软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2015,26(9):2326−2338 [doi: 10.13328/j.cnki.jos.004736]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

增强覆盖度与非相似性的标签选择多样化方法*

汪美玲[1,2], 周 翔[1], 陶秋铭[1], 赵 琛[1]

[1](中国科学院 软件研究所,北京 100190)

[2](中国科学院 研究生院,北京 100049)

通讯作者: 汪美玲, E-mail: meiling@nfs.iscas.ac.cn

摘 要: 标签云是社交网站提供在线资源说明与导航功能的一种流行机制.标签选择即从大量标签中选出有代表性的有限标签,是创建标签云的核心任务.标签选择结果的多样性,是影响用户满意度的一个重要因素.信息覆盖度与标签非相似性是在标签选择中引入多样性的两个主要角度.为了进一步提高标签选择结果的信息覆盖度与标签非相似性,提出了 3 种标签选择方法.在每种方法中,定义了目标函数以同时量化标签集合的信息覆盖度与标签非相似性,并设计了近似算法以求解相应的最大化问题;同时,还分析了近似算法的近似比.利用 CiteULike 网站与 Last.fm 网站的标注数据集,将所提出的方法与已有方法进行了比较.实验结果表明,所提出的方法在信息覆盖度与标签非相似性方面都具有较好的效果.

关键词: 标签云;标签选择;结果多样化;信息覆盖度;非相似性

中图法分类号: TP311

中文引用格式: 汪美玲,周翔,陶秋铭,赵琛.增强覆盖度与非相似性的标签选择多样化方法.软件学报,2015,26(9):2326−2338. http://www.jos.org.cn/1000-9825/4736.htm

英文引用格式: Wang ML, Zhou X, Tao QM, Zhao C. Diversifying tag selection result by improving both coverage and dissimilarity. Ruan Jian Xue Bao/Journal of Software, 2015,26(9):2326−2338 (in Chinese). http://www.jos.org.cn/1000-9825/4736.htm

Diversifying Tag Selection Result by Improving Both Coverage and Dissimilarity

WANG Mei-Ling[1,2], ZHOU Xiang[1], TAO Qiu-Ming[1], ZHAO Chen[1]

[1](Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Tag cloud has been a popular facility used by social networks for online resource summarization and navigation. Tag selection, which aims to select a limited number of representative tags from a large set of tags, is the core task for creating tag clouds. Diversity of tag selection result is an important factor that affects user satisfaction. Information coverage and tag dissimilarity are two major perspectives for introducing diversity in tag selection. To improve information coverage and tag dissimilarity of tag selection result, this paper proposes three new tag selection approaches. In each approach, an objective function is defined to quantify both information coverage and tag dissimilarity of tags, and an approximate algorithm is designed to solve the corresponding maximization problem. Further the approximate ratio for each approximate algorithm is analyzed. The proposed and existing approaches are compared using tagging datasets extracted from the websites of CiteULike and Last.fm. The experimental results show that the new approaches perform better in terms of both information coverage and tag dissimilarity.

Key words: tag cloud; tag selection; result diversification; information coverage; dissimilarity

标注(tagging)是目前众多社交网站的一个基本功能,用户在这些网站中可以使用自定义的标签描述各种信息资源(例如文本、图片和视频).利用用户贡献的标签,社交网站提供了多种访问信息资源的机制,而标签云是其

* 基金项目: 国家自然科学基金(61100067); 中国科学院先导专项(XDA06010600)
收稿时间: 2013-09-22; 修改时间: 2014-07-09; 定稿时间: 2014-09-26

中的重要一种,已被应用于 Flickr、CiteULike、Delicious 和豆瓣等网站中.在标签云中,一组具有代表性的标签以可视化的方式呈现给用户.借助于这组标签,用户不仅可以了解资源整体的信息,还可以通过点击其感兴趣的标签访问与之相关联的个别资源.在标签云的创建过程中,从所有与资源相关联的标签中选择一组有代表性的标签是一个关键步骤[1−4],本文针对该标签选择问题进行研究.

标签云主要通过所包含的一组标签描述资源的信息,因而,为使标签云更富信息量,多样化标签选择结果十分必要.信息覆盖程度与标签非相似性是目前标签选择方法引入多样性的两个主要角度[4],文献[1−3]分别从单一角度提出了若干标签选择方法,而综合考虑信息覆盖度与标签非相似性为标签选择多样化提供了新的思路,我们在文献[4]中对这一思路进行了初步尝试,提出一种新的标签选择方法.与只考虑一种角度的方法相比,该方法在信息覆盖度与标签非相似性方面都表现较好.在本文中,我们仍沿用综合考虑信息覆盖度与标签非相似性的思路,研究不同的标签选择多样化方法,以进一步提高选择结果的多样化程度.

本文将标签选择多样化问题视作特定目标函数在基数约束条件下的最大化问题,在该问题框架下,我们提出 3 种标签选择方法.本文采用 3 种不同的策略来量化标签集合的信息覆盖度与标签非相似性,据此定义了相应的目标函数,并依据目标函数的性质设计并实现了最大化问题的近似求解算法,同时还分析了相应算法的近似比.利用 CiteULike 网站与 Last.fm 网站的标注数据集,我们将所提出的方法与已有方法进行了比较.实验结果表明:所提出的方法在信息覆盖度与标签非相似性方面都表现较好;与文献[4]中的方法相比表现更好,进一步提高了标签选择结果的多样化程度.

本文第 1 节介绍相关工作.第 2 节介绍预备知识.第 3 节描述所提出的 3 种方法.第 4 节给出在 CiteULike 的标注数据集与 Last.fm 的标注数据集上将所提出方法与已有方法进行比较的实验结果.最后给出全文总结.

1 相关工作

多样化问题在 Web 搜索[5−8]、数据库[9−11]、文档摘要[12−14]、推荐[15−17]等领域得到了广泛研究.文献[18−22]对多样化的一般问题进行了研究.

目前已有的标签选择方法在不同程度上考虑了多样化:

- 最早的标签选择方法(以下简称 POP)根据标签所标注资源的总数(即流行度)选择标签.方法 POP 的选择结果通常包含很多相似标签,覆盖的信息也不够全面;
- 文献[1]中所提出的标签选择方法(以下简称 USE)以降低选择结果中不同标签之间的资源重叠(相似性)为目标,基于标签效能的概念选择标签;
- 文献[3]中所提出的标签选择方法之一(以下简称 COV)以最大化选择结果所覆盖的资源数为目标,根据相对于当前已选择标签所覆盖的新资源的数目选择标签;
- 文献[2]中提出两种显式考虑多样化的方法:一种方法(以下简称 POP+DIS)是根据标签流行度和与当前已选择标签的最短距离的凸组合选择标签,以使选择结果中的标签尽量互不相似;另一种方法(以下简称 NOV)是根据相对于当前已选择标签的新颖性选择标签.当方法 NOV 在最大程度上强调新颖性时,其求得与方法 COV 相同的选择结果;
- 我们在文献[4]中所提出的标签选择方法(以下简称 COV+SUSE)以提高选择结果所覆盖的资源数并降低选择结果中不同标签之间的资源重叠为目标,根据相对于当前已选择标签所覆盖的新资源的数目和所增加的标签效能的凸组合选择标签.

