计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2015.20140805
52(11):2488-2495,2015

一种基于社会化标注的查询扩展研究

董华磊　王　健　林鸿飞　王　浩
(大连理工大学计算机科学与技术学院　辽宁大连　116024)
(donghl@mail.dlut.edu.cn)

A Study of Query Expansion Based on Social Tagging

Dong Hualei, Wang Jian, Lin Hongfei, and Wang Hao
(*School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning* 116024)

Abstract　With the development of Web 2.0, many websites allow users to create and manage their social tags. A lot of searches show that social annotations can be used to improve search quality, but the real tagging system is often sparse, uncategorized, lack of structure and of low quality, therefore traditional SimRank algorithm is so difficult to work. Introducing Jaccard index to SimRank algorithm, we put forward the improvement of social tagging Jaccard SimRank (JSR) similarity calculation method which automatically analyzes the similarity of user-input social annotations and expands them to increase the density. JSR algorithm can make full use of the information of social tagging to achieve effective retrieval and to describe similarity between any two tags intuitively. The experimental datasets come from bibsonomy website, and we have applied Jaccard index, SimRank and JSR algorithms against the test datasets. Experimental results show that the JSR algorithm is more effective in improving search quality than the traditional algorithms.

Key words　query expansion; social tagging; SimRank algorithm; Jaccard index; similarity

摘　要　采用社会化标签可以提高检索质量,但真实的标注系统往往比较稀疏,并且标签存在无序性、不规范性和低效性等特点,因此单纯使用传统的 SimRank 等相似度算法难以奏效.为此,在 SimRank 算法基础上融入 Jaccard 系数计算,提出一种改进的社会化标签的相似度计算方法,称作 Jaccard SimRank (JSR)算法,更加直观地描述社会化标签之间的相似度,在用户标注网络资源时自动对标签集进行扩展,增加标注密度,并在检索时对标签集进行扩展,因而能够更充分利用社会化标注系统的信息实现有效检索.实验结果表明,与传统的相似度算法相比,JSR 方法有效提高了查询扩展系统的性能.

关键词　查询扩展;社会化标注;SimRank 算法;Jaccard 系数;相似度

中图法分类号　TP391

在信息检索应用中,用户使用的查询词往往不能准确表达用户意图,即便是用户与网络资源编辑者有相同的意图,如果理解方式不同或是知识背景不同,就很容易出现描述方式不同的现象,进而导致词不匹配、信息过载和信息迷向等问题.例如用户对1个资源标注"苹果",如果没有语境,仅仅从标签词很难理解指的是水果还是手机,这样会对检索性能产生不同程度的负面影响.

收稿日期:2014-09-03;**修回日期**:2015-05-05
基金项目:国家自然科学基金项目(61340020,61272373,61300088);辽宁省自然科学基金项目(201202031)
通信作者:王　健(wangjian@dlut.edu.cn)

随着互联网的快速发展,Web 已成为人们获取信息的一个重要途径,社会化标注作为典型的 Web2.0 时代的网络资源,由用户群体自下而上对网络资源进行评价和分享.在用户对网络资源进行标注时,可以收集和分析用户的信息来学习用户的行为和兴趣.所实施的标注不需要专业的知识背景和特定的文化背景,更贴合用户群体的表达方式.很多研究者对社会化标注的应用产生了兴趣,社会化标注可以应用于生成网页摘要[1],也可以改善网页检索[2-4],一些研究者利用其进行标签推荐[5-6].将社会化标注引入到信息检索当中可以在一定程度上弥补传统信息检索中的词不匹配问题.

用户使用的标签不规范、无序,很多时候相当低效.标签词中存在同义词、一词多义、大量流行用语等现象.为了更高效地利用这些标签词,很多标签词扩展的方法被提出[7-8],这些扩展方法往往都基于标签相似度算法,在用户对网络资源进行标注时,利用标签词查询网页对标签集进行扩展来改进查询质量.

大多数的经典算法,如 Cosine 相似度、Jaccard 系数、Pearson 相关分析等,在标签对网络资源的密集标注时是有效的[9-11],但在实际应用中用户的标注往往非常稀疏[12-13].例如,在 1 个标注系统中使用少量的标签词标注大多数网络资源,剩余的大量标签标注剩余的少量网络资源.这就意味着在极端情况下用户选取标签词进行查询时,可能该标签词标注的网络资源数量几乎为零,导致很难查询出有用结果,而传统算法在标注稀疏的时候几乎是无效的.

