基于复杂网络的社会化标签语义相似度分析

张昌利[1]，龚建国[2]，闫茂德[2]

(1. 长安大学信息工程学院 西安 710064; 2. 长安大学电子与控制工程学院 西安 710064)

【摘要】针对社会化标签系统所对应的标签共现复杂网络，引入标签语义相似度权值和抽象权值算子，建立了标签语义相似度计算模型。相比基于“用户-对象-标签”三元组的统计性计算公式或基于复杂网络拓扑结构的节点相似性计算公式，本模型可以在标签语义相似度计算中将标签标注行为的统计特性与复杂网络的拓扑特性有机地结合起来，形成一个具有良好数学性质的形式化系统。仿照模糊逻辑中T范数、S范数给出了抽象权值算子的具体化实现，形成具体化算子簇，可以通过调节参数(如参数h和阶数l)形成不同类型或不同全局性的具体化算子。设计实验方案，利用复杂网络链路预测的AUC指标、Precision指标对典型算子及算子簇进行了综合分析。分析结果表明，这些具体化算子同时具有“语义补充”、“语义破坏”两种相反作用，在算子阶数较低(如$2\leq l\leq 5$)时能明显提高标签语义相似度计算的准确性，在社会化标签系统的高精确性个性化推荐算法设计中具有应用价值。

关 键 词 复杂网络； 链路预测； 算子； 社会化标签系统； 标签语义相似度

中图分类号 TP391 文献标识码 A doi:10.3969/j.issn.1001-0548.2012.05.001

Complex Network Based Semantic Similarity Measure for Social Tagging Systems

ZHANG Chang-li[1], GONG Jian-guo[2], and YAN Mao-de[2]

(1. School of Information Engineering, Chang'an University Xi'an 710064;

2. School of Electronics & Control Engineering, Chang'an University Xi'an 710064)

Abstract Regarding to the complex network composed of the vast amount of tags in social tagging systems in Internet with their co-occurrences, the weights as the statistical semantic similarity of tag-tag edges and two abstract operators for weights computation were introduced, and a model of tag semantic similarity measurement is established. Comparing with traditional "users-items-tags" tripartite graph based statistic measures or network topology focused nodes similarity measures, this model provides a well defined formal system, which explicitly addresses both the statistical influential factors and the topological influential factors in computation of tag semantic similarities. A cluster of concrete implementations of the abstract operators are devised, which have similar format with T norms and S norms in fuzzy logics. In this cluster, concrete operators of different types or addressing different scopes of network topological factors are configurated with particular parameters (e.g., parameter h and order l). By incorporating the AUC index and precision index in link prediction of complex network, an experiment is conducted to analyze the effectiveness and feasibility of these concrete operators. The experimental results show that these concrete operators introduce the effects of "semantic complementation" as well as the effects of "semantic destruction" when they are applied, but lower ordered calculations (e.g., $2\leq l\leq 5$ in the model) with these operators are helpful for precise analysis of tag semantic similarities, therefore they are useful in devising high accurate tag-aware recommendation algorithms for social tagging systems.

Key words complex network; link prediction; operators; social tagging system; tag semantic similarity

社会化标签系统是Web2.0的核心构件，鼓励互联网用户自发地创建、选择和运用标签，可以充分地发挥用户的集体智慧[1]。目前，社会化标签系统作为自动描述、组织和挖掘海量互联网资源的有效工具，得到了科学界和工业界的广泛关注[2]。例如，由于标签中所包含的丰富的语义知识和用户个性化信息，基于社会化标签系统的个性化推荐被认为是解决Web2.0信息过载问题的有效手段[3]。但是，用

收稿日期： 2011 – 11 – 28; 修回日期：2012 – 06 – 18

基金项目：交通运输部应用基础研究项目(2011319812400)；中央高校基本科研业务费专项资金(CHD2012JC022).

作者简介：张昌利(1979 –)，男，博士，主要从事于复杂网络与复杂系统、互联网信息计算和交通信息集成等方面的研究.

户操作的随意性又导致社会化标签系统中标签组织混乱和语义模糊，如何提取标签的隐含语义，并重建社会化标签系统所蕴含的结构化知识体系，成为目前互联网发展所面临的重要问题，对于构建高精确性的个性化推荐算法也具有非常重要的意义[4]。对此，现阶段研究主要利用用户、标签、网络资源(对象)之间形成的三元关系，构造社会化标签系统的三部图模型，从图的结构、演化和功能等方面展开。例如，利用社会化标签系统的“用户-对象-标签”三部图模型所蕴涵的标签属性信息，统计计算成对标签的相似度，从而在标签之间建立起松散的语义关系[5-7]。