2 预备知识

2.1 社交标注系统

研究者通常将社交标注系统(social tagging system)[23,24]建模为三分超图 $G=(U\cup T\cup R,E)$,其中,U,T 和 R 分别为用户集、标签集和资源集且互不相交,而边集 $E\subseteq U\times T\times R$.边 $(u,t,r)\in E$ 表示用户 u 使用标签 t 标注资源 r.

- 对任意的 $t\in T$,令 $R(t)=\{r\in R|\exists u(u\in U\wedge(u,t,r)\in E)\}$,表示标签 t 所标注的全部资源的集合;
- 对任意的 $r\in R$,令 $T(r)=\{t\in T|\exists u(u\in U\wedge(u,t,r)\in E)\}$,表示标注资源 r 的全部标签的集合.

2.2 次模函数及单调函数

设 X 为有限集合.令 $\mathcal{F}:2^X\to\mathbb{R}$.若对任意的 $A\subseteq B\subset X$ 及任意的 $x\in X\backslash B$,$\mathcal{F}(A\cup\{x\})-\mathcal{F}(A)\geqslant\mathcal{F}(B\cup\{x\})-\mathcal{F}(B)$成立,则称函数$\mathcal{F}$满足次模性(submodularity)[25],其中,称$\mathcal{F}(A\cup\{x\})-\mathcal{F}(A)$为给定集合 A 时 x 关于$\mathcal{F}$的边际收益(marginal returns).次模函数具有边际收益递减的性质.

易证次模函数满足如下性质:

引理 1. 设 X 为有限集合,$\alpha_1,\alpha_2,\ldots,\alpha_n\in\mathbb{R}$且$\alpha_i\geqslant 0(0\leqslant i\leqslant n)$,$\mathcal{F}_1,\mathcal{F}_2,\ldots,\mathcal{F}_n:2^X\to\mathbb{R}$为次模函数.令$\mathcal{F}:2^X\to\mathbb{R}$,且对任意的 $A\subseteq X$,定义 $\mathcal{F}(A)=\sum_{i=1}^{n}\alpha_i\cdot\mathcal{F}_i(A)$,则$\mathcal{F}$也是次模函数.

令$\mathcal{G}:2^X\to\mathbb{R}$.若对任意的 $A\subseteq B\subseteq X$,$\mathcal{G}(A)\leqslant\mathcal{G}(B)$成立,则称函数$\mathcal{G}$满足单调性(monotonicity)[25].

2.3 距离函数

设 X 为集合.令 $d:X\times X\to\mathbb{R}$.若对任意的 $x,y\in X$,d 满足如下性质:

(1) $d(x,y)\geqslant 0$;

(2) $d(x,y)=0$ 当且仅当 $x=y$;

(3) $d(x,y)=d(y,x)$,

则称 d 为 X 上的距离函数(distance function)[26].若对任意的 $x,y,z\in X$,距离函数 d 还满足 $d(x,z)\leqslant d(x,y)+d(y,z)$,则称 d 为度量距离函数(metric distance function)[26].

3 提出的方法

定义 1(标签选择多样化问题). 令 $G=(U\cup T\cup R,E)$为社交标注系统,目标函数$\mathcal{D}:2^T\to\mathbb{R}$量度 T 的子集的多样性.给定一个正整数 k,标签选择多样化求解一个集合 S,$S\subseteq T$,使得$\mathcal{D}(S)$最大,且$|S|=k$.

在该问题框架下,我们提出 3 种标签选择多样化方法.在每种方法中,采用相应的、同时量化信息覆盖度与标签非相似性的策略定义了目标函数,并依据目标函数的性质设计了近似求解算法,同时分析了近似算法的近似比.下面详细描述这 3 种方法.

3.1 方法COV-SIM

3.1.1 目标函数定义

考虑到信息覆盖度与标签非相似性可能互相冲突,将本方法的目标函数$\mathcal{D}_1:2^T\to\mathbb{R}$定义为如下凸组合:

$$\mathcal{D}_1(S)=\lambda\cdot\mathcal{H}_1(S)+(1-\lambda)\cdot\mathcal{L}_1(S).$$

其中,$\mathcal{H}_1:2^T\to\mathbb{R}$是量化信息覆盖度的子目标函数;$\mathcal{L}_1:2^T\to\mathbb{R}$是量化标签非相似性的子目标函数;参数$\lambda\in[0,1]$,用于控制$\mathcal{H}_1$与$\mathcal{L}_1$的权重.下面详细介绍函数$\mathcal{H}_1$与$\mathcal{L}_1$的定义.

(1) 在社交标注系统中,可以将每个资源看作一块信息,因此,一个标签集合中所有标签所标注的资源的总数是该标签集的信息覆盖程度的一种量化.由于不同标签可能标注相同的资源,随着一个标签集合中元素的增多,该标签集中所有标签所标注的资源的总数将会增加,但是增加的速度将会下降.可见,函数$\mathcal{H}_1$应该满足单调次模性.使用文献[4]中的覆盖函数作为$\mathcal{H}_1$的定义,即:对任意的 $S\subseteq T$:

$$\mathcal{H}_1(S)=\sqrt{\left|\bigcup_{t\in S}R(t)\right|}.$$

$\mathcal{H}_1$ 满足单调次模性[4].在$\mathcal{H}_1$ 的定义中,平方根函数一方面用于保证$\mathcal{H}_1$ 满足单调次模性,另一方面用于避免$\mathcal{H}_1$与$\mathcal{L}_1$相差过多.

(2) 因为标签的作用是标注资源,所以借助所标注的资源集来刻画标签是比较自然的想法.这里,首先将标

签的相似性定义为其所标注的资源集的 Jaccard 相似度[27],即:对任意的 $t_i,t_j\in T$ 有

$$s_J(t_i,t_j)=\frac{|R(t_i)\cap R(t_j)|}{|R(t_i)\cup R(t_j)|}.$$

进一步地,我们定义$\mathcal{L}_1$如下:对任意的 $S\subseteq T$:

$$\mathcal{L}_1(S)=-\sum_{t_i,t_j\in S,t_i\neq t_j}s_J(t_i,t_j)$$

表示选择标签时,使结果集中标签相似性的和尽量的小.