针对上述问题,本文提出了一种计算社会化标签相似度的算法——Jaccard SimRank(JSR)算法,在用户标注网络资源时自动对标签集进行扩展,增加标注密度,并在检索时对标签集进行扩展以提高检索精度.

1 相关背景

1.1 研究背景

社会化标注系统存在标注稀疏问题.以 1 个社会化标注系统 BibSonomy[14-15] 为例,它允许用户对网络资源自由地进行标注.BibSonomy 中约 58%的网络资源只有不足 3 个标签词,约 23%的网络资源有不足 5 个标签词.所有标签词中约 81%的标签词被不足 5 个用户使用过,剩余的 19%被使用的次数大于 5.可见大多数的网络资源只被很少量的用户标注,即该网络资源的标注只反映了少量用户的观点和喜好,往往比较片面并且无法反映出用户群体的观点和喜好,没有充分利用社会化标注系统的优点.

为了解决上述问题,可以使用查询扩展技术对社会化标注系统进行改进.即可以在用户标注网络资源时对用户使用的标注词进行扩展,通过选取扩展词引入其他用户的观点,缓解因为标签词少造成的标注稀疏问题,标签词描述的问题比较片面,可以在用户查询时对查询词进行扩展,通过选取新的查询词缓解信息检索领域的词不匹配问题.

传统的查询扩展技术研究很多.但是目前基于社会化标注的查询扩展研究还较少,其中晋松[16]提出了基于排序学习的扩展词挖掘方法、基于词依赖的扩展词挖掘方法和基于词共现统计的扩展词挖掘方法;张志强等人[17]提出了标签分析扩展法、社会化标签查询词共现分析扩展法和词频(term frequency,TF)分析扩展法.

本文方法根据标签语义分析的相关研究[18-20],利用"标注到同一网络资源的标签词之间存在一定的语义相关性"这个假设,为查询词选取扩展词,如表 1 所示:

Table 1 Tagged Tags from Different Sites

表 1 不同网站被标注的标签词

Website	Tags of User Tagging
www.google.com	Search Google Searchengine Web Engine Reference Bookmarkbar Searchengines Tools
www.facebook.com	Social Facebook Networking Friends Community Web2.0 Entertainment Network Bookmarkbar
www.flickr.com	Photos Flickr Photography Photo Sharing Images Web2.0 Community Social Tools
delicious.com	Delicious Social Web2.0 SNS Tools Bookmarking Del.icio.us Web Social Networking Search
www.bibsonomy.org	Bibliography Bibtex Tagging Folksonomy Web2.0 Social Research Bookmarking Tags Tools
www.bbc.co.uk	News Bbc Media Tv Radio Uk English Gibenchmark Sport Ingles

表 1 中的数据来自多个网站,其中一些是流行的网站,标签词是当前用户标注该网络资源时所使用的最频繁的词汇.根据这些数据可以看出,这些标签词从不同的角度描述相应的网站,排名靠前的

标签词描述的往往是相应网站比较突出的特性,排名靠后的标签词从其他角度对相应网站进行描述,并且同一个网站中的标签词之间存在语义相关性,这也印证了之前的假设.目前基于社会化标注的查询扩展方法主要由传统的查询扩展方法发展而来.本文采用的查询扩展方法属于全局分析法,是基于社会化标注相似度的一种查询扩展方法.

1.2 传统的相似度算法

在目前的社会化标注系统中,一般允许用户对网络资源标注多个标签词,也允许用户对网络资源实施多次标注.社会化标注系统中的信息有这样的特点,1 个用户对 1 个网络资源可以标注多个标签词 t,且数量没有限制.我们可以将这些信息看成是由许多书签组成的,每个书签包含 1 个用户 u、1 个网络资源 r 和用户 u 对该网络资源标注的所有标签词的集合 S,三元组形式为$\langle u,r,S\rangle$.为了更好地对算法进行描述,定义某个社会化标注系统中的用户总数为 n_u,标签总数为 n_t,网络资源总数为 n_r.系统中的标签词和网络资源可以生成一个 $n_t\times n_r$ 的关联矩阵称为 $\boldsymbol{T}$,T_{ij} 代表矩阵中第 j 个网络资源被第 i 个标签词标注的次数.$t_r(i)$为矩阵 $\boldsymbol{T}$ 中第 i 行元素的集合;$r_t(i)$为矩阵 $\boldsymbol{T}$ 中第 i 列元素的集合.标签之间的相似度矩阵表述为 $\boldsymbol{S}_t$,网络资源之间的相似度矩阵表述为 $\boldsymbol{S}_r$.