复杂网络是对复杂系统的高度抽象，是建模和分析复杂动态系统的一个强大而有效的工具[8]。已有研究表明，海量社会化标签之间通过共现关系(同一用户标注同一互联网资源而同时使用的标签具有共现关系)形成了复杂网络[9-10]。标签相似度也代表了一种开世界语义，因此标签的语义相似度分析还应基于复杂网络理论从总体上考虑众多标签之间根据复杂网络拓扑结构所产生的交互影响。鉴于此，本文将基于“用户-对象-标签”三部图统计计算的标签语义相似度看作标签共现复杂网络的权值，引入抽象权值算子建立了标签语义相似度的形式化计算模型，在利用统计计算结果的同时，还可以融入标签语义复杂网络的拓扑结构影响。进而，参考模糊逻辑中的T范数、S范数给出了抽象权值算子的具体化实现，并利用复杂网络链路预测的AUC指标、Precision指标对不同具体化算子下标签语义相似度的综合计算结果进行了实验分析。实验结果表明，这些具体算子会同时带来“语义补充”和“语义破坏”两种相反作用，但在计算的阶数较低(如$2\le l\le 5$)时能够提高标签语义相似度计算的准确性。

1　标签共现复杂网络与标签语义相似度计算综述

根据用户向互联网资源标注标签的行为，社会化标签系统可抽象为三部图模型$\mathbb{F}(U,I,T;Y)$，其前3个元素分别表示用户、对象和标签的有限集合，Y为标签标注行为的三元关系集合[5]。对于任意的三元组$(u,i,t)\in U\times I\times T$，$Y(u,i,t)=1$表示该三元组满足标注关系，否则$Y(u,i,t)$取值为0。

标签语义相似度是对标签之间共同语义特征的量化表示[11]。利用社会化标签系统的三部图模型，许多研究从标注行为统计的角度给出了标签语义相似度的计算方法[5-7]。例如，假设

$$B(x)=\{(u,i)\,|\,(u,i)\in U\times I\wedge Y(u,i,x)=1\}\qquad(1)$$

表示标签$x\in T$对应的标注行为集合(其元素为用户、资源二元组)，文献[6]根据集合重叠比例提出了标签语义相似度计算公式：

$$S(x,y)=\frac{|B(x)\cap B(y)|}{\sqrt{|B(x)|\times|B(y)|}}\qquad(2)$$

式中，$|\cdots|$用于计算集合大小。但是，社会化标签系统的三部图模型将用户、对象、标签看做3类不同的节点，边仅存在于不同类别的节点之间，从而割裂了同类节点之间的共现关系，且不可避免地造成了信息丢失。例如，式(2)对不共现的标签节点的计算结果均为0，显然与实际情况不符。

在社会化标签系统三部图模型的基础上，还可定义标签共现网络$\mathbb{G}(T,E)$，其中T为标签集，$E\subseteq T\times T$为标签共现关系集。显然$\mathbb{G}$是仅由社会化标签系统中标签节点组成的无向图。文献[8,11]通过实证分析，分别独立证明了$\mathbb{G}$具有小世界和无标度等特征，是一种典型的复杂网络。还可以发现，$\mathbb{G}$的拓扑结构也刻画了标签的语义关联关系。因此，根据复杂网络的节点相似性理论[5,11]，在$\mathbb{G}$上还可进一步定义基于复杂网络拓扑结构的标签语义相似度计算公式如表1所示。

表1　10种基于复杂网络拓扑结构的标签语义相似度计算公式(假设$x,y\in T$)

名称	定义
共同邻居公式	$s^{\mathrm{CN}}(x,y)=\lvert\Gamma(x)\cap\Gamma(y)\rvert$
Salton 公式	$s^{\mathrm{Salton}}(x,y)=\dfrac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\sqrt{k(x)+k(y)}}$
Jaccard 公式	$s^{\mathrm{Jaccard}}(x,y)=\dfrac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Sørensen 公式	$s^{\mathrm{Sørensen}}(x,y)=\dfrac{2\lvert\Gamma(x)\cap\Gamma(y)\rvert}{k(x)+k(y)}$
大度节点有利公式	$s^{\mathrm{HPI}}(x,y)=\dfrac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\min\{k(x),k(y)\}}$
大度节点不利公式	$s^{\mathrm{HDI}}(x,y)=\dfrac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\max\{k(x),k(y)\}}$
Adamic-Adar 公式	$s^{\mathrm{AA}}(x,y)=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\dfrac{1}{\lg k(z)}$
网络资源分配公式	$s^{\mathrm{RA}}(x,y)=\sum_{z\in\Gamma(x)\cap\Gamma(y)}\dfrac{1}{k(z)}$
Katz 公式	$s^{\mathrm{Katz}}(x,y)=\sum_{l=1}^{\infty}\beta^l(\boldsymbol{A}^l)_{xy}=(\boldsymbol{I}-\beta\boldsymbol{A})^{-1}-\boldsymbol{I}$
LP 公式	$s^{\mathrm{LP}}(x,y)=(\boldsymbol{A}^2)_{xy}+\varepsilon(\boldsymbol{A}^3)_{xy}$