基于如上定义,目标函数$\mathcal{D}_1$满足如下性质:

命题 1. $\mathcal{D}_1$是非单调次模函数.

证明:根据定义,$\mathcal{L}_1(S)=-\sum_{t_i,t_j\in S,t_i\neq t_j}s_J(t_i,t_j)$是次模函数.根据文献[4]中的结论,$\mathcal{H}_1(S)=\sqrt{\left|\bigcup_{t\in S}R(t)\right|}$是单调次模函数.根据引理 1,$\mathcal{D}_1(S)=\lambda\cdot\mathcal{H}_1(S)+(1-\lambda)\cdot\mathcal{L}_1(S)(0\leqslant\lambda\leqslant1)$是次模函数.根据定义,$\mathcal{D}_1(S)$不是单调函数. □

3.1.2 近似算法与近似比

因为以$\mathcal{D}_1$为目标函数的标签选择多样化问题包含了最大覆盖问题(maximum coverage problem)[28],所以该标签选择多样化问题是 NP-hard 的.我们采用算法 1 近似求解该问题,该算法包含一种贪心算法 1_1 和一种局部搜索算法 1_2.

- 算法 1_1 从空集起逐步构造一个包含 k 个标签的集合:在每个循环中,该算法从剩余标签中选择给定当前结果集时关于$\mathcal{D}_1$的边际收益最大的标签,将其添加到结果集中,并继续下个循环;
- 算法 1_2 首先利用关于$\mathcal{H}_1$的贪心策略构造一个初始结果集,然后从这个初始结果集开始进行局部搜索,即:不断交换当前结果集内、外的各一个标签以改进结果集的目标函数值,直到没有任何可执行的交换为止.在交换过程中,换出标签 t_{out} 是从结果集内、按照对目标函数值的贡献从小到大的顺序选择,换入标签 t_{in} 是从结果集外选择的、与当前换出标签交换后使得目标函数值最大的标签.

算法 1 是上述局部搜索算法和贪心算法的结合:首先,以 T 和 k 为输入执行局部搜索算法 1_2,求得包含 k 个标签的集合 S_1;然后,以 $T\backslash S_1$ 和 k 为输入执行贪心算法 1_1,求得包含 k 个标签的集合 S_2;最后,将 S_1 与 S_2 中目标函数值较大的一个作为结果返回.

算法 1.

输入:标签集合 T,正整数 k;

输出:$S\subseteq T$ 且$|S|=k$.

1: $S_1=P_{localsearch}(T,k)$

2: $S_2=P_{greedy}(T\backslash S_1,k)$

3: if $\mathcal{D}_1(S_1)\geqslant\mathcal{D}_1(S_2)$ then $S=S_1$ else $S=S_2$ end if

4: return S

算法 1_1. P_{greedy}.

输入:标签集合 T,正整数 k;

输出:$S\subseteq T$ 且$|S|=k$.

1: $S=\varnothing$

2: while $|S|<k$ do

3: $t^*=\arg\max_{t\in T\backslash S}(\mathcal{D}_1(S\cup\{t\})-\mathcal{D}_1(S))$

4: $S=S\cup\{t^*\}$

5: end while

6: return S

算法 1_2. $P_{localsearch}$.

输入:标签集合 T,正整数 k;

输出:$S\subseteq T$ 且$|S|=k$.

1: $S=\varnothing$

2: while $|S|<k$ do

3: $t^* = \arg\max_{t\in T\setminus S}(\mathcal{H}_1(S\cup\{t\})-\mathcal{H}_1(S))$

4: $S=S\cup\{t^*\}$

5: end while

6: $T_{out}=S$;

7: while $|T_{out}|>0$ do

8: $t_{out} = \arg\max_{t\in T_{out}} \mathcal{D}_1(S\setminus\{t\})$

9: $t_{in} = \arg\max_{t\in T\setminus S} \mathcal{D}_1(S\setminus\{t_{out}\}\cup\{t\})$

10: if $\mathcal{D}_1(S\setminus\{t_{out}\}\cup\{t_{in}\})>\mathcal{D}_1(S)$ then

11: $S=S\setminus\{t_{out}\}\cup\{t_{in}\}$

12: $T_{out}=S$

13: else

14: $T_{out}=T_{out}\setminus\{t_{out}\}$

15: end if

16: end while

17: return S

定理 1[29]. 设$\mathcal{D}:2^T\to\mathbb{R}$为非单调次模函数,对任意的 $S\subseteq T$,都有$\mathcal{D}(S)\geqslant 0$ 且$\mathcal{D}(\varnothing)=0$.若使用如下算法框架构造集合 $\hat{S}$:

步骤 1:使用局部搜索算法求得 $S_1,S_1\subseteq T$ 且$|S_1|=k$:初始化 S_1,使其包含 T 中 k 个标签,当存在 $t_1\in S_1,t_2\in T\setminus S_1$ 使得$\mathcal{D}(S_1\setminus\{t_1\}\cup\{t_2\})>\mathcal{D}(S_1)$时,$S_1=S_1\setminus\{t_1\}\cup\{t_2\}$;

步骤 2:使用贪心算法求得 $S_2,S_2\subseteq T\setminus S_1$ 且$|S_2|=k$:初始化 S_2 为空集,从当前 $T\setminus(S_1\cup S_2)$中选出给定 S_2 时关于$\mathcal{D}$的边际收益最大的标签添加到 S_2 中,直到$|S_2|=k$ 为止;

步骤 3:取 $\hat{S}$ 为 S_1 与 S_2 中目标函数值较大者.

则 $\mathcal{D}(\hat{S})\geqslant\frac{1}{4}\mathcal{D}(S^*)$,其中, $S^*\in\arg\max_{S\subseteq T,|S|=k}\mathcal{D}(S)$.

在本方法中,算法 1 是定理 1 中所述的算法框架的一个实例;根据命题 1,$\mathcal{D}_1$ 是非单调次模函数;根据定义,$\mathcal{D}_1(\varnothing)=0$;在给定的社交标注系统中,当$\lambda$取适当的值,使得对任意的 $S\subseteq T,\mathcal{D}_1(S)\geqslant 0$ 成立时,本方法具有如定理 1 所述的理论保证,即,算法 1 所生成的标签选择结果的目标函数值至少是最优目标函数值的 0.25 倍.在实验中我们发现:在一个给定的社交标注系统中,大多数λ的取值都能使上述非负条件成立.

另外,在定理 1 所述的算法框架中,步骤 1 的运行效率决定了整个算法的运行效率;而一般来说,局部搜索算法在最坏情况下的运行时间为指数级[30].为了提高步骤 1 的运行效率,我们实例化该步骤时在算法 1_2 中采用如下策略:

(1) 在初始化阶段,利用关于函数$\mathcal{H}_1$ 的贪心算法构造初始结果集.$\mathcal{H}_1$ 与方法 COV 的目标函数只相差平方根函数,算法 1_2 中关于$\mathcal{H}_1$的贪心算法与方法 COV 中的贪心算法求得相同的结果集,且该结果集在信息覆盖度方面表现较好[3].