传统的相似度算法如 Cosine 相似度计算见式(1),Jaccard 系数计算见式(2):

$$s_t(t_i,t_j)=\frac{\langle t_r(i),t_r(j)\rangle}{\sqrt{\langle t_r(i),t_r(i)\rangle\times\sqrt{\langle t_r(j),t_r(j)\rangle}}},\quad(1)$$

$$s_t(t_a,t_b)=\frac{\sum_{i=1}^{n_r}\min(T_{ai},T_{bi})}{\sum_{i=1}^{n_r}T_{ai}+\sum_{i=1}^{n_r}T_{bi}-\sum_{i=1}^{n_r}\min(T_{ai},T_{bi})}.\quad(2)$$

这 2 种算法均基于标签词共现原理,即只有标注过同一个网络资源的标签词之间才具有相似度,但如 1.1 节所述真实的社会化标注系统中存在标注稀疏现象,由此可知矩阵 $\boldsymbol{T}$ 是相当稀疏的.在 Cosine 相似度和 Jaccard 系数中任何 2 个标签的相似度,比如 $s_t(t_a,t_b)$,必须比较 $t_r(a)$和 $t_r(b)$之间标注了同一网络资源的元素,但由于矩阵 $\boldsymbol{T}$ 的稀疏性,导致了 $t_r(a)$和 $t_r(b)$之间基本上没有共同标注同一网络资源的元素,即使存在,数量也很少,这样 Cosine 相似度和 Jaccard 系数在计算时大多数为 0 或者接近于 0.所以传统的算法 Cosine 相似度及 Jaccard 系数计算出来的标签相似度矩阵 $\boldsymbol{S}_t$ 也是稀疏的,基于这 2 种社会化标注的相似度算法的查询扩展方法几乎是无效的.

因为真实的社会化标注系统存在稀疏性,我们可以尝试使用 SimRank[21-22] 相似度算法,如式(3)(4)所示:

$$s_t^k(t_a,t_b)=\frac{C_1}{|r(t_a)|\times|r(t_b)|}\sum_{r_i\in r(t_a)}\sum_{r_j\in r(t_b)}s_r^{k-1}(r_i,r_j),\quad(3)$$

$$s_r^k(r_i,r_j)=\frac{C_2}{|t(r_a)|\times|t(r_b)|}\sum_{t_i\in t(t_a)}\sum_{t_j\in t(r_b)}s_t^{k-1}(t_i,t_j),\quad(4)$$

其中,$s_t^k(t_a,t_b)$和 $s_r^k(r_a,r_b)$分别表示第 k 次迭代后标签 t_a 与标签 t_b 以及网络资源 r_a 和 r_b 之间的相似度;C_1 和 C_2 是介于 0 和 1 之间的常量.SimRank 算法利用共有元素相互增强的原则,通过迭代计算能在一定程度上缓解数据稀疏性问题,但在社会化标注系统中使用 SimRank 算法计算相似度是不恰当的.

图 1 是利用社会化标注系统中的数据信息构造的二部图.

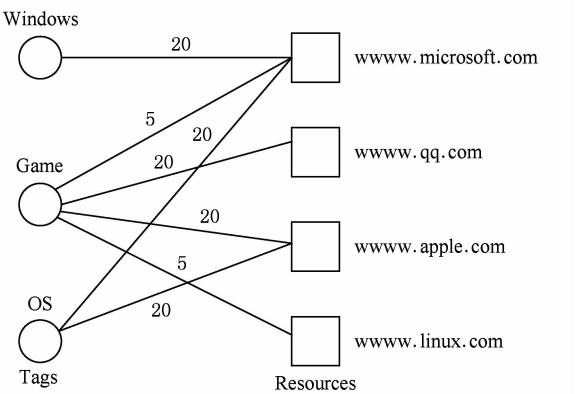


Fig. 1 Bipartite graph based on social tagging.

图 1 基于社会化标注的二部图

从图 1 可以看出,社会化标注系统中标注过相同网络资源的标签之间有一定的相似度,被同一标签标注过的网络资源之间也存在一定的相似度;标签词与网络资源连接的边的权值,表示将该标签词标注到该网络资源上的用户数量.SimRank 算法中没有利用用户数量作为权值,而是将所有标签对网络资源标注过的边的权值都置为 1,并且 SimRank 算法没有区分标注过相同网络资源的标签和标注过有相似度但不是相同的网络资源的标签.基于二部图结构的 SimRank 算法忽视了标签词之间和网络资源之间共现度的作用.例如 2 个标签 MP3 和 Music

之间虽然存在着很高的语义相关性,但随着使用社会化标注系统人数的增多,这 2 个标签将被标注到越来越多的网络资源上,导致了 MP3 和 Music 之间的相似度越来越小,这显然是不合理的. 基于以上 2 点,SimRank 算法虽然能在一定程度上缓解标注稀疏性问题,但效果并不显著.