其中，$\Gamma(x)$表示任意标签$x\in T$的相邻节点集，$k(x)=|\Gamma(x)|$；$\boldsymbol{A}$表示$\mathbb{G}$的邻接矩阵，$(\boldsymbol{A}^l)_{xy}$表示x、y之

间长度为l的路径的数目，$\boldsymbol{I}$为单位矩阵；β、ε为权值因子。

已有研究表明，基于复杂网络拓扑结构的节点相似性计算公式往往侧重刻画了复杂网络的某一侧面拓扑结构特征[11]。例如，表1中前8个公式更多考虑了共同邻居等局部拓扑结构信息，后两个公式侧重考虑了全局拓扑结构的影响。总体上看这些公式各有优劣，但实验分析表明CN、AA、RA、Katz等指标对于常见复杂网络具有更佳的表现。但是，对于社会化标签系统而言，标签语义相似度的取值同时依赖于标签标注行为的统计特性及标签共现复杂网络的拓扑结构特性，前者起主导作用，后者反映了众多标签语义关联随网络拓扑结构所产生的交叉影响。而现有统计性计算公式或基于复杂网络拓扑结构的计算公式仍不能将两个侧面关联起来，无法对标签语义相似度值进行综合计算。

2 基于复杂网络的标签语义相似度计算模型

利用基于“用户-对象-标签”三部图统计计算的标签语义相似度计算结果，为标签共现复杂网络$\mathbb{G}$引入权值。定义抽象权值算子，用于在标签语义相似度计算中融入$\mathbb{G}$拓扑结构的影响因素。仿照模糊逻辑中T范数、S范数的定义，给出了抽象权值算子的具体化实现，形成了标签语义相似度计算的具体化算子簇。

2.1 模型定义

将标签共现复杂网络$\mathbb{G}$扩展为加权无向网络$\mathbb{G}'(T,E;s)$。其中，$s{:}E\to[0,1]$为边的权值函数，其值是对基于“用户-对象-标签”三部图统计计算的标签语义相似度(如式(2)的计算结果)的[0,1]标准化处理，反映了标签标注行为对共现标签语义相似度的影响程度。并为$\mathbb{G}'$引入抽象权值算子⊗和⊕，满足：对于任意标签a、$b\in T$，某长度为n的路径p包含的标签有$t_0=a$, t_1, t_2,…, $t_{n+1}=b\in T$，则该路径对应的综合权值为$s(p)=\otimes_{k=0}^{n}s(t_k,t_{k+1})$；假设$a$、$b$间的所有路径有$p_1$，$p_2$, …，$p_m$，则综合权值为$s_\Sigma(a,b)=\oplus_{k=0}^{m}s(p_k)$。例如，假设某标签共现复杂网络中，标签$a$、$b$、$d$组成长度为2的路径，两个边的权值依次为$s(a,b)$和$s(b,d)$，则$s_{abd}=s(a,b)\otimes s(b,d)$为该路径的等效权值，其大小反映了该路径对标签$a$、$d$语义相似度的影响能力；假设$a$、$d$之间另一条长度为2的路径由标签$a$、$c$、$d$组成，其等效权值为$s_{acd}=s(a,c)\otimes s(c,d)$，则两条路径的综合权值为$s_{abd}\oplus s_{acd}$，其大小反映了两条路径对$a$、$d$语义关联的综合影响。

将上述算子扩展到矩阵形式。对于n阶矩阵$\boldsymbol{A}$、$\boldsymbol{B}$，仿照传统的矩阵乘、矩阵加运算法则定义矩阵运算$\boldsymbol{A}\otimes\boldsymbol{B}$、$\boldsymbol{A}\oplus\boldsymbol{B}$。记$\mathbb{G}'$的邻接矩阵为$\boldsymbol{G}$，则$\boldsymbol{G}$是对角线为1的对称阵，其余元素为相应边的权值。令$\boldsymbol{G}^0=\boldsymbol{I}$、$\boldsymbol{G}^{l+1}=\boldsymbol{G}^l\otimes\boldsymbol{G}$，则$\boldsymbol{G}^l(l>0)$反映了长度为$l$的路径所产生的综合语义效应。令$\boldsymbol{G}^{(0)}=\boldsymbol{G}^0$、$\boldsymbol{G}^{(l+1)}=\boldsymbol{G}^{(l)}\oplus\boldsymbol{G}^l$，则$\boldsymbol{G}^{(l)}(l>0$，称为阶数)表示所有长度不大于$l$的路径的综合语义效应，其中$\boldsymbol{G}^{(1)}$正好对应于基于统计的共现标签语义相似度计算结果。进而，$\mathbb{G}'$整体拓扑结构所产生的综合权值则对应于公式