(2) 在交换阶段,尽可能选择使目标函数值改进较大的标签对进行交换.

基于上述策略,算法 1_2 所需进行的交换次数大大减少.在实验中我们发现:在实际运行中,该交换次数与选择标签数 k 在同一个量级.

3.2 方法COV+DIS

3.2.1 目标函数定义

将本方法的目标函数$\mathcal{D}_2$:$2^T\to\mathbb{R}$定义为如下凸组合:

$$\mathcal{D}_2(S)=\lambda\cdot\mathcal{H}_1(S)+(1-\lambda)\cdot\sum_{t_i,t_j\in S}d_J(t_i,t_j),$$

其中,

- d_J:$T\times T\to\mathbb{R}$是标签之间的距离函数,对任意的 $t_i,t_j\in T$: $d_J(t_i,t_j)=1-\dfrac{|R(t_i)\cap R(t_j)|}{|R(t_i)\cup R(t_j)|}$;
- 参数$\lambda\in[0,1]$,用于控制$\mathcal{H}_1$与 $\sum_{t_i,t_j\in S}d_J(t_i,t_j)$ 的权重.

在本方法中,我们采用与方法 COV-SIM 中相同的子目标函数$\mathcal{H}_1$量化信息覆盖度,但是对于标签非相似性,我们直接采用标签集合中标签之间的距离之和量化[19−21].这与在方法 COV-SIM 中所采用的子目标函数$\mathcal{L}_1$不同,后者使标签集合中标签相似性之和尽量地小.

3.2.2 近似算法与近似比

以$\mathcal{D}_2$为目标函数的标签选择多样化问题同样是 NP-hard 问题,我们采用算法 2 近似求解该问题.算法 2 是一个贪心算法,与算法 1_1 的不同之处在于:在每个循环中,选择标签的标准不是给定当前结果集时关于$\mathcal{D}_2$的边际收益,而是两部分的凸组合:一部分是给定当前结果集时关于$\mathcal{H}_1$的边际收益的$\frac{1}{2}$倍,另一部分是与当前已选择标签的距离的和.

算法 2.

输入:标签集合 T,正整数 k;

输出:$S\subseteq T$ 且$|S|=k$.

1: $S=\varnothing$

2: while $|S|<k$ do

3: $t^*=\arg\max_{t\in T\setminus S}\left(\lambda\cdot\frac{1}{2}(\mathcal{H}_1(S\cup\{t\})-\mathcal{H}_1(S))+(1-\lambda)\cdot\sum_{t_i\in S}d_J(t,t_i)\right)$

4: $S=S\cup\{t^*\}$

5: end while

6: return S

定理 2[21]. 设$\mathcal{D}$:$2^T\to\mathbb{R}$定义如下:对任意的 $S\subseteq T$, $\mathcal{D}(S)=\lambda\cdot\mathcal{H}(S)+(1-\lambda)\cdot\sum_{t_i,t_j\in S}d(t_i,t_j)$,其中,$0\leqslant\lambda\leqslant1$;而$\mathcal{H}$:$2^T\to\mathbb{R}$为单调次模函数,且$\mathcal{H}(\varnothing)=0$,$d$:$T\times T\to\mathbb{R}$为度量距离函数.若使用如下贪心算法构造集合 $\hat{S}$:

步骤 1:初始化 $\hat{S}$ 为空集;

步骤 2:针对当前的 $\hat{S}$,选择一个标签 $t, t\in T\setminus\hat{S}$ 且使 $\lambda\cdot\frac{1}{2}(\mathcal{H}(\hat{S}\cup\{t\})-\mathcal{H}(\hat{S}))+(1-\lambda)\cdot\sum_{t_i\in\hat{S}}d(t,t_i)$ 最大,并将 t 添加到 $\hat{S}$ 中;

步骤 3:重复执行步骤 2,直到 $|\hat{S}|=k$ 为止.

则 $\mathcal{D}(\hat{S})\geqslant\frac{1}{2}\mathcal{D}(S^*)$, 其中, $S^*\in\arg\max_{S\subseteq T,|S|=k}\mathcal{D}(S)$.

在本方法中,算法 2 是定理 2 中所述的贪心算法的一个实例.根据文献[4]中的结论,$\mathcal{H}_1$满足单调次模性且$\mathcal{H}_1(\varnothing)=0$.根据文献[31]中的结论,$d_J$是度量距离函数.综上,本方法具有如定理 2 所述的理论保证,即,算法 2 所生成的标签选择结果的目标函数值至少是最优目标函数值的 0.5 倍.

3.3 方法GAIN

3.3.1 目标函数定义

标签云通过所包含的标签为用户提供资源信息说明与资源空间导航的功能.本方法从标签给用户带来的效用的角度来定义目标函数,其基本思想借鉴于文献[12].

我们首先引入效用函数 $v:R\times 2^T\to\mathbb{R}$,其描述一个标签集合相对于一个资源的效用:对任意的 $S\subseteq T$ 和任意的 $r\in R$:

$$v(r,S)=\begin{cases}0, & \text{如果对任意的}t\in S, r\notin R(t)\\ \lambda^{x-1}, & \text{如果}|\{t\in S: r\in R(t)\}|=x\text{且}x\geqslant 1\end{cases}.$$

其中,$\lambda\in(0,1]$.进一步地,我们定义一个标签集合相对于整个资源集的效用为该标签集相对于所有资源的效用之和:对于任意的 $S\subseteq T$:

$$\mathcal{D}_3(S)=\sum_{r\in R}v(r,S).$$

将函数$\mathcal{D}_3$作为本方法的目标函数.

目标函数$\mathcal{D}_3$符合我们提高标签选择结果的资源覆盖度并降低选择结果中标签相似性的意图.

- 首先,从效用函数的定义来看,若标签集合 S 中存在标签覆盖资源 r,则 S 对 r 的效用大于 0;否则,S 对 r 的效用为 0.因而,依据$\mathcal{D}_3$ 选择效用较大的标签集合时,我们会比较倾向于选择覆盖更多资源的标签集合;
- 其次,在效用函数的定义中,当资源 r 被标签集合 S 中的标签覆盖时,S 的效用是随着 r 被覆盖次数的增加而减少的.因而,依据$\mathcal{D}_3$ 选择效用较大的标签集合时,我们很自然地要使选择结果中的标签尽量覆盖不同的资源集,这实际上就降低了选择结果中标签的相似性;
- 最后,我们在效用函数的定义中使用了参数λ,该参数用于权衡信息覆盖度与标签非相似性:当λ变大时,信息覆盖度的权重变大而标签非相似性的权重变小;当λ=1 时,对任意标签集合 S,函数$\mathcal{D}_3(S)$计算的是 S 所覆盖资源的数量,从而依据$\mathcal{D}_3$选择标签时仅考虑了选择结果的资源覆盖度.

