文章编号:1671 - 6833(2012)05 - 0114 - 04

一种基于社会化标注的查询扩展方法

王 健,李志斌,林鸿飞

(大连理工大学 计算机科学与技术学院,辽宁 大连 116024)

摘 要:利用互联网上的社会化标注信息来改善查询扩展效果,是目前信息检索领域的一个研究热点.根据社会化标注系统中数据的特点,提出了一种改进的加权社会化相似度算法,称作 Weighted SimRank (WSR)算法,用于改善查询扩展效果.WSR 方法在计算标签和网页之间边的权值时,既考虑与标签和网页共现的用户数量,又兼顾到被同一标签所标注过的不同网页数.所有的实验都是在从 del.icio.us 网站上抽取的真实标注数据集上进行的.实验结果表明,WSR 方法能够有效地衡量标签之间的相似度,与其他几种基于社会化标注的方法相比,可以获得更有用的查询扩展信息,明显地改善了查询扩展的效果.

关键词:查询扩展;社会化标注;SimRank 算法;标签相似度

中图分类号:TP391 **文献标志码**:A doi:10.3969/j.issn.1671 - 6833.2012.05.025

0 引言

在 web2.0 时代,互联网上涌现出很多社会化标注系统,例如 Last.fm,YouTube,Flickr,del.icio.us 等.在这些系统里,用户可以用标签来描述不同类型的互联网资源.系统中的数据是以四元组的形式存储的,包括时间、用户、标签和资源等四部分数据信息.社会化标注系统中的这些数据也逐渐吸引着研究者的兴趣,如文献[1]将其用于生成网页摘要,文献[2 - 4]利用社会化标注来改善网页检索性能.

作者利用社会化标注数据来改善网页检索中的查询扩展效果.首先需要计算标签之间的相似度,然后依据相关性从标签集合中选取扩展词.很多传统的方法可直接用于计算标签之间的相似度,例如 Cosine 相似度、Jaccard 系数和 Pearson 相关分析等.这些方法对于分布紧密的网络图形往往具有较好的效果,但在社会化标注系统中,用户的标注行为具有幂律分布的特性,具体描述见文献[5 - 6],而传统的相似度计算方法要求标签之间必须有共现的用户或者资源,才能得到一定的相似度,这样大大降低了查询扩展的效果.本文中提出一种改进的加权 SimRank (Weighted SimRank, WSR)算法,可以缓解标注系统中数据的稀疏性,从而能更好地计算标签之间的相关性.

1 基于社会化标注的查询扩展

1.1 WSR 算法的提出

如图 1 所示,网页和标签表示为两种类型的节点,节点之间的连边表示用户的标注行为.

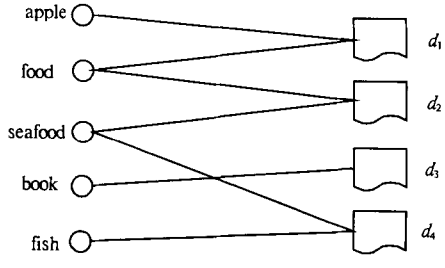

图 1 基于社会化标注的二部图

Fig.1 Bipartite graph of social tagging

图 1 中的标签“apple”和“food”都标注了网页“d_1”,标签“food”同时标注了网页“d_1”和“d_2”.以往的工作中,往往直接用与标签和网页共现的用户数量来表示标签与网页的边权值.WSR 算法采用的是一种新的边权值算法,它同时考虑了每个标签所标注过的网页数,如式(1)所示.

$$w(t,d) = tf_{t,d} * \log_2(N/n_t) \quad (1)$$

式中:$w(t,d)$为标签与网页之间边的权值;$tf_{t,d}$为

收稿日期:2012 - 03 - 28;**修订日期**:2012 - 05 - 17

基金项目:国家自然科学基金资助项目(60673039,60973068);国家“863”高科技计划资助项目(2006AA01Z151);教育部留学回国人员科研启动基金和高等学校博士学科点专项科研基金资助课题(20090041110002)

作者简介:王健(1967 -),女,黑龙江双城人,副教授,硕士,研究方向为搜索引擎、文本挖掘,wangjian@dlut.edu.cn.

标签 t 标注到网页 d 上的不同用户的数量;n_t 为标签 t 所标注过的网页总数;N 为系统中的网页总数.