$$\boldsymbol{G}^{(+)}=\lim_{l\to\infty}\boldsymbol{G}^{(l)} \tag{3}$$

2.2 抽象权值算子的具体化实现

结合$\mathbb{G}'$的拓扑结构可以发现，抽象算子⊗、⊕应具有类似于模糊数学中T范数、S范数的良好数学性质，如表2所示。以算子的分配性为例，假设某标签共现复杂网络中标签a、e之间有两条长度为3的路径，路径a、b、d、e的等效权值为$s_{abde}=s_{abd}\otimes s(d,e)$，路径$a$、$c$、$d$、$e$的等效权值为$s_{acde}=s_{acd}\otimes s(d,e)$，两条路径的综合权值为$s_{abde}\oplus s_{acde}$；按照另一种思路，$s_{abd}\oplus s_{acd}$为路径$a$、$b$、$d$和路径$a$、$c$、$d$综合的等效权值，这样$(s_{abd}\oplus s_{acd})\otimes s(d,e)$也应表示$a$、$e$间同样的综合权值。

因此，可以借鉴S范数和T范数来定义⊗、⊕算子的具体化实现，如表3所示。其中，Zadeh算子具有最好的数学特性，必然存在一个$k<=n$使得$\boldsymbol{G}^{(+)}\equiv\boldsymbol{G}^{(k)}$。对于概率、边界、突变等其它具体算子，可取$\boldsymbol{G}^{(+)}\approx\boldsymbol{G}^{(n)}$。进而，仿照文献[12]对T范数簇、S范数簇的定义，还可通过式(4)和式(5)建立⊗、⊕的具体化算子簇。式中，参数h取值为1、0.75、0.5和0时分别对应于Zadeh、概率、边界、突变4种具体化算子，ite{…}为条件表达式。

$$\begin{aligned}x\otimes_h y=\mathrm{ite}\{&(4h-3)(x\otimes_{\mathrm{Z}}y)+\\&(4-4h)(x\otimes_{\mathrm{P}}y)\,|\,h\geq0.75;\\&(4h-2)(x\otimes_{\mathrm{P}}y)+\\&(3-4h)(x\otimes_{\mathrm{B}}y)\,|\,h\geq0.5;\\&(2h)(x\otimes_{\mathrm{B}}y)+(1-2h)(x\otimes_{\mathrm{D}}y)\}\end{aligned} \tag{4}$$

$$\begin{aligned}x\oplus_h y=\mathrm{ite}\{&(4h-3)(x\oplus_{\mathrm{Z}}y)+\\&(4-4h)(x\oplus_{\mathrm{P}}y)\,|\,h\geq0.75;\\&(4h-2)(x\oplus_{\mathrm{P}}y)+\\&(3-4h)(x\oplus_{\mathrm{B}}y)\,|\,h\geq0.5;\\&(2h)(x\oplus_{\mathrm{B}}y)+(1-2h)(x\oplus_{\mathrm{D}}y)\}\end{aligned} \tag{5}$$

表2　标签共现复杂网络上抽象算子 $\otimes$、$\oplus$ 的数学特性(假设 $x, y, z \in [0,1]$)

数学特性	算子 $\otimes$	算子 $\oplus$
交换性		$x \oplus y = y \oplus x$
结合性	$(x \otimes y) \otimes z = x \otimes (y \otimes z)$	
折扣性	$x \otimes y \leq \min\{x, y\}$	
聚合性		$x \oplus y \geq \max\{x, y\}$
保序性	$x \geq y \rightarrow x \otimes z \geq y \otimes z, z \otimes x \geq z \otimes y$	$x \geq y \rightarrow x \oplus z \geq y \oplus z$
吸收性	$x \otimes 0 = 0 \otimes x = 0$	$x \oplus 1 = 1$
不变性	$x \otimes 1 = 1 \otimes x = x$	$x \oplus 0 = x$
分配性	$z \otimes (x \oplus y) = (z \otimes x) \oplus (z \otimes y), (x \oplus y) \otimes z = (x \otimes z) \oplus (y \otimes z)$	

表3　具体化算子示例(假设 $x, y \in [0,1]$)

名称	算子 $\otimes$	算子 $\oplus$
Zadeh 算子	$x \otimes_{Z} y = \min(x, y)$	$x \oplus_{Z} y = \max(x, y)$
概率算子	$x \otimes_{P} y = xy$	$x \oplus_{P} y = x + y - xy$
边界算子	$x \otimes_{B} y = \max(0, x + y - 1)$	$x \oplus_{B} y = \min(1, x + y)$
突变算子	$x \otimes_{D} y = \text{ite}\{\min(x, y) \mid \max(x, y) = 1; 0\}$	$x \oplus_{D} y = \text{ite}\{\max(x, y) \mid \min(x, y) = 0; 1\}$