另外,目标函数$\mathcal{D}_3$在λ=1 时满足如下性质:

命题 2. 当λ=1 时,$\mathcal{D}_3$为单调次模函数.

证明:当λ=1 时,对于任意的 $S\subseteq T$ 和任意的 $r\in R$:

$$v(r,S)=\begin{cases}0, & \text{如果对任意的}t\in S, r\notin R(t)\\ 1, & \text{如果}|\{t\in S: r\in R(t)\}|=x\text{且}x\geqslant 1\end{cases}.$$

此时, $\mathcal{D}_3(S)=\sum_{r\in R}v(r,S)=\left|\bigcup_{t\in S}R(t)\right|$.根据定义,$\mathcal{D}_3$ 是单调次模函数. □

3.3.2 近似算法与近似比

以$\mathcal{D}_3$为目标函数的标签选择多样化问题同样是一个 NP-hard 问题,本方法使用算法 3 近似求解该问题.算法 3 为贪心算法,与算法 1_1 的区别在于目标函数不同.

算法 3.

输入:标签集合 T,正整数 k;

输出:$S\subseteq T$ 且$|S|=k$.

1: $S=\varnothing$

2: while $|S|<k$ do

3: $t^*=\arg\max_{t\in T\setminus S}(\mathcal{D}_3(S\cup\{t\})-\mathcal{D}_3(S))$

4: $S=S\cup\{t^*\}$

5: end while

6: return S

定理 3[32]. 设$\mathcal{D}:2^T\to\mathbb{R}$为单调次模函数且$\mathcal{D}(\varnothing)=0$.若使用如下贪心算法构造集合$\hat{S}$:

初始化$\hat{S}$为空集,从当前$T\setminus\hat{S}$中选出给定$\hat{S}$时关于$\mathcal{D}$的边际收益最大的标签添加到$\hat{S}$中,直到$|\hat{S}|=k$为止.则$\mathcal{D}(\hat{S})\geqslant(1-1/e)\mathcal{D}(S^*)$, 其中, $S^*\in\arg\max_{S\subseteq T,|S|=k}\mathcal{D}(S)$.

在本方法中,算法 3 是定理 3 中所述的贪心算法的一个实例;根据命题 2,当λ=1 时,$\mathcal{D}_3$满足单调次模性且$\mathcal{D}_3(\varnothing)$=0,所以此时本方法具有如定理 3 所述的理论保证;而当 0<λ<1 时,不能得到相同的结论.一般情形下,算法 3 不一定具有理论上的性能保证.

4 实 验

在从 CiteULike 网站(http://www.citeulike.org)与 Last.fm 网站(http://www.last.fm/)抽取的标注数据集上我们将本文所提出的标签选择方法与已有方法进行了比较.

4.1 数据准备

从 CiteULike 网站的整个标注数据集(http://www.citeulike.org/faq/data.adp)中抽取出一个子数据集用于实验.CiteULike 网站的整个标注数据集包括 17 481 632 个标注,每个标注由一个用户的 id、一个标签、一个文章(article)的 id 和一些其他信息组成.该数据集一共包括 761 674 个不同标签,其中,software 是流行度较高的一个标签.我们将 software 看作主题标签[33],抽取出与 software 相关的子数据集:首先抽取出 software 所标注的所有资源,然后抽取出所有与这些资源相关联的标注(包含 software 的标注除外),并将所抽取到的标注集作为与 software 相关的子数据集.为了使该数据集所包含的标签更有意义,我们将包含噪音标签(例如符号、停词)的标注删除,之后,该数据集的统计信息见表 1 中第 1 行所示.

Table 1 Dataset statistics

表 1 数据集的统计信息

数据集	标注数	标签数	资源数	用户数
CiteULike	242 008	26 651	16 854	8 097
Last.fm	186 479	9 749	12 523	1 892

我们用于实验的 Last.fm 标注数据集是由明尼苏达大学(University of Minnesota)的 GroupLens 研究实验室(http://grouplens.org/)从 Last.fm 网站抽取并发布(http://grouplens.org/datasets/hetrec-2011/)的,该数据集包括 186 479 个标注,每个标注由一个用户的 id、一个标签、一个艺术家(artist)的 id 和一些时间信息组成.该数据集的统计信息见表 1 中第 2 行.

图 1 描述了在 CiteULike 数据集与 Last.fm 数据集中标签关于标注资源数的分布.

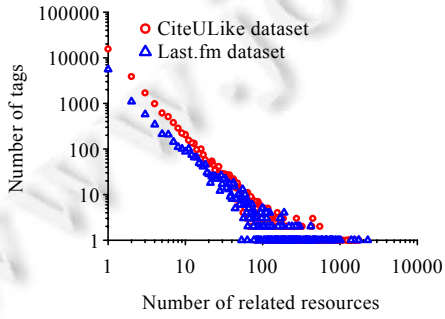


Fig.1 Distribution of tags across the number of related resources

图 1 标签关于标注资源数的分布

如图 1 所示:在这两个数据集中,大多数标签所标注的资源的数目都较小;并且在 Last.fm 数据集中,具有特定数量标签的资源的数目更小.

4.2 评价标准

在已有的评价标签选择结果的标准中,覆盖度与重叠度较好地反映了信息覆盖度与标签非相似性[4].本文使用覆盖度与重叠度评价各个标签选择方法的结果的多样性.

标签选择结果 S 的覆盖度量度 S 中所有标签所覆盖的资源占资源集 R 中全部资源的比例.在文献[2]中,将 S 的覆盖度定义为

$$coverage(S)=\frac{\left|\bigcup_{t\in S}R(t)\right|}{|R|}.$$

$coverage(S)$越大,S 所覆盖的信息越多.

标签选择结果 S 的重叠度量度 S 中所包含的标签的冗余程度.在文献[1]中,将 S 的重叠度定义为

$$overlap(S)=\frac{\sum_{t_i,t_j\in S,t_i\neq t_j}s_J(t_i,t_j)}{|S|\cdot(|S|-1)/2},$$

其中,$s_J(t_i,t_j)=\frac{|R(t_i)\cap R(t_j)|}{|R(t_i)\cup R(t_j)|}$.$overlap(S)$越小,$S$ 所包含的标签越不相似.