2 基于社会化标注的查询扩展

2.1 JSR 算法提出

本文对提出的 JSR 算法考虑数据集里的潜在信息,并尝试使用最直观的共现度衡量标签对的相似度方法. 首先需要使矩阵 $\boldsymbol{S}_t$ 更加密集,运用 SimRank 算法的思想,利用以下观点进行迭代计算使矩阵 $\boldsymbol{S}_t$ 更加密集:标注过相同网络资源的标签之间存在相似性;被相同标签标注过的网络资源之间也存在相似性. 其次运用 Jaccard 系数的思想,使用最直观的共现度衡量相似度,也就是如果 2 个集合的交集越大则它们的相似度越大,并集越大则相似度越小. 下面以计算 $s_t(t_a,t_b)$ 为例,讨论 $t_r(a)$ 与 $t_r(b)$ 交集的概念.

首先狭义交集 $I_s(t_r(a),t_r(b))$ 的定义是 $t_r(a)$ 与 $t_r(b)$ 最直观的交集,计算公式如下:

$$I_s(t_r(a),t_r(b))=\sum_{i=1}^{n_r}\min(T_{ai},T_{bi}). \tag{5}$$

由式(5)可见,狭义交集没有利用 $t_r(a)$ 与 $t_r(b)$ 中具有相似度的网络资源来标注信息,计算狭义交集的信息是恒定的,无论多少次计算都不能改变该集合,所以矩阵 $\boldsymbol{S}_t$ 经过多少次迭代计算都不能够使矩阵密集,因此需要对交集重定义,能够对交集进行扩充,狭义交集可认为是 $t_r(a)$ 与 $t_r(b)$ 中相似度等于 1 的网络资源对的交集,$t_r(a)$ 与 $t_r(b)$ 中相似度在 0 和 1 之间的网络资源对也可以扩充为交集,因此广义交集 $I_g(t_r(a),t_r(b))$ 即是我们扩充后的交集.

首先讨论 $t_r(a)$ 中的 T_{ai} 与 $t_r(b)$ 的广义交集,T_{ai} 与 $t_r(b)$ 中的 T_{bj} 之间的相似度为 $s_r(r_i,r_j)\times\min(T_{ai},T_{bj})$,则 T_{ai} 与 $t_r(b)$ 可能存在的广义交集定义为潜在交集 $I_l(T_{ai},t_r(b))$,计算公式如下:

$$I_l(T_{ai},t_r(b))=\sum_{j=1}^{|t_r(b)|}(s_r(r_i,r_j)\times\min(T_{ai},T_{bj})), \tag{6}$$

但 T_{ai} 与 $t_r(b)$ 的潜在交集并不等于 T_{ai} 与 $t_r(b)$ 的广义交集,因为潜在交集不应大于 T_{ai} 与 $t_r(b)$ 的最大交集 $I_m(T_{ai},t_r(b))$,潜在交集公式如下:

$$I_m(T_{ai},t_r(b))=T_{ai}\times\min\Big(\sum_{j=1}^{|t_r(b)|}s_r(r_i,r_j),1\Big), \tag{7}$$

所以 T_{ai} 与 $t_r(b)$ 的广义交集 $I_g(T_{ai},t_r(b))$,见式(8):

$$I_g(T_{ai},t_r(b))=\min(I_l(T_{ai},t_r(b)),I_m(T_{ai},t_r(b))). \tag{8}$$

可以计算出 $t_r(a)$ 与 $t_r(b)$ 的广义交集 $I_g(t_r(a),t_r(b))$:

$$I_g(t_r(a),t_r(b))=\min\Big(\sum_{i=1}^{|t_r(a)|}I_g(T_{ai},t_r(b)),\sum_{j=1}^{|t_r(b)|}I_g(T_{ai},t_r(a))\Big). \tag{9}$$

需要注意的是此处计算出的广义交集并非精确的广义交集,广义交集在很多情况下无法精确计算,所以本文用广义交集的近似值进行实验.

$t_r(a)$ 的全集为 $C(t_r(a))$,公式如下:

$$C(t_r(a))=\sum_{i=1}^{|t_r(a)|}T_{ai}. \tag{10}$$

为了区分狭义交集与引入广义交集后额外产生的交集之间的差别,在相似度公式中引入衰减因子 α 来控制引入广义交集后产生的影响.

