基于兴趣度变化的社区网站用户性格相似度计算①

张晓滨, 庞海燕

(西安工程大学 计算机科学学院, 西安 710048)

摘 要: 针对社区网站中通过衡量用户静态信息的一致性和共同好友数量, 忽视其动态信息以及动态信息变化过程实现好友推荐这一问题, 提出基于兴趣集、兴趣度持续时间、兴趣集序列构造性格模型, 比较用户性格相似度实现性格相似的判断. 实验结果显示, 该模型实现好友推荐的效果良好.

关键字: 性格相似度; 兴趣度; 兴趣集; 兴趣度持续时间

Character Similarity Calculation of Community Site Based on User's Interest Change

ZHANG Xiao-Bin, PANG Hai-Yan

(School of Computer Sicense, Xi'an Polytechnic University, Xi'an 710048, China)

Abstract: In view of SNS, the way to achieve friends recommended by measures the consistency of user's static information or the number of common friends. But it ignores dynamic information and the change process of it. For this problem, this paper based on interest sets, the duration of interest and the interest sequence to structure a character model. Compares the similarity to achieve character similarity judgment. The results show that it is a good model to get friends.

Key words: character similarity; interest; interest series; the duration of interest

传统的社区网站通过用户静态信息的一致性和共同好友的数目来实现好友推荐, 但这种推荐方法[1-3]仅仅把用户的静态信息作为相似性的唯一标准, 忽视了用户的动态信息. 该方法虽然能够为用户寻找静态信息一致和有共同好友的好友, 但是静态信息一致和共同好友达到一定值不是好友的充分条件; 而且该方法局限性于静态信息和共同好友, 不能找到真正志同道合的人. 基于此, 本文将兴趣爱好作为性格的特征项[4], 通过计算兴趣爱好及其变化过程的相似度来实现好友推荐.

用户发表和分享的日志作为数据来源, 通过中文分词和同义词近义词处理过程[5-8], 得到兴趣度[9-10]及其对应的频数二元组. 基于该二元组, 通过兴趣集、兴趣度持续时间、兴趣集序列构造性格模型. 文献[11]描述了时间序列的相似度检测模型, 用兴趣度曲线的倾斜数组表示兴趣度序列. 但是忽略了曲线长度对曲线形状的影响. 本文在此基础上, 引入曲线在时间轴上的投影, 即时间间隔比较函数,最终以斜率比较函数和时间间隔函数的乘积作为序列相似度[12]的判断依据.

1 算法描述

性格相似度用来衡量用户的性格相似程度, 在社区网站的实际运用中, 设定一定的性格相似度判断阈值. 比较其计算结果与阈值大小, 判断用户是否可以作为好友推荐给用户. 社区网站用户集合为W, 用户P和Q是W的元素, 且P不是Q, 即$P \in W$, $Q \in W$, 且$P \neq Q$. 其算法描述如下:

步骤一 通过兴趣度及其对应的频数二元组等日志预处理结果, 得到兴趣集 E、兴趣度持续时间 F、兴趣集序列 G.

步骤二 计算用户 P 和Q的兴趣集相似$Sim_E(P,Q)$、兴趣度持续时间的相似度$Sim_F(P,Q)$、兴趣集序列的

① 基金项目:教育部春晖计划(Z2009-1-71001)

收稿时间:2012-07-23;收到修改稿时间:2012-08-27

相似度 $Sim_G(P,Q)$.

步骤三 根据各特征项相似度 $Sim_{I_i}(P,Q)$ 与该相似度在性格相似度中所占比例 $P(I_i)$ 求出性格相似度 $Sim(P,Q)$.

步骤四 比较性格相似度判断阈值 σ 与性格相似度 $Sim(P,Q)$ 的大小. 若 $Sim(P,Q)\geq\sigma$,则满足好友推荐的条件, 成为一个好友对. 否则, 不满足好友推荐条件.

步骤五 当对社区网站中的用户 W 进行遍历, 执行上述过程. 即当用户 M,N 满足 $M\in W$, $N\in W$, 且 $M\bigcup N\neq P\bigcup Q$ 时, 重复步骤一到步骤四. 否则, 计算结束, 输出好友对.