4.3 实验结果

我们分别实现了在本文第 3 节所提出的 3 种标签选择方法(COV-SIM,COV+DIS 和 GAIN)、在文献[4]中所提出的方法 COV+SUSE 以及相关工作中介绍的另外 5 种标签选择方法(POP,USE,COV,POP+DIS 和 NOV).我们在第 4.1 节中所介绍的 CiteULike 数据集与 Last.fm 数据集上比较这 9 种方法的选择结果的多样性.使用 MySQL Server 5.0 存储数据集,基于 Java 1.6 实现各个标签选择方法并计算标签选择结果的覆盖度与重叠度.在各个方法的实现和实验中,令选择的标签数 k=30.对于方法 NOV,将一般情形下的函数$\gamma()$定义为

$$\gamma(n_{r,S})=\frac{|T(r)|-n_{r,S}}{|T(r)|},$$

其中,$r\in R$,S 中包含当前已选择的标签,$n_{r,s}=|\{t\in S:r\in R(t)\}|$(方法 NOV 的定义详见文献[2];当方法 NOV 在最大程度上强调新颖性时,其选择结果与方法 COV 的选择结果相同,此处考虑该方法的一般情形).

根据第 4.2 节所介绍的定义,计算方法 POP,USE,COV 和 NOV 的选择结果的覆盖度与重叠度.令方法 COV-SIM,COV+DIS,GAIN,POP+DIS 和 COV+SUSE 中参数λ的值从 0.05 开始,以间隔 0.05 增加到 0.95(方法 POP+DIS 与方法 COV+SUSE 中,参数λ的定义见文献[2,4]),并计算当λ取特定值时各个方法的选择结果的覆盖度与重叠度.图 2、图 3 分别显示了在 CiteULike 数据集与 Last.fm 数据集上这 9 种方法的选择结果的覆盖度与重叠度随λ值的变化.

从图 2 与图 3 可见:

(1) 与已有的 5 种方法 POP,USE,COV,POP+DIS 和 NOV 相比,在 CiteULike 数据集与 Last.fm 数据集上,对于λ的很多取值,本文所提出的 3 种方法在覆盖度与重叠度方面都表现较好:

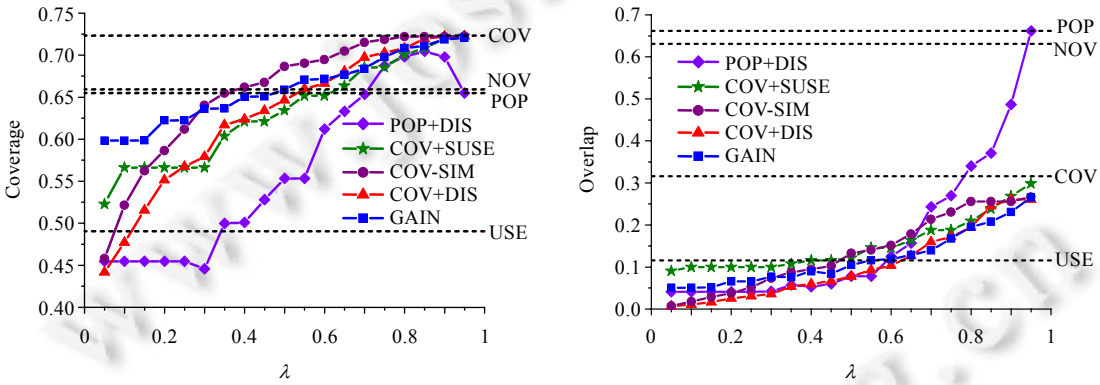
- 如图 2(a)和图 3(a)所示:在 CiteULike 数据集与 Last.fm 数据集上,方法 COV 的选择结果的覆盖度都最大;当λ从 0.05 增加到 0.95 时,本文所提出的 3 种方法的选择结果的覆盖度都在不断增长,且当λ增长到一定值之后,与已有的 5 种方法相比,这 3 种方法的选择结果的覆盖度都更加接近于方法 COV 的选择结果的覆盖度;
- 如图 2(b)和图 3(b)所示:在 CiteULike 数据集与 Last.fm 数据集上,当λ从 0.05 增加到 0.95 时,与已有的 4 种方法 POP,COV,POP+DIS 和 NOV 相比,本文所提出的 3 种方法的选择结果的重叠度都始终处于较低水平,只当λ增长到一定值之后才高于方法 USE 的选择结果的重叠度.

(2) 与我们在文献[4]中所提出的方法 COV+SUSE 相比,在 CiteULike 数据集与 Last.fm 数据集上,对于λ的很多取值,本文所提出的 3 种方法在覆盖度与重叠度方面都表现较好:

- 在 CiteULike 数据集上,如图 2(a)、图 2(b)所示:对于绝大多数的λ值,与方法 COV+SUSE 相比,本文所提出的 3 种方法的选择结果都具有更高的覆盖度与更低的重叠度;
- 在 Last.fm 数据集上,如图 3(a)所示:当λ从 0.05 增加到 0.95 时,最初方法 COV+SUSE 的选择结果的覆盖度高于本文所提出的 3 种方法的选择结果的覆盖度,之后,本文所提出的 3 种方法的选择结果的覆盖度快速逼近于方法 COV+SUSE 的选择结果的覆盖度,且当λ增长到一定值之后,这 3 种方法的选择结果的覆盖度高于方法 COV+SUSE 的选择结果的覆盖度.如图 3(b)所示:对于所有的λ值,本文所提出的方法 COV+DIS 和方法 GAIN 的选择结果的重叠度都远低于方法 COV+SUSE 的选择结果的重叠度,对于除了 0.95 之外的所有λ值,本文所提出的方法 COV-SIM 的选择结果的重叠度都远低于方法 COV+SUSE 的选择结果的重叠度.

(3) 本文所提出的 3 种方法各有特点:

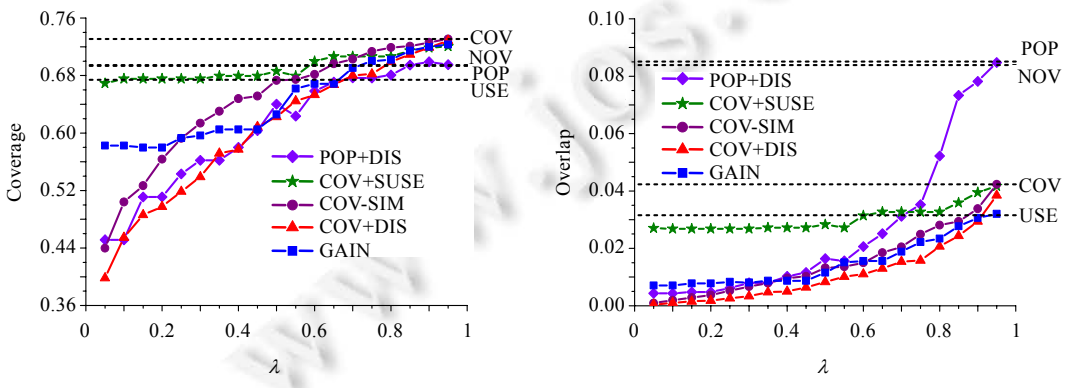
- COV-SIM 在 CiteULike 数据集与 Last.fm 数据集上都在覆盖度方面表现较好;
- COV+DIS 在 CiteULike 数据集与 Last.fm 数据集上都在重叠度方面表现较好;
- GAIN 在 CiteULike 数据集上在重叠度方面表现较好,在 Last.fm 数据集上在覆盖度方面表现较好.