JSR 算法流程如算法 1 所示:

算法 1. JSR 算法.

Step1. For each 标签对$(t_a,\ t_b)$和网络资源对$(r_a,\ r_b)$ do {

if $t_a=t_b\{s_t^0(t_a,t_b)=1;\}$

otherwise $s_t^0(t_a,\ t_b)=0$;

if $r_a=r_b\{s_r^0(r_a,r_b)=1;\}$

otherwise $s_r^0(r_a,\ r_b)=0;\}$

Step2. Repeat

For each 标签对$(t_a,\ t_b)$ do {

$$s_t^k(t_a,t_b)=(I_s(t_r(a),t_r(b))+\alpha(I_g^{k-1}(t_r(a),t_r(b)))-I_s(t_r(a),t_r(b)))/(C(t_r(a))+C(t_r(b))-I_s(t_r(a),t_r(b))-\alpha(I_g^{k-1}(t_r(a),t_r(b)))-I_s(t_r(a),t_r(b)));\} \tag{11}$$

For each 网络资源对$(r_a,\ r_b)$ do {

$$s_r^k(r_a,r_b)=(I_s(r_t(a),r_t(b))+\alpha(I_g^{k-1}(r_t(a),r_t(b)))-I_s(r_t(a),r_t(b)))/(C(r_t(a))+C(r_t(b))-I_s(r_t(a),r_t(b))-\alpha(I_g^{k-1}(r_t(a),r_t(b)))-I_s(r_t(a),r_t(b)));\} \tag{12}$$

Until $s_t(t_a,\ t_b)$收敛;

Step3. 输出 $s_t(t_a,\ t_b)$.

其中,$s_t^k(t_a,t_b)$和$s_r^k(r_a,r_b)$分别表示k次迭代后标签t_a与t_b以及网络资源r_a和r_b之间的相似度.

2.2 JSR算法的时间复杂度及收敛性

JSR算法的时间复杂度为$O(k(n_r^2\times n_t^2))$,k代表算法迭代的次数.该算法的时间复杂度主要取决于标签和网络资源的数目,因为在实际应用中标签对网络资源的标注非常稀疏,算法可以在较少的迭代次数内收敛.δ_t^k与δ_r^k的计算公式如下:

$$\delta_t^k=\frac{\|s_t^k-s_t^{k-1}\|_1}{\|s_t^k\|_1},\tag{13}$$

$$\delta_r^k=\frac{\|s_r^k-s_r^{k-1}\|_1}{\|s_r^{k-1}\|_1}.\tag{14}$$

如图2所示为迭代次数(从第1～10次)与δ_t^k和δ_r^k的关系.从图2可以看出,经过4次迭代计算后δ_t^k和δ_r^k的值趋于平稳,并且都小于0.1,证明该算法的收敛性是可靠的.

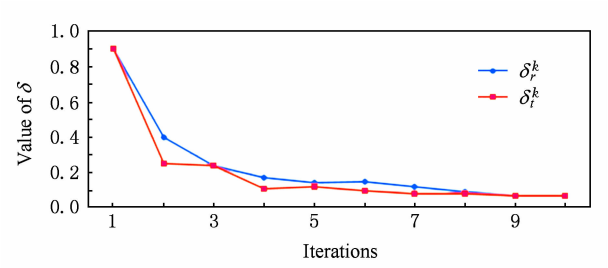


Fig. 2 Value of δ_t^k and δ_r^k after one to ten times iteration.

图2 第1～10次迭代计算后δ_t^k和δ_r^k的值

2.3 扩展方法

本文运用标签词之间的相似度进行查询扩展,相似度的计算比较耗时,但是相似度可以离线计算所以时间消耗可以忽略不计.考虑了标注过网络资源的标签词频率和标签词被使用过的次数,能够体现标签词与查询词之间的相关度及在数据集中的重要性.

假设$J=\{t_1,t_2,\cdots,t_n\}$为1个由n个词构成的查询,$t_j\in J$,$t_i\notin J$,则扩展词跟该查询的相关度计算公式如下:

$$s_c(t_i,t_j)=s_t(t_i,t_j)\times\lg count(t_i)\times R(t_i),\tag{15}$$

其中,R表示逆网络资源频率.设计思想类似于逆文档频率(inverse document frequency, IDF),R的计算公式如下:

$$R(t_i)=\lg\frac{r(t_i)}{n_r},\tag{16}$$

其中,$r(t_i)$表示被标签词t_i标注过的网络资源数量,n_r表示数据集中网络资源的数量.

$$M(t_i,J)=\sum_{t_j\in S}s_c(t_i,t_j),\tag{17}$$

其中,$M(t_i,J)$表示扩展标签词t_i与查询J之间的相关度.根据式(9)选取与查询J最相关的前k个标签词作为扩展词.如果$|J|>6$,$k=[0.5\times|J|]$;如果$|J|\leqslant 6$,则$k=3$.