表4　从Flickr网站抓取的实验数据示例(对角线元素为单个标签数据，下三角阵数据为成对标签数据)

标签数据	band	concert	festival	live	music	show
band	3 323 458					
concert	1 000 882	6 089 510				
festival	229 341	629 268	6 465 155			
live	1 943 969	3 184 463	1 097 228	5 565 027		
music	2 770 926	3 560 618	2 470 157	4 913 849	9 820 872	
show	454 319	903 002	217 186	858 353	919 628	5 214 940

3　实验分析

Flickr是雅虎公司旗下的一款在线数字照片共享网站，用户在发布照片的同时要求为照片附加多个描述性标签。为了验证上述基于复杂网络的社会化标签语义相似度计算的准确性，本文针对Flickr网站一直以来的最热门标签(共143个，见网址www.flickr.com/photos/tags/)，设计爬虫程序分别抓取了与每个单独标签及成对标签所对应的照片个数(正好对应于标签标注行为的数量)。作为示例，表4给出了一小部分所抓取的数据。利用抓取的Flickr热门标签数据，本节基于复杂网络的链路预测技术设计实验方案，通过客观的链路预测指标对表3的4种具体化算子及式(4)、式(5)所对应的具体化算子簇进行实验分析，并与表1所列的基于复杂网络拓扑结构的节点相似性计算公式进行对比。

3.1　实验方案

复杂网络的链路预测指基于已知的网络拓扑结构预测未知或未来出现的网络链路的技术，可以用于客观地分析和评价复杂网络节点相似性算法的准确性[11]。以文献[13]提出的链路预测的10重交叉验证为蓝本，本文实验的实施步骤如下：

1) 假设 $\mathbb{G}'(T, E; s)$ 表示抓取的Flickr热门标签所形成的加权复杂网络，根据式(2)计算 $\mathbb{G}'$ 中所有边的权值。计算结果显示，$\mathbb{G}'$ 几乎为全互联网络，仅极少量节点间无链路，这与所选用标签为Flickr热门标签有关。

2) 对 $\mathbb{G}'$ 进行稀疏化处理，使 $\mathbb{G}'$ 最终共包含7 000条边，其它边均看作不存在边(共3 153条)。将 $\mathbb{G}'$ 的所有边随机地划分到10个子集，每个子集均为700条边。

3) 针对10个子集循环执行如下操作：依次选择当前子集作为测试集 E^{P}，其余9个子集合并形成训练集 E^{T}，显然 $E = E^{\mathrm{T}} \cup E^{\mathrm{P}}$ 且 $E^{\mathrm{T}} \cap E^{\mathrm{P}} = \varnothing$；利用标签语义相似度公式计算 $\mathbb{G}'$ 所有节点之间的相似度，作为对应边的分数值，并计算AUC(area under the receiver operating characteristic curve)、Precision等指标的取值。

4) 分别计算10重循环操作所得AUC指标、Precision指标的平均值，作为标签语义相似度公式的最终评价依据。

这里，AUC指标[14]表示测试集中的边的分数值比随机选择的不存在的边的分数值高的概率，其计算公式为：

$$\mathrm{AUC}=\frac{n'+0.5n''}{n} \tag{6}$$

式中，n表示总的独立比较次数，n'表示测试集中边的分数值更大的次数、n''表示测试集中边与不存在边分数值相等的次数。Precision指标[15]表示分值靠前的预测边中测试边被预测准确的比例。假设前L条预测边中有m条边在测试集，则Precision指标定义为

$$\mathrm{Precision}=\frac{m}{L} \tag{7}$$

实验中，式(6)中选择 n=700×3 153，表示对所有测试边与所有不存在边进行两两对比；式(7)中选择 L=700，等于测试集的大小。

3.2 具体化算子实验

首先，针对表3所给出的4种具体化算子，按照1≤l≤143的不同阶数计算$\boldsymbol{G}^{(l)}$所对应的AUC、Precision指标。其中，$\boldsymbol{G}^{(1)}$恰好为基于统计的标签语义相似度计算结果，从$\boldsymbol{G}^{(2)}$到$\boldsymbol{G}^{(143)}$的递增表明更多地考虑了标签共现复杂网络全局拓扑结构的影响。实验结果分别如图1和图2所示。

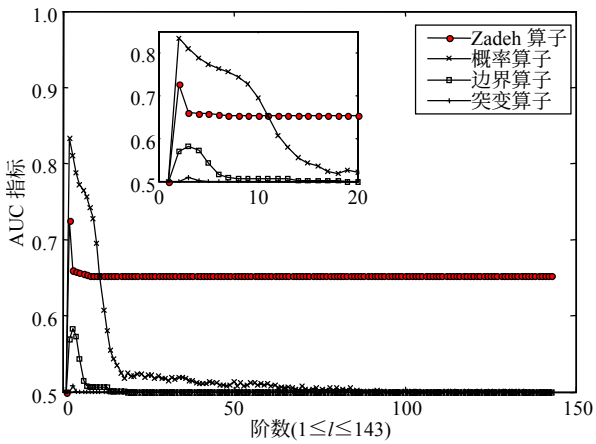

图1 4种具体化算子在不同阶数下的AUC指标
(内嵌图：阶数l≤20时的详细情况)