WSR 算法描述如下:

Step1: For each tag pair (t_i, t_j) and each page pair (d_m, d_n) do

if $t_i = t_j$ $S_T^0(t_i, t_j) = 1$

otherwise $S_T^0(t_i, t_j) = 0$

if $d_m = d_n$ $S_T^0(d_m, d_n) = 1$

otherwise $S_T^0(d_m, d_n) = 0$

Step2: Repeat

For each tag pair (t_i, t_j) do

$$S_T^{K+1}(t_i,t_j) = C_T \sum_{m=1}^{|D(t_i)|}\sum_{n=1}^{|D(t_j)|} \frac{w(t_i,d_m)}{\sum_{t\in T(d_m)} w(t,d_m)} \frac{w(t_j,d_n)}{\sum_{t\in T(d_n)} w(t,d_n)} S_D^K(d_m,d_n) \quad (2)$$

For each web pair (d_m, d_n) do

$$S_D^{K+1}(d_m,d_n) = C_D \sum_{i=1}^{|T(d_m)|}\sum_{j=1}^{|T(d_n)|} \frac{w(t_i,d_m)}{\sum_{d\in D(t_i)} w(t_i,d)} \frac{w(t_j,d_n)}{\sum_{d\in D(t_j)} w(t_j,d)} S_T^{K+1}(t_i,t_j) \quad (3)$$

Until $\frac{||S_T^{K+1} - S_T^K||_2}{||S_T^K||_2} \leqslant \sigma$ and $\frac{||S_D^{K+1} - S_D^K||_2}{||S_D^K||_2} \leqslant \sigma$

Step3: Output $S_T(t_i,t_j)$ and $S_D(d_m,d_n)$

其中,C_T 和 C_D 是算法中相似度迭代计算过程中的消退因子,$w(t_i,d_m)$ 为标签 t_i 和网页 d_m 的边权值. $S_T^K(t_i,t_j)$ 和 $S_D^K(d_i,d_j)$ 分别为第 K 次迭代之后标签 t_i 与 t_j 以及网页 d_i 和 d_j 之间的相似度. $D(t_i)$ 表示标签 t_i 标注过的文档集. $T(d_i)$ 表示标注到文档 d_i 上的标签集. σ 是一个预定义的阈值,作者设置为 0.001.

1.2 WSR 算法的时间复杂度与收敛性

WSR 算法的时间复杂度为 $O(n(T^2N^2))$,其中 N 代表系统中的网页总数,T 表示标签的总数目,n 为算法迭代次数. 该算法的时间复杂度主要取决于标签和网页的数目,因为在实际应用中,算法往往能在较短的迭代次数内达到收敛.

WSR 算法的收敛性可用类似文献[7]中提出的方法进行证明.

1.3 查询扩展方法

假设 $Q = \{q_1, q_2, \cdots q_n\}$ 为一个由 n 个词构成的查询,则扩展词与该查询的相关性计算公式为

$$rel(t,Q) = \sum_{q_i\in Q} s_T(t,q_i) \quad (4)$$

式中:$rel(t,Q)$ 为扩展标签 t 与查询 Q 的相关性;$S_T(t,q_i)$ 为 WSR 算法得到的标签间的相似度.

根据公式(4)选取与查询 Q 最相关的前 K 个标签作为扩展词,然后再进行检索. 本文中 K 取值为 Q 中原始查询词的半数.

2 实验方法和结果分析

2.1 实验语料

本实验采用文献[8]所提供的数据集,下载自 del. icio. us 网站. 从初始数据集中过滤了网页内容少于 20 个词,以及标签数少于 5 个的网页,以减少数据的稀疏性. 同时,也过滤了使用次数少于 5 个的标签词和停用词. 对所有的标注词进行了词干化处理.

由于 del. icio. us 网上要求用户的一次标注行为只能有一个标注词,很多用户就会通过某种方式将多个词连接起来构成一个合成词,例如"social tagging"会被不同的用户组合成诸如"socialtagging"、"social - tagging"、"social_tagging"等不同形式的合成词. 作者对这种合成词进行了拆分. 经过预处理之后的数据集包含 38 060 个文档,4407 个用户和 26 556 个标签.

2.2 检索模型和评价方法

用户在将一组标签词标注到一个网页上时,可以理解为在该用户看来,该资源就是与其提供的标签词间相关的网页. 作者根据用户的标注心理来模拟检索过程,采用十倍交叉验证的方式对实验结果进行评价.

2.2.1 实验方法

首先,将系统中的原始数据整理成三元组的形式,即 <用户,标签集,网页>,并将所有的三元组随机分割成十份,每次都取其中的九份作为训练集,剩余的一份作为测试集. 在检索之前,需要为训练集中每一个三元组中的标签集选取扩展词,标签集经过扩展后的三元组作为新的训练集.