在真实的社会化标注系统中标注十分稀疏,本文模拟用户对网络资源标注时,对用户标注的标签词进行扩展,提高矩阵标注密度从而增加每个书签中标签词的数量,并在用户查询时对用户所使用的标签词进行扩展,增加查询长度,降低因为用户标注的随意性、稀疏性和差异性带来的影响.

3 实验设计与结果分析

3.1 实验语料

本文实验所使用的数据集是由文献[23]的作者提供,采用真实的社会化标注系统网站①于2009年6月的网页快照,共包含190 147个文档、4 969个用户、168 686个标签和648 924个书签.

3.2 检索模型和评价方法

首先对数据集进行整理,表示成书签形式:$\langle u,r,S\rangle$;其次将数据集分成10份,其中9份为训练集,剩余1份为测试集.利用训练集进行训练计算出所有标签词之间的相似度,为训练集和测试集的所有书签选取k个标签词进行扩展,此过程模拟用户对网络资源标注标签集S时自动对S进行扩展,增加书签中标签词的数量,并利用测试集中每个书签的标签集S查询相关网络资源时也自动对S进行扩展,增加查询长度.

当用户将1组标签词标注到1个网络资源上,模拟这一过程使用查询词对网络资源进行检索,通过计算平均检索成功率对JSR算法进行评价时,形成1个书签$\langle u,r,S\rangle$,可以理解为用户u认为该标签集S是与网络资源r最相关的描述.如果使用标签集S作为查询,那么最相关的结果应该是网络资源r.

标签与网络资源之间的相关性计算使用TF-IDF(term frequency-inverse document frequency)

① http://www.bibsonomy.org

公式,标签集与网络资源的相关性为所有标签词的和值,公式如下:

$$tf_{t_a,r_i}=\frac{T_{ai}}{\left|\sum_{m=1}^{t(r_i)}T_{mi}\right|},\tag{18}$$

$$idf_{t_a,r_i}=\lg\left(\frac{n_r}{|r_i(t_a)|}\right),\tag{19}$$

$$p(S,r_i)=\sum_{t_a\in S}tf_{t_a,r_i}\times idf_{t_a,r_i}.\tag{20}$$

对每个查询分别计算出 q 个相关的网络资源,$q\in\{5,10,20\}$;再分别将这些网络资源与该书签中的网络资源 r 进行对比,包含 r 则查询成功,不包含网络资源 r 则查询失败;最后计算算法的平均查询成功率.数据集分成了 10 份,所以需要以每份为测试集计算 10 次,并算出平均值.本文以不进行查询扩展的检索模型作为 Baseline,同时实现了 Cosine 算法、Jaccard 系数、SimRank 算法的比较.

3.3 实验结果分析

实验结果如图 3 所示,记录了检索结果中取前 N 个网络资源时,各个算法对应的平均成功率.实验绘制了 N 分别取 5,10,20 时的实验结果,实验证明 JSR 算法中的衰减因子 $\alpha=0.3$ 时效果较好.

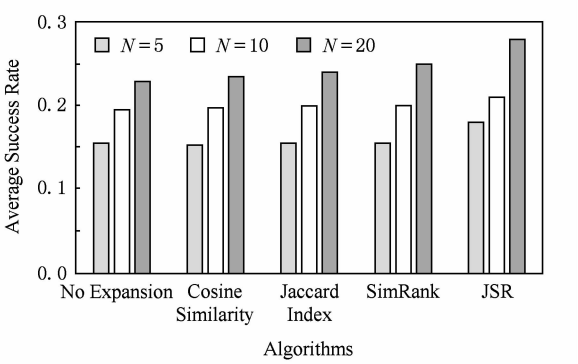


Fig. 3 Average success rate of retrieval by five algorithms.

图 3 5 种算法的平均检索成功率

从图 3 的实验结果显示,当 $N=20$ 时,平均检索成功率相对于其他算法提高更多,没有进行查询扩展的检索模型是最不理想的检索模型,而其他任何一种社会化相似度算法相对于不进行查询扩展都能提高检索结果.社会化标注是用户在使用网络资源时对网络资源内容做的简单描述,网络资源的用户越多,他们的相似标注词就越能体现用户群对这一网络资源的集体认同.而查询也是出自互联网中的同一群体,所使用的查询词与标签词往往存在着一定的重合性,通过对查询词进行扩展可以扩大查询范围,可以明显提高检索效果.