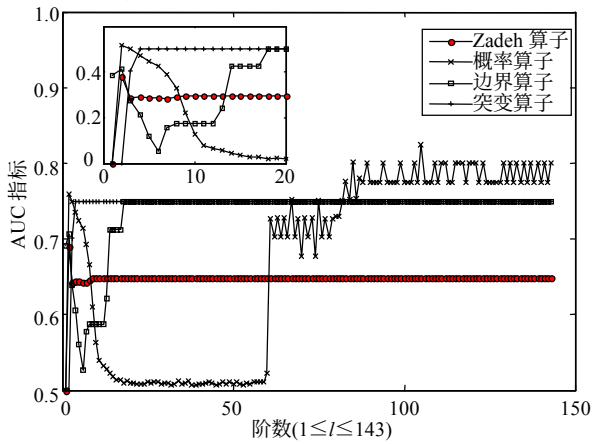

图2 4种具体化算子在不同阶数下的Precision指标
(内嵌图：阶数l≤20时的详细情况)

从图1和图2可见，4种具体算子都在低阶(约2≤l≤5)时有更佳的链路预测表现。随着阶数的上升，所有算子的AUC指标、Precision指标均迅速降低。Zadeh算子的AUC指标和Precision指标均在阶数l≥10后趋于固定，与前述该算子的收敛特性相一致。在阶数较高(l≥10)时，概率算子、边界算子、突变算子的AUC指标值取向0.5，表明这些高阶算子的链路预测能力降低至与随机预测水平相当；其Precision指标取值则异常地固定于0.5或在0.5上下跳变，观察实验数据发现针对某些测试集该指标值趋近于1，针对其它测试集该指标又趋近于0，平均值则在0.5左右。本实验说明：标签共现复杂网络对标签语义相似度的影响更多地体现在局部拓扑结构上；4种算子在调节不同全局程度的拓扑结构因素的影响时会同时带来“语义补充”和“语义破坏”两种相反效果，在阶数小时主要体现为“语义补充”，阶数大时更多产生的是“语义破坏”；Zadeh算子和低阶概率算子相对有更好的表现，高阶的概率算子、边界算子、突变算子的Precision指标出现异常值，可能是这些算子与某些未知的网络拓扑特征具有强相关性所致。

3.3 扩展算子簇实验

针对$\otimes_h$、$\oplus_h$算子簇执行类似的实验过程，得到如图3、图4所示的实验结果。图3所示为 AUC 指标随阶数 l 和参数 h 的变化情况，在阶数较低(约2≤l≤5) 或 h 参数偏大(约 h≥0.75)时 AUC 指标取值较好，在 l=2且 h=0.75时取到最大值，其它情况下 AUC 指标值迅速下降到0.5左右，此时链路预测效果仅与随机预测相当。图4所示为 Precision 指标随 l、h 的变化情况，图中曲面中存在一个明显的低谷，将整个曲面划分为两个部分。由图4可见，在阶数较低(约2≤l≤5)或 h 参数偏大(约 h≥0.75)时 Precision 指标也有较好的取值；但 h<0.75时在高阶(约 l≥10)形成一个取值约0.5的平台，与图2中出现的异常情况类似，此时实验数据中针对某些测试集的 Precision 指标趋近于1，针对其它测试集该指标又趋近于0，平均值则在0.5左右。

通过该实验可以得到与3.1节类似的结论。例如，标签共现复杂网络对标签语义相似度的影响更多地体现在局部的拓扑结构上，随着算子阶数的上升，AUC 指标、Precision 指标均迅速降低。导致上述现象的另一个可能的原因是，这些算子会同时带来“语义补充”和“语义破坏”效果，在阶数小时主要体现为“语义补充”，阶数大时更多产生的是“语义破坏”。此外，该实验还有助于优选具有最佳效果