然后,将测试集中的每一个三元组中的标签集也进行扩展后作为查询. 例如,测试集中存在一个三元组 $<u, tSet, d>$,其中 $tSet$ 作为查询,网页 d 为该查询的相关文档. 网页与查询的相关性计算公式如下:

$$p(tSet|d) = \sum_{q\in tSet} tf_{q,d} * \log(N/n_q) \quad (5)$$

其中,$tf_{q,d}$ 表示将查询词 q 标注到网页 d 上的用户数,n_q 表示查询词 q 所标注过的网页总数,

N 为系统中网页的总数量.

根据公式(5)计算出的查询与网页的相关性对网页进行排序,然后查看排名最靠前的 N 个网页里是否包含该查询的相关网页 d. 如果包含网页 d,则说明扩展算法对于该查询来说是成功的.

最后,通过计算测试集中所有三元组的平均成功率来衡量一个扩展算法的好坏.

2.2.2 算法实现

为了检验查询扩展和标注信息的作用,首先实现了一个未进行扩展的检索模型作为比较对象. 为了验证采用 WSR 方法的有效性,我们在相同语料集上同时实现了其他三种扩展算法并进行了对比实验,三种算法分别是 Cosine 相似度衡量算法,SimRank 算法,以及文献[3]给出的基于 SimRank 的改进算法 SSR.

2.3 实验结果分析

实验结果如图 2 所示,其中记录了检索结果取前 N 个网页时,各个算法对应的平均成功率. 图 2 中为 N 分别取 5,10 和 20 的实验结果,针对这些结果做以下分析.

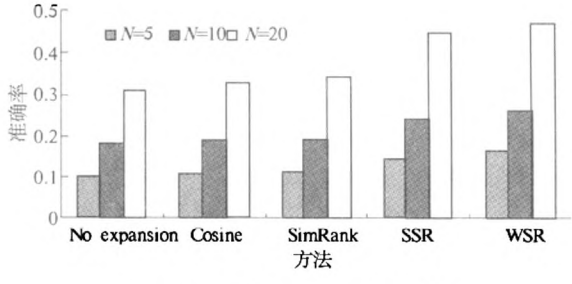

图 2 几种算法的平均检索准确率

Fig. 2 Average retrieval accuracy of each algorithm

2.3.1 查询扩展算法的检索效果

从图 2 中可以看出,没有进行查询扩展的实验效果最不理想,本文实现的任何一种查询扩展算法都能够在一定程度上提高检索效果.

一方面,社会化标注是标注者对网页的内容所做的概述,其标签词往往能从语义层面上概括网页的内容. 对于网页来说,标签是一种很好的数据源. 同时,社会化标注来源于互联网上的真实用户,相比较网页开发人员写出的网页,它在内容与形式上都更为丰富,同时查询一般也是出自互联网上的日常用户,它与标签的重合性往往非常高. 所以,将网页的标签信息用于查询扩展,能够较好地改善网页检索的性能.

另一方面,在社会化标注系统里,用户的标注行为会产生大量的链接关系. 例如,同一个用户可能会标注多个网页,这些网页就可以通过该用户而得到一定的相关性. 同样地,被同一个标签词标注过的网页也会根据标签而得到一定的相关性. 这些隐含在用户标注行为中的信息具有很高的价值.

2.3.2 几种算法的比较

在社会化标注系统里,用户的标注行为具有长尾效应,导致标注数据非常稀疏,而 Cosine 相似度算法本身并不适合数据稀疏的情况,所以该算法对应的实验效果并不理想.

SimRank 算法本身就是为缓解数据稀疏性而提出来的,但 SimRank 算法并没有将标签和网页的边权值进行量化. 而实际上标注到同一个网页上的多个标签与该网页之间的相关性是不同的,所以 SimRank 虽然较 Cosine 效果要好,但提升效果并不明显.

文献[3]的 SSR 算法根据与标签和网页共现的用户数对标签和网页的边权值进行了量化并引入到 SimRank 算法中. 可以看出,在引入边的权值后,算法的提升效果较为明显.

与 SSR 算法相比较,本文的 WSR 在计算标签和网页之间边的权值时,除了考虑与标签和网页共现的用户数外,还考虑了标签本身的区分度,即如果一个标签所标注的网页数量越多,它的区分度越低,则它与其所标注的网页之间的相关性越小. 另外,WSR 算法也是将边的权值直接引入 SimRank 算法中,使得边的权值能够很好的体现在算法的迭代过程当中. 从实验效果来看,WSR 算法要比 SSR 算法取得更好的检索效果.