在社会化标注系统中,数据集一般非常稀疏,而 Cosine 相似度算法和 Jaccard 系数只有在数据集密集的情况下才会有较好的效果,所以在社会化标注系统中这 2 种相似度算法在实验中效果一般.SimRank 算法在数据集稀疏的情况下效果较好,但是在社会化标注系统中该算法公式没有使用矩阵 $\boldsymbol{T}$ 中的权值,忽视了同一个标签词对 1 个网络资源的标注次数,无法体现网络资源较为认可的标签词.所以 SimRank 算法提升实验效果不是很明显.

为了更直观地观察多种相似度算法的合理性,本文随机抽取了一些标签词,并使用各相似度算法对其进行扩展,如表 2 所示:

Table 2 Effective Comparison of the Related Tags Mined by Five Algorithms

表 2 5 种算法挖掘出的相关标签词比较

Tag	Algorithm	Relevant Tags
Music	JSR	MP3 Audio Radio Free Service
	SimRank	MP3 Web2.0 Community Film Blog
	Jaccard Index	MP3 Software Radio P2P Tool
	Cosine	MP3 Media Flash News Sharing
Windows	JSR	WinXP Microsoft Office Guide Homepage
	SimRank	WinXP Service Security Guide Nasa
	Jaccard Index	Security Freeware Linux Free Download
	Cosine	Web Opensource Game WinXP MicroSoft
Web2.0	JSR	Blog Share Folksonomy Search Recommendations
	SimRank	Share Blog Delicious Search Cool
	Jaccard Index	Blog Webapp Share Community Design
	Cosine	Blog Design Delicious Music Cool

从表 2 可以看出,使用 JSR 算法计算得到的扩展标签词排名更合理,不相关的词也较少;JSR 算法利用了社会化标注系统的潜在信息,如同一个标签词对 1 个网络资源的标注次数、2 个标签词分别标注的网络资源集的重合度和 2 个网络资源分别被标注的标签集的重合度,算法的提升效果较为明显.

4 结束语

本文根据真实的社会化标注系统特点,提出了一种新的计算社会化标签相似度的 JSR 算法,尝试使用最直观的共现度衡量 2 个标签词之间的相似度,并且在某种程度上缓解了社会化标注系统的标注稀疏问题.将 JSR 算法应用于查询扩展后,实验结果优于传统的 Cosine 相似度算法、Jaccard 系数和 SimRank 算法.

数据集中存在着垃圾标注现象,具体表现为单个用户对大量的网页标注相同的标签词,这种标注几乎没有任何价值,或者用户对网络资源标注毫无关系的标签词,都极大地干扰了实验效果.如何降低垃圾标注对社会化标注系统的负面影响是今后一段时期的主要工作.

参 考 文 献

[1] Shang Shujie, Wang Can, Zhu Junyan. Tag-based Web page summarization approach [J]. Computer Engineering, 2010, 36(21): 260-261, 264 (in Chinese)
(尚书杰, 王灿, 朱俊彦. 一种基于标签的网页摘要方法[J]. 计算机工程, 2010, 36(21): 260-261, 264)

[2] Sarwar S M, Abedin M A, Ullah A H M, et al. Personalized query expansion for Web search using social keywords [C]// Proc of Int Conf on Information Integration and Web-based Applications & Services. New York: ACM, 2013: 610-614

[3] Bao S, Xue G, Wu X, et al. Optimizing Web search using social annotations [C]//Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 501-510

[4] Xu S, Bao S, Cao Y, et al. Using social annotations to improve language model for information retrieval [C] //Proc of the 16th ACM Conf on Information and Knowledge Management. New York: ACM, 2007: 1003-1006

[5] Wang T, Wang H, Yin G, et al. Tag recommendation for open source software [J]. Frontiers of Computer Science, 2014, 8(1): 69-82

[6] Yuan X, Huang J, Zhong N. Preference structure and similarity measure in tag-based recommender systems [G] // Active Media Technology. Berlin: Springer, 2013: 193-202

[7] Tso-Sutter K H L, Marinho L B, Schmidt-Thieme L. Tag-aware recommender systems by fusion of collaborative filtering algorithms [C]//Proc of the 2008 ACM Symp on Applied Computing. New York: ACM, 2008: 1995-1999

