| 宫崎骏 | 悬崖上的金鱼公主 | 龙猫 | 天空之城 | 魔女宅急便 |

表 3 是原始网站针对影片《天空之城》基于统计得到的标签列表。从表中可见,"动画、动漫、动画片"都是动画片的同义词,"宫崎駿"是"宫崎骏"的繁体形式,"天空之城"是电影名称。

表 3 《天空之城》原始标签

| 宫崎骏 | 天空之城 | 动画 | 日本 |
|---|---|---|---|
| 动漫 | 动画片 | 电影 | 宫崎駿 |

针对同一部电影《天空之城》对比表 2 与表 3,显然原始的 8 个标签有较大的冗余性并缺乏深层次信息,而第 95 标签基中的标签能更好地帮助用户从多个角度认识《天空之城》。特别是对一些只是初步知道宫崎骏及其作品的用户,标签基中包含的标签能帮助用户从《天空之城》这一部电影扩展开,了解与宫崎骏有关的其他信息,也方便用户进一步通过搜索引擎有的放矢地查找他们感兴趣的内容。

图 5 展示了系数矩阵 $C$ 中另外一种情况。动作演员甄子丹的电影《导火线》在第 19 和 49 标签基上占有较高权重,表明这部电影在标签子空间 $U$ 中主要由 19 和 49 这两组标签基重构而成。

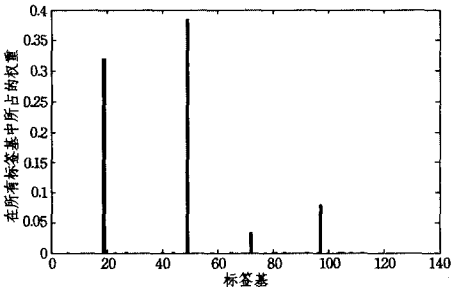


图 5 甄子丹的动作电影《导火线》

第 49 标签基中包含的标签如表 4 所列(篇幅所限,只列举 10 个标签)。该标签基的核心标签是"动作",包含了相关动作电影明星及指示该类影片常见题材的标签。用户通过该标签基可以了解到国内外其他的动作电影明星,进而了解他们的代表作品,方便用户从整体上认识动作电影。

表 4 电影《导火线》的第 49 标签基

| 标签 | 所占权重比例 | 标签 | 所占权重比例 |
|---|---|---|---|
| 动作 | 0.560985 | 施瓦辛格 | 0.015470 |
| 李连杰 | 0.048005 | 甄子丹 | 0.011584 |
| 暴力 | 0.017752 | 李小龙 | 0.010771 |
| 史泰龙 | 0.017674 | 冒险 | 0.009951 |
| 警匪 | 0.015926 | 布鲁斯·威利斯 | 0.009501 |

由此可见,标签子空间 $U$ 反映了多个标签在语义层级上的聚类,揭示了标签彼此之间的联系。用户通过标签基中的一个标签可以看到与之相关的其他标签,从中受到启示,这就方便了用户进一步添加整理自己的收藏,以及在网络中寻找感兴趣的产品资源,增强了用户体验。

### 3.5 矩阵的稀疏性

非负矩阵分解算法在稀疏性和局部性上的优势使得分解得到的基矩阵 $B$ 和系数矩阵 $C$ 具有非负稀疏和结构化的特点。

#### 3.5.1 基矩阵的稀疏性

图 6 展示了从基矩阵 $B$ 得到的标签基"惊悚"中选取的用户使用最多的前 2000 位标签(标签空间 $T$)的权重分布情况。

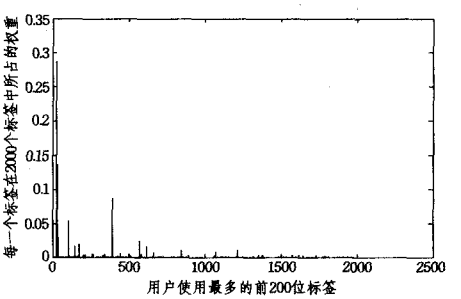


图 6 标签基"惊悚"中 2000 个标签的权重分布

从图 6 可见,除表 1 中列出的相关标签拥有较高的权值外,标签空间 $T$ 中绝大多数标签的权值近似为 0,并且表 1 列举的 13 个标签已占有 89%以上的权重。

#### 3.5.2 系数矩阵的稀疏性

从图 4 宫崎骏的动漫电影《天空之城》可见,该产品分布

(下转第 191 页)

Proc. IEEE International Conference on Neural Networks, 1995 (4): 1942-1948

[8] Eberhart R C, Shi Y. Particle swarm optimization: developments, applications and resources [J]. Pro. Congress on Evolutionary Computation, 2001(1): 81-86

[9] 李宁，邹彤，孙德宝. 带时间窗车辆路径问题的粒子群算法[J]. 系统工程理论与实践，2004，4(4)：130-135

[10] 高尚，韩斌，等. 求解旅行商问题的混合粒子群优化算法[J]. 控制与决策，2004，19(11)：1286 -1289

[11] 莫愿斌，陈德钊，胡上序. 粒子群复形法求解旅行商问题[J]. 浙江大学学报，2007，43(3)：370-375

[12] Cagnina L, Esquivel S, Gallard R. Particle swarm optimization for sequencing problems: A case study[J]. Proceeding of the Congress on Evolutionary Computation, 2004(1): 536-541