为了进一步衡量采用上述几种算法进行查询的效果,随机抽取了一些标签词,分别用这些算法挖掘与之相关的标签词,按相关性大小排序的挖掘结果见表 1.

从表 1 中可以看出,与其他几种方法相比,WSR 算法挖掘到的相关性标签词排名更合理些,不相关的词也相对少些. 例如,对于标签词"program",普通 SimRank 算法得到的相关词依次为:"java new c xml linux dotnet good";SSR 算法得到的为:"script java exampl dotnet book linux",显然都不如用 WSR 算法获得的"java c databas php script xml python"合理与准确.

3 结束语

作者根据社会化标注系统中数据信息的特点,提出了一种改进的 SimRank 算法 WSR,将标

表 1　几种算法挖掘词的关联性比较

Tab. 1　The relevance contrast effect of mining word of each algorithm

标签词	算法	挖掘到的相关标签词
computer	WSR	pc internet os microsoft windows
	SSR	pc mac inform shop internet blog
	SimRank	pc Microsoft dell hp inform
	Cosine	pc like cool new dell internet
program	WSR	java c databas php script xml python
	SSR	script java exampl dotnet book linux
	SimRank	java new c xml linux dotnet good
	Cosine	c java good new linux dotnet book
cook	WSR	kitchen delicious food wine diet
	SSR	food diet easi tea healthi brunch
	SimRank	food healthi art rice easi diet kitchen
	Cosine	food rice diet breakfast delicious

签和网页之间边的权值引入 SimRank 算法中,解决了数据稀疏性问题. WSR 在量化边的权值时,不但考虑了与标签和网页共现的用户数量,还考虑了每个标签所标注的网页数量对相关性的降低影响. 实验结果表明,社会化标注网络是对查询扩展非常有用的信息源;采用本文的 WSR 算法进行查询扩展,可获取更为合理可靠的相似度信息,有效地提高了查询扩展系统的性能. 此外,本文的方法还可用于标签推荐和网页聚类等,这也是我们未来要继续探讨的工作.

参考文献:

[1] 尚书姐,王灿,朱俊彦. 一种依据标签的网页摘要方法[J]. 计算机程,2010,36(21):260 - 261,264.

[2] HOTHO A, JAESCHKE R, SCHMITZ C, et al. Information Retrieval in Folksonomies: Search and Ranking [C]// Proceedings of Extended Semantic Web Conference, Budva, Serbia Monteneg. 2006:411 - 426.

[3] BAO Sheng-hua, XUE Gui-rong, WU Xiao-yuan, et al. Optimizing web search using social annotations [C]// Proceedings of the 16th international conference on World Wide Web. Banff, Alberta, Canada. 2007:501 - 510.

[4] XU Sheng-liang, BAO Sheng-hua, CAO Yun-bo, et al. Using Social Annotations to Improve Language Model for Information Retrieval [C]. Lisbon, Portugal: The 16nd ACM CIKM Conference, 2007. 1003 - 1006.

[5] CATTUTO C, SCHMITZ C, BALDASSARRI A, et al. Network properties of folksonomies [J]. AI Communications, 2007, 20(4):245 - 262.

[6] MEO P D, 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.

[7] GLEN J, JENNIFER W. SimRank: A measure of structural-contextsimilarity [C]. Edmonton, Canada: Proceedings of SIGKDD, 2002. 538 - 543.

[8] LU Cai-mei, HU Xiao-hua, PARK E K. Exploit the tripartite network of social tagging for web clustering [J]. IEEE Transactions on Systems, Man and Cybernetics, Part A (Systems and Humans), 2011, 41(5): 840 - 852.

An Approach of Query Expansion Based on Social Tagging

WANG Jian, LI Zhi-bin, LIN Hong-fei

(School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China)

Abstract: How to use the social tagging information to improve the effect of query expansion is a current research hotspot in the information retrieval field. In this paper, according to the data characteristics of social tagging systems, we propose a modified social similarity algorithm called "Weighted SimRank" (WSR), which is used to improve the effect of query expansion. When the edge weighted values between labels and web pages are calculated, the WSR algorithm takes into account the number of co-occurrence users with tags and web pages as well as the number of different web pages labeled by every same tag. All the experiments are carried out on a real-world annotation data set which is sampled from the website del. icio. us. The experimental results show that our proposed WSR method can effectively measure the similarity of annotations. Compared to the other social-annotation-based methods, WSR produces more useful query expansion information and achieves better performance.

Key words: query expansion; social tagging; SimRank algorithm; tag similarity