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

[9] Gemmell J, Schimoler T, Ramezani M, et al. Adapting K-Nearest neighbor for tag recommendation in folksonomies [C]//Proc of the 7th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems. Menlo Park, CA: AAAI, 2009: 528-539

[10] Jäschke R, Marinho L, Hotho A, et al. Tag recommendations in folksonomies [G] //Knowledge Discovery in Databases. Berlin: Springer, 2007: 506-514

[11] Wang Hao. The study of query expansion in social tagging [D]. Dalian, Liaoning: Dalian University of Technology, 2013 (in Chinese)
(王浩. 基于社会化标注的查询扩展研究[D]. 辽宁大连: 大连理工大学, 2013)

[12] Cattuto C, Schmitz C, Baldassarri A, et al. Network properties of folksonomies [J]. AI Communications, 2007, 20(4): 245-262

[13] De Meo P, Quattrone G, Ursino D. Exploitation of semantic relationships and hierarchical data structures to support a user in his annotation and browsing activities in folksonomies [J]. Information Systems, 2009, 34(6): 511-535

[14] Hotho A, Jäschke R, Schmitz C, et al. BibSonomy: A social bookmark and publication sharing system [C] //Proc of the 14th Int Conf on Conceptual Structures. Berlin: Springer, 2006: 87-102

[15] Jäschke R, Hotho A, Schmitz C, et al. Analysis of the publication sharing behaviour in BibSonomy [G] //Conceptual Structures: Knowledge Architectures for Smart Applications. Berlin: Springer, 2007: 283-295

[16] Jin Song. Query expansion based on social annotation [D]. Dalian, Liaoning: Dalian University of Technology, 2010 (in Chinese)
(晋松. 基于社会化标注的查询扩展技术研究[D]. 辽宁大连: 大连理工大学, 2010)

[17] Zhang Zhiqiang, Meng Qinghai, Xie Xiaoqin, et al. Research on personalized social tag query expansion techniques [J]. Journal of Frontiers of Computer Science and Technology, 2010, 4(9): 812-829 (in Chinese)
(张志强, 孟庆海, 谢晓芹, 等. 个性化的社会标签查询扩展技术研究[J]. 计算机科学与探索, 2010, 4(9): 812-829)

[18] Gabriel H H, Spiliopoulou M, Nanopoulos A. Summarizing dynamic social tagging systems [J]. Expert Systems with Applications, 2014, 41(2): 457-469

[19] Li Ruimin, Lin Hongfei, Yan Jun. Mining latent semantic on user-tag-item for personalized music recommendation [J]. Journal of Computer Research and Development, 2014, 51(10): 2270-2276 (in Chinese)
(李瑞敏, 林鸿飞, 闫俊. 基于用户-标签-项目语义挖掘的个性化音乐推荐[J]. 计算机研究与发展, 2014, 51(10): 2270-2276)

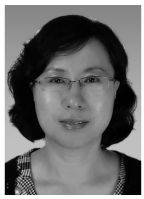
[20] Wu X, Zhang L, Yu Y. Exploring social annotations for the semantic Web [C] //Proc of the 15th Int Conf on World Wide Web. New York: ACM, 2006: 417-426

[21] Camarille R P, Sanchez L A, Nunez R D. Towards a semantic social network [C] //Proc of 2013 Int Conf on Electronics, Communications and Computing. Piscataway, NJ: IEEE, 2013: 74-77

[22] Quattrone G, Capra L, De Meo P, et al. Effective retrieval of resources in folksonomies using a new tag similarity measure [C]//Proc of the 20th ACM Int Conf on Information and Knowledge Management. New York: ACM, 2011: 545-550

[23] Yu W, Lin X, Zhang W. Towards efficient SimRank computation on large networks [C]//Proc of the 29th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2013: 601-612

Dong Hualei, born in 1990. Received her master's degree from Dalian University of Technology. Student member of China Computer Federation. Her main research interests include data mining, information retrieval and QA system (donghl@mail.dlut.edu.cn).



Wang Jian, born in 1967. PhD and associate professor in Dalian University of Technology. Senior member of China Computer Federation. Her main research interests include information retrieval, biomedical text mining.

Lin Hongfei, born in 1962. Professor and PhD supervisor in Dalian University of Technology. Member of China Computer Federation. His main research interests include information retrieval, data mining and natural language understanding (hflin@dlut.edu.cn).

Wang Hao, born in 1981. Received his master's degree from Dalian University of Technology. His main research interests include data mining and information retrieval (jerry8210@sina.com).