(a) 覆盖度随λ值的变化 (b) 重叠度随λ值的变化

Fig.2 Coverage and overlap of nine approaches for increasing λ on the CiteULike dataset

图 2 CiteULike 数据集上 9 种方法的覆盖度与重叠度随λ值的变化



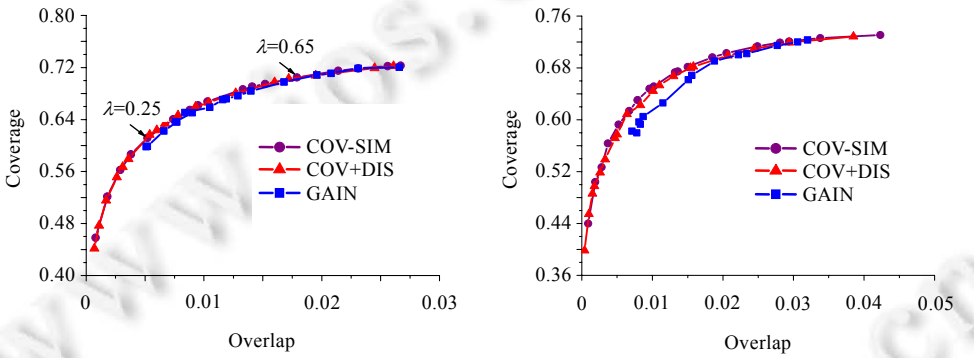
(a) 覆盖度随λ值的变化 (b) 重叠度随λ值的变化

Fig.3 Coverage and overlap of nine approaches for increasing λ on the Last.fm dataset

图 3 Last.fm 数据集上 9 种方法的覆盖度与重叠度随λ值的变化

说明:在本文所提出的 3 种标签选择方法中,参数λ用于权衡信息覆盖度与标签非相似性,即:当λ变大时,信息覆盖度的权重变大而标签非相似性的权重变小;相应地,在我们的实验中,当λ从 0.05 增加到 0.95 时,标签选择

结果的覆盖度变大而重叠度变小.我们可以为所提出方法确定参数λ的具体取值,使选择结果同时具有较好的覆盖度与重叠度.图 4 显示了所提出的 3 种方法在 CiteULike 数据集与 Last.fm 数据集上相对于λ的覆盖度-重叠度权衡曲线,其中,曲线上的点分别以λ取特定值时方法的选择结果的重叠度与覆盖度为横坐标与纵坐标.从图 4 可见,这 3 种方法在 CiteULike 数据集与 Last.fm 数据集上的覆盖度-重叠度权衡曲线都明显具有弯曲部分.例如如图 4(a)中方法 COV-SIM 在 CiteULike 数据集上的覆盖度-重叠度权衡曲线:当λ从 0.05 增加到 0.25 时,权衡曲线急剧上升,即,覆盖度增加较快同时重叠度增加较慢;当λ从 0.25 增加到 0.65 时,权衡曲线变弯曲,即,覆盖度增加变慢同时重叠度增加变快;当λ从 0.65 增加到 0.95 时,权衡曲线趋于平稳,即,覆盖度增加更慢同时重叠度增加更快.可见,当为这 3 种方法取其权衡曲线的弯曲部分所对应的λ值时,所得到的选择结果的覆盖度接近最大、重叠度接近最小.我们一般可以通过两种方式确定所提出方法在特定数据集上的参数λ的具体取值:一是借鉴文献[2,4]中的做法,直接取权衡曲线的弯曲部分所对应的λ值中的一个,例如,为方法 COV-SIM 在 CiteULike 数据集上取λ=0.5;二是根据用户反馈的偏好信息分析出权衡曲线的弯曲部分所对应的λ值中的最优者.我们将后者作为未来工作的一部分.



(a) CiteULike 数据集上的覆盖度-重叠度权衡曲线 (b) Last.fm 数据集上的覆盖度-重叠度权衡曲线

Fig.4 Coverage-Overlap tradeoff curves of our approaches for λ on the datasets of CiteULike and Last.fm

图 4 CiteULike 数据集与 Last.fm 数据集上所提方法相对于λ的覆盖度-重叠度权衡曲线

5 结 论

本文在最大化问题框架下提出了 3 种标签选择多样化方法,以增强标签选择结果的信息覆盖度与标签非相似性.在各个方法中,采用不同的策略以量化标签集合的信息覆盖度与标签非相似性,并定义了目标函数.针对这些目标函数,设计了近似算法以解决相应的最大化问题,并分析了近似算法的近似比.实验结果表明:与已有方法相比,本文所提出的方法的选择结果具有较好的信息覆盖度与标签非相似性.未来,我们计划进行如下 3 方面工作:一是在更多的数据集上评估本文所提出的方法;二是对标签选择结果的多样性的评价方法作深入研究;三是根据用户反馈的偏好信息为所提出方法的参数λ确定较合适的取值.

References:

[1] Hassan-Montero Y, Herrero-Solana V. Improving tag-clouds as visual information retrieval interfaces. In: Proc. of the Int'l Conf. on Multidisciplinary Information Sciences and Technologies. 2006. 25−28.

[2] Skoutas D, Alrifai M. Tag clouds revisited. In: Berendt B, de Vries A, Fan WF, Macdonald C, Ounis I, Ruthven I, eds. Proc. of the 20th ACM Int'l Conf. on Information and Knowledge Management. ACM Press, 2011. 221−230. [doi: 10.1145/2063576. 2063613]

[3] Venetis P, Koutrika G, Garcia-Molina H. On the selection of tags for tag clouds. In: Proc. of the 4th ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2011. 835−844. [doi: 10.1145/1935826.1935855]

[4] Wang M, Zhou X, Tao QM, Wu W, Zhao C. Diversifying tag selection result for tag clouds by enhancing both coverage and dissimilarity. In: Lin X, ed. Proc. of the 14th Int'l Conf. on Web Information System Engineering. Berlin, Heidelberg: Springer-Verlag, 2013. 29−42. [doi: 10.1007/978-3-642-41154-0_3]

[5] Zhai CX, Cohen WW, Lafferty J. Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2003. 10−17. [doi: 10.1145/860435.860440]