的标签语义相似度算子。例如，通过对实验结果的综合比较可见，参数 l=2、h=0.75所对应算子的总体表现最佳。

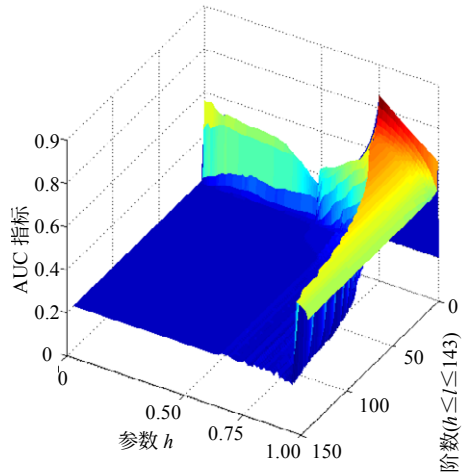

图3　不同h值、不同阶数下的AUC指标分布

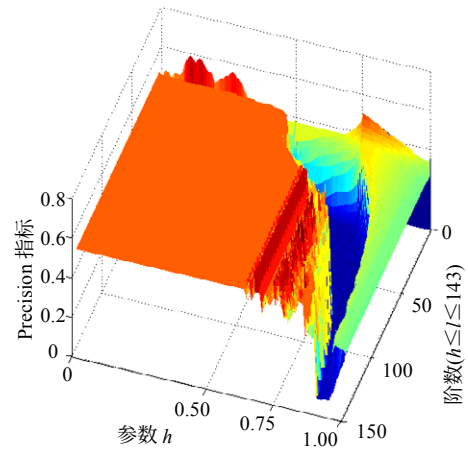

图4　不同h值、不同阶数下的Precision指标分布

3.4　部分算子与典型方法的比较

参照3.2节的实验结论，本节选择一些典型的标签语义相似度算子，与表1中的公式进行对比。根据实验中所有算子在求取标签语义相似度的总体表现，本文从低阶($2\leq l\leq 5$)范围选择算子。图5给出了低阶算子中所出现的最大 AUC 指标、最大 Precision 指标随参数 h 的变化情况。如图5所示，根据两条曲线中的5个极值点可以得到4个标签语义相似度算子，其中右侧两个极值点同时对应于2阶概率算子(h=0.75，l=2)。

表5分别给出了4个算子及 CN、AA、RA、Katz 等公式所计算的 AUC 指标和 Precision 指标。通过对比可见，AA、RA 公式具有最佳表现；4号算子(即2阶概率算子)的计算结果好于 Katz 公式，与 CN 公式相比则各有优劣；2号算子具有最大的 Precision 指标值，但 AUC 指标接近0.5，与1号算子、3号算子相比则各有优劣。总体来看，本文所选4个算子中2阶概率算子最优。

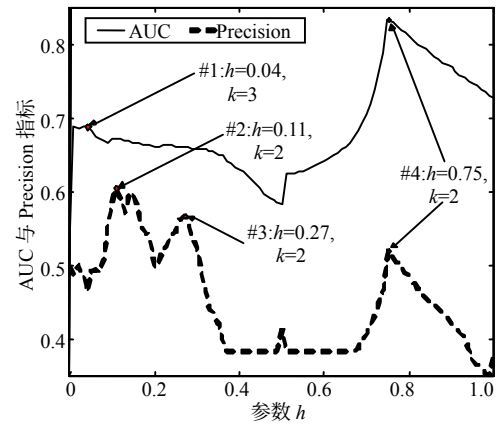

图5　利用低阶($2\leq k\leq 5$)算子计算所的最大AUC指标、最大Precision指标随参数h的变化情况

表5　不同标签语义相似度计算算子(公式)的AUC指标、Precision指标对比

算子	AUC	Precision	公式	AUC	Precision
#1	0.687 9	0.363 9	CN	0.850 0	0.403 2
#2	0.506 4	0.604 6	AA	0.850 2	0.601 9
#3	0.534 3	0.567 5	RA	0.850 0	0.602 0
#4	0.833 0	0.519 3	Katz	0.831 7	0.518 0

4　结 束 语

针对社会化标签系统所对应的标签共现复杂网络，引入标签语义相似度权值和抽象的权值算子，提出了综合考虑标签标注行统计特性与复杂网络拓扑特性的标签语义相似度计算模型，具有良好的数学性质。在此基础上，给出了典型的具体化算子及具体化算子簇，可以通过参数调节形成不同类型或不同全局性的具体化算子。并且，这些模型、算子及算子簇同时也适用于其他形式复杂网络的节点相似度计算。

本文模型与Katz公式、LP公式等在形式上具有相似之处。不同的是，本文模型同时考虑了标签标注行统计特性与复杂网络拓扑特性两方面因素，因此参与计算的矩阵为权值矩阵，且要求所有矩阵元素取值在[0,1]范围。如果将抽象算子⊗、⊕直接解释为算术加、乘运算后，本文模型与Katz公式会更为类似，但运算产生的中间矩阵的元素取值会超出[0,1]范围，表2中部分数学性质相应失效。此外，Katz公式、LP公式均使用参数限定高阶矩阵在形成最终相似度值时的权重，为了保持良好的数学性质，本文模型暂未采取类似方式，仅能通过优选算子达到这一目的。