[13] 蔡延光，魏明. 一种新型自适应混沌粒子群算法在联盟运输调度问题中的研究[J]. 系统工程，2008，26(8)：32-37

[14] Clerc M. Discrete particle swarm optimization[M]. New Optimization Techniques in Engineering Springer-Verlag, 2004: 219-240

[15] 钟一文，杨建刚，宁正元. 求解 TSP 问题的离散粒子群优化算法[J]. 系统工程理论与实践，2006，6(6)：88-95

[16] Liu B D, Iwamura K. Chance constrained programming with fuzzy parameters [J]. Fuzzy Set and Systems, 1998, 94(2): 227-237

[17] 方述诚，汪定伟. 模糊数学与模糊优化[M]. 北京：科学出版社，1997，7(3)：34-44

[18] 赵晓煜，汪定伟. 供应链中二级分销网络优化设计的模糊机会约束规划模型[J]. 控制理论与应用，2002，19(2)：249-252

(上接第 174 页)

于 120 组标签基上的重构系数中，对应于描述语义“宫崎骏”的第 95 标签基的系数最大，所占权重超过 70%。另外还有 4 个比较大的系数，对应于 4 组标签基，其累加权重比例接近 30%，这也意味着剩余的所有 115 个系数所占的权重近似为 0。类似地，图 5 甄子丹的动作电影《导火线》中最大的 4 个系数的累加权重比例也超过 81%，其余系数接近于 0，权重非常集中。

**结束语** 如何挖掘标签的语义信息，更有效地为用户提供精确的信息服务，是当前研究的热点。本文提出了一种基于非负矩阵分解的语义相关标签挖掘算法，对产品与用户标注标签间的关系进行了挖掘，从标签空间中发现潜在的语义，为用户有效地管理添加网站收藏提供参考。未来将进一步深入测试本方法的性能，并针对 del. icio. us 等网站中网页链接与用户标签间的关系开展研究工作。

## 参 考 文 献

[1] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization [J]. Nature, 1999, 401: 788-791

[2] Lee D D, Seung H. Algorithms for non-negative matrix factorization [J]. Advances in Neural Information Processing Systems, 2001, 13: 556-562

[3] Xu W, Liu X, Gong Y H, et al. Document clustering based on non-negative matrix factorization [C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2003: 267-273

[4] Zanardi V, Capra L. Social ranking: uncovering relevant content using tag-based recommender systems [C]// Proceedings of the 2008 ACM Conference on Recommender Systems. New York: ACM, 2008: 51-58

[5] Wu X, Zhang L, Yu Y. Exploring social annotations for the semantic Web [C]// Proceedings of the 15th International World Wide Web Conference. New York: ACM, 2006: 417-426

[6] 徐雁斐. 基于协同标记的个性化信息服务[D]. 上海：复旦大学，2006

[7] Yeung C A, Gibbins N, Shadbolt N. Understanding the semantics of ambiguous tags in folksonomies [C] // International Workshop on Emergent Semantics and Ontology Evolution at ISWC/ASWC. Busan, South Korea, 2007

[8] Gemmell J, Shepitsen A, Mobasher B. Personalizing Navigation in Folksonomies Using Hierarchical Tag Clustering [J]. Data Warehousing and Knowledge Discovery, 2008, 5182: 196-205

[9] Szomszor M N, Cantador I, Alani H. Correlating user profiles from multiple folksonomies [C]// Proceedings of the 19th ACM Conference on Hypertext and Hypermedia. New York: ACM, 2008: 33-42

[10] Michlmayr E, Cayzer S. Learning user profiles from tagging data and leveraging them for personal (ized) information access [C]// Proceedings of the Workshop on Tagging and Metadata for Social Information Organization at 16th International World Wide Web Conference. New York: ACM, 2007

[11] 刘维湘，郑南宁，游屈波. 非负矩阵分解及其在模式识别中的应用 [J]. 科学通报，2006，51(3)：241-250

[12] Wu Z L, Cheng C W, Li C H. Social and semantics analysis via non-negative matrix factorization [C]// Proceedings of the 17th International World Wide Web Conference. New York: ACM, 2008: 1245-1246

[13] Wu Z L, Li C H. Topic Detection in Online Discussion Using Non-negative Matrix Factorization [C]// Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Workshops. IEEE, 2007: 272-275

[14] Wang D D, Li T, Zhu S H. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization [C]// Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2008: 307-314

[15] Wang D D, Zhu S H, Li T. Integrating clustering and multi-document summarization to improve document understanding [C]// Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM, 2008: 1435-1436

[16] Park S, Lee J H, Kim D H. Multi-document Summarization Based on Cluster Using Non-negative Matrix Factorization [C]// SOFSEM 2007. Theory and Practice of Computer Science. Berlin/Heidelberg: Springer, 2007, 4362: 761-770

[17] Park S, Lee J H, Kim D H. Document Summarization Using Non-negative Matrix Factorization and Relevance Feedback [C]// International Conference on Convergence and Hybrid Information Technology. 2008: 301-306

[18] Park S. Personalized Document Summarization Using Non-negative Semantic Feature and Non-negative Semantic Variable [C]// Intelligent Data Engineering and Automated Learning-IDEAL 2008. 2008, 5326: 298-305

[19] Lee J H, Park S, Ahn C M. Automatic generic document summarization based on non-negative matrix factorization [J]. Information Processing & Management, 2009, 45(1): 20-34

[20] Begelman G, Keller P, Smadja F. Automated Tag Clustering: Improving Search and Exploration in the Tag Space [C]// the 15th International World Wide Web Conference. Edinburgh, Scotland, 2006