[6] Clarke CLA, Kolla M, Cormack GV, Vechtomova O, Ashkan A, Büttcher S, MacKinnon I. Novelty and diversity in information retrieval evaluation. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2008. 659−666. [doi: 10.1145/1390334.1390446]

[7] Agrawal R, Gollapudi S, Halverson A, Ieong S. Diversifying search results. In: Ricardo Baeza-Yates PB, Ribeiro-Neto B, Barla Cambazoglu B, eds. Proc. of the 2nd ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2009. 5−14. [doi: 10.1145/1498759.1498766]

[8] Bansal N, Jain K, Kazeykina A, Naor J. Approximation algorithms for diversified search ranking. In: Abramsky S, Gavoille C, Kirchner C, Meyer auf der Heide F, Spirakis P, eds. Proc. of the Automata, Languages and Programming. Berlin, Heidelberg: Springer-Verlag, 2010. 273−284. [doi: 10.1007/978-3-642-14162-1_23]

[9] Demidova E, Fankhauser P, Zhou X, Nejdl W. DivQ: Diversification for keyword search over structured databases. In: Proc. of the 33rd Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2010. 331−338. [doi: 10.1145/1835449.1835506]

[10] Fraternali P, Martinenghi D, Tagliasacchi M. Top-*k* bounded diversification. In: Proc. of the 2012 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2012. 421−432. [doi: 10.1145/2213836.2213884]

[11] Ranu S, Hoang M, Singh A. Answering top-*k* representative queries on graph databases. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. ACM Press, 2014. 1163−1174. [doi: 10.1145/2588555.2610524]

[12] Liu K, Terzi E, Grandison T. Highlighting diverse concepts in documents. In: Proc. of the 9th SIAM Int'l Conf. on Data Mining. Society for Industrial and Applied Mathematics. 2009. 545−556.

[13] Lin H, Bilmes J. A class of submodular functions for document summarization. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2011. 510−520.

[14] Tsaparas P, Ntoulas Alexandros, Terz E. Selecting a comprehensive set of reviews. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2011. 168−176. [doi: 10.1145/2020408.2020440]

[15] Yu C, Lakshmanan L, Amer-Yahia S. It takes variety to make a world: diversification in recommender systems. In: Kersten M, Novikov B, Teubner J, Polutin V, Manegold S, eds. Proc. of the 12th Int'l Conf. on Extending Database Technology: Advances in Database Technology. ACM Press, 2009. 368−378. [doi: 10.1145/1516360.1516404]

[16] Hurley N, Zhang M. Novelty and diversity in top-*N* recommendation—Analysis and evaluation. ACM Trans. on Internet Technology (TOIT), 2011,10:1−30. [doi: 10.1145/1944339.1944341]

[17] Cui CR, Ma J. An image tag recommendation approach combining relevance with diversity. Chinese Journal of Computers, 2013, 36:654−663 (in Chinese with English abstract).

[18] Carbonell J, Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. of the 21st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 1998. 335−336. [doi: 10.1145/290941.291025]

[19] Gollapudi S, Sharma A. An axiomatic approach for result diversification. In: Proc. of the 18th Int'l Conf. on World Wide Web. ACM Press, 2009. 381−390. [doi: 10.1145/1526709.1526761]

[20] Drosou M, Pitoura E. Search result diversification. ACM SIGMOD Record, 2010,39:41−47. [doi: 10.1145/1860702.1860709]

[21] Borodin A, Lee HC, Ye Y. Max-Sum diversification, monotone submodular functions and dynamic updates. In: Krötzsch M, ed. Proc. of the 31st Symp. on Principles of Database Systems. ACM Press, 2012. 155−166. [doi: 10.1145/2213556.2213580]

[22] Drosou M, Pitoura E. Disc diversity: Result diversification based on dissimilarity and coverage. In: Proc. of the VLDB Endowment. 2012. 13−24. [doi: 10.14778/2428536.2428538]

[23] Mika P. Ontologies are us: A unified model of social networks and semantics. In: Gil Y, Motta E, Benjamins VR, Musen M, eds. Proc. of 4th Int'l Semantic Web Conf. Berlin, Heidelberg: Springer-Verlag, 2005. 522−536. [doi: 10.1007/11574620_38]

[24] Halpin H, Robu V, Shepherd H. The complex dynamics of collaborative tagging. In: Proc. of the 16th Int'l Conf. on World Wide Web. ACM Press, 2007. 211−220. [doi: 10.1145/1242572.1242602]

[25] Fujishige S. Submodular Functions and Optimization. 2nd ed., Elsevier Science, 2005.

[26] Deza E, Deza MM. Dictionary of Distances. Elsevier Science, 2006.

[27] Markines B, Cattuto C, Menczer F, Benz D, Hotho A, Stumme G. Evaluating similarity measures for emergent semantics of social tagging. In: Proc. of the 18th Int'l Conf. on World Wide Web. ACM Press, 2009. 641−650. [doi: 10.1145/1526709.1526796]

[28] Hochba DS. Approximation algorithms for NP-hard problems. ACM SIGACT News, 1997,28:40−52.

[29] Fadaei S, Fazli M, Safari M. Maximizing non-monotone submodular set functions subject to different constraints: Combined algorithms. Operations Research Letters, 2011,39:447−451. [doi: 10.1016/j.orl.2011.10.002]

[30] Lee J, Mirrokni VS, Nagarajan V, Sviridenko M. Non-Monotone submodular maximization under matroid and knapsack constraints. In: Proc. of the 41st Annual ACM Symp. on Theory of Computing. ACM Press, 2009. 323−332. [doi: 10.1145/1536414.1536459]

[31] Levandowsky M, Winter D. Distance between sets. Nature, 1971,234:34−35. [doi: 10.1038/234034a0]

[32] Nemhauser GL, Wolsey LA, Fisher ML. An analysis of approximations for maximizing submodular set functions—I. Mathematical Programming, 1978,14:265−294. [doi: 10.1007/BF01588971]

[33] Song Y, Zhuang ZM, Li HJ, Zhao QK, Li J, Lee WC, Giles CL. Real-Time automatic tag recommendation. In: Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM Press, 2008. 515−522. [doi: 10.1145/1390334.1390423]

附中文参考文献:

[17] 崔超然,马军.一种结合相关性和多样性的图像标签推荐方法.计算机学报,2013,36:654−663.

汪美玲(1982－),女,吉林伊通人,博士生,主要研究领域为标注系统,社交网络,数据分析.

陶秋铭(1979－),男,博士,副研究员,主要研究领域为编译优化技术,软件工程.

周翔(1980－),男,博士,助理研究员,主要研究领域为形式化方法,多核算法和程序设计.

赵琛(1967－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为编译优化,软件测试,人工智能.