针对 Flickr 热门标签数据设计实验方案，利用复杂网络链路预测的 AUC 指标、Precision 指标对典型的具体化算子及算子簇进行了综合分析。分析表明，标签共现复杂网络对标签语义相似度的影响主要体现于局部拓扑结构，低阶(如 $2 \le l \le 5$)算子能明显提高标签语义相似度计算的准确性，其中，2 阶概率算子具有最佳的表现，其标签语义相似度计算结果可以在个性化推荐系统中予以应用。

实验分析还表明，优选的标签语义相似度算子与典型的复杂网络节点相似性计算公式的计算效果相当，但不够突出。其原因在于：1) 本文计算模型虽然引入了标签标注行统计计算结果作为标签共现复杂网络的权值，但权值对最终标签语义相似度的影响情况仍比较复杂，例如存在“弱连接效应”[16]等现象；对此目前尚无成熟的研究结论可供借鉴，因此现有算子仍不能显著地体现出引入权值的优势；2) 在计算标签语义相似度时，本文模型仅依靠具体化算子限定不同全局性的复杂网络拓扑特征的影响能力，现有算子大多同时导致“语义补充”和“语义破坏”两重作用，其中高阶算子的“语义破坏”作用更为明显，而低阶算子(特别是 2 阶概率算子)则主要体现“语义补充”为作用。因此，在后期研究中仍有必要扩展现有模型或进一步寻找更佳的具体化算子。

此外，本文抓取 Flickr 网站的最热门标签数据作为实验数据，这些标签由于在足够多的标注行为中出现，因此其数据能够充分地反映这些标签的语义关联。与热门标签相比，冷门标签同样包含有丰富的语义信息，本文为了模拟这一情况，在实验中对热门标签所形成的标签共现复杂网络进行了稀疏化处理，最终选择的边数约为原总边数的 70%。在后期研究中，作者还将针对冷门标签抓取更多的实验数据，通过更为翔实的实验分析进一步验证本文的结论。

参 考 文 献

[1] CATTUO C, LORETO V, PIETRONERO L. Semantic dynamics and collaborative tagging[J]. Proceedings of the National Academy of Sciences, 2007, 104(5): 1461-1464.

[2] ISABELLA P. Folksonomies, indexing and retrieval in Web 2.0[M]. Berlin: De Gruyter Saur, 2009.

[3] ZHANG Z K, ZHOU T, ZHANG Y C. Tag-aware recommender systems: a state-of-the-art survey[J]. Journal of Computer Science and Technology, 2011, 26(5): 767-777.

[4] LIMPENS F, GANDON F, BUFFA M. Bridging ontologies and folkosonomies to leverage knowledge sharing on the social web: a brief survey[C]//The 2008 IEEE/ACM International Conference on Automated Software Engineering. New York: IEEE/ACM, 2008.

[5] ZHANG Z K, ZHOU T, ZHANG Y C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graph[J]. Physica A, 2010, 389(1): 179-186.

[6] LINDSEY R, VEKSLER V D, GRINTSVAYG A, et al. Effects of corpus selection on semantic relatedness[C]//The 2007 Soar Technology International Conference of Cognitive Modeling. Ann Arbor: MI, 2007.

[7] HALPIN H, ROBU V, SHEPHERD H. The complex dynamics of collaborative tagging[C]//The 2007 ACM International Conference on World Wide Web. New York: ACM, 2007: 211-220.

[8] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用[M]. 北京: 清华大学出版社, 2006.
WANG Xiao-fan, LI Xiang, CHEN Guan-rong. Complex network theory and application[M]. Beijing: Tsinghua University Press, 2006.

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

[10] 吴超, 周波. 基于复杂网络的社会化标签分析[J]. 浙江大学学报(工学版), 2010, 44(11): 2194-2197.
WU Chao, ZHOU Bo. Complex network analysis of tag as a social network[J]. Journal of Zhejiang University (Engineering Science), 2010, 44(11): 2194-2197.

[11] LV LY, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A, 2011, 390(1): 1150-1170.

[12] 何华灿, 王华, 刘永怀, 等. 范逻辑学原理[M]. 北京: 科学出版社, 2001.
HE Hua-can, WANG Hua, LIU Yong-huai, et al. Universal logic principle[M]. Beijing: Science Press, 2001.

[13] KOHAVI R. A study of cross-valisation and bootstrap for accuracy estimation and model selection[C]//Proceedings of the International Joint Conference on Artificial Intelligence. Quebec, Canada: Morgan Kaufmann Publisher, 1995: 1137-1143.

[14] HANELY J A, MCNEIL B J. The meaning and use of the area under a receiver operating characteristic (ROC) curve [J]. Radiology, 1982, 143(1): 29-36.

[15] HERLOCKER J L, KONSTANN J A, TERVEEN K, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transaction on Information Systems, 2004, 22(1): 5-53.

[16] GRANOVETTER M S. The strength of weak ties[J]. American Journal of Sociology, 1973, 78(6): 1360-1380.

编 辑 蒋 晓