2 性格相似度的计算框架

2.1 兴趣集相似度 $Sim_E(P,Q)$

兴趣元素集合记为 $H=\{h_1,h_2,h_3,\cdots,h_n\}$, 频数集合记为 $F=\{f_1,f_2,f_3,\cdots,f_n\}$. 某一时刻, 对应的兴趣元素与对应的频数组成的二元构成为这一时刻的兴趣集 E. 则

$$E=\{\ <e_1=(h_1,f_1),e_2=(h_2,f_2),\cdots e_n=(h_n,f_n)>\}$$

用户 P 和 Q 兴趣度的交集记为 H_{com}, $H_{com}=H(P)\bigcap H(Q)=\{h_{com1},h_{com2},\cdots,h_{comn}\}$,其对应的兴趣集二元组为:

$$E_{com}=\{<e_{com1}=(h_{com1},f_{com1}),e_{com2}=(h_{com2},f_{com2}),\cdots,e_{comn}=(h_{comn},f_{comn})>\}$$

用户兴趣集的相似度表征二者兴趣度交集频数的相似程度. 当 H_{com} 不为空时, 二者有共同的兴趣度, 但是对于不同用户, 其共同兴趣度的频数存在差异. 此时, 通过各元素的频数相似度与该元素在兴趣集交集中所占比例求和得出. 当 H_{com} 为空时, 即二者无共同的兴趣度, 兴趣集的相似度为 0. 公式表达如下:

$$Sim_E(P,Q)=\begin{cases}\sum_{i=1}^{n}\dfrac{\min(f_{comi}(P),f_{comi}(Q))}{\max(f_{comi}(P),f_{comi}(Q))}\bullet\overline{P_{comi}} & ,\ H_{com}\neq\phi\\ 0 & ,\ H_{com}=\phi\end{cases}\quad(1)$$

式中, $\min(f_{comi}(P),f_{comi}(Q))$ 表示 $E_{com}(P)$ 和 $E_{com}(Q)$ 中第 i 个兴趣度的频数最小值, $\max(f_{comi}(P),f_{comi}(Q))$ 表示 $E_{com}(P)$ 和 $E_{com}(Q)$ 中第 i 个兴趣度的频数最大值, $\overline{P_{comi}}$ 表示 E_{com} 中第 i 个元素在该兴趣集中所占的比重,

$$\overline{P_{comi}}=\frac{P_{comi}(P)+P_{comi}(Q)}{2}\quad(2)$$

$P_{comi}(P)$, $P_{comi}(Q)$ 分别表示用户 P 和 Q 第 i 个共同兴趣度在 $E_{com}(P)$ 和 $E_{com}(Q)$ 中所占的比重, 即

$$P_{comi}(P)=f_{comi}(P)\Big/\sum_{i=1}^{n}f_{comi}(P),\quad P_{comi}(Q)=f_{comi}(Q)\Big/\sum_{i=1}^{n}f_{comi}(Q)$$

2.2 兴趣度持续时间相似度

T_1 到 T_N 时刻用户兴趣度集合为:

$$H_{sum}(T_1\sim T_N)=H(T_1)\bigcap H(T_2)\bigcap\ \cdots\bigcap H(T_N)=\{h_{sum1},h_{sum2},h_{sum3,\cdots,}h_{sumn}\}$$

兴趣度持续时间的相似度用来表征一段时间内兴趣度在不同时刻存在性的相似程度. 兴趣及持续时间用兴趣集集合中各元素持续时间和的平均值来表示, 但是考虑到时间粒度对时间结果的影响, 因此引入时间粒度 Δt, 某一元素的持续时间就表示为 Δt 与该时间粒度下持续时间的乘积. 即为:

$$T(T_1\sim T_N)=\sum_{i=1}^{n}\frac{1}{n}t(h_{sumi})\bullet\Delta t\quad(3)$$

其中, $t(h_{sumi})$ 表示兴趣度 i 持续的时间, Δt 表示 T_N 与 T_{N+1} 的时间间隔. $t_j(h_{sumi})$ 表示兴趣度 h_{sumi} 是否存在与连续时刻, 若 h_{sumi} 存在于连续的时刻中, 则该兴趣度的持续, 记为 1, 否则, 说明该兴趣度不持续, 记为 0.

$$t_j(h_{sumi})\begin{cases}1 & ,\ h_{sumi}\in H(T_N)\bigcap H(T_{N+1})\\ 0 & ,\ h_{sumi}\notin H(T_N)\bigcap H(T_{N+1})\end{cases}\quad(4)$$

由公式(3)和(4), 可得出, 兴趣度持续时间为:

$$T(T_1\sim T_N)=\sum_{i=1}^{n}\frac{1}{n}\sum_{j=1}^{N-1}t_j(h_{sumi})\bullet\Delta t\quad(5)$$

根据兴趣度持续时间的相似度定义, 当 $\max(\{T_P(T_1\sim T_N),T_Q(T_1\sim T_N)\neq 0$ 时, 二者的持续时间有交集, 但是对于不同用户, 其持续时间有差异. 此时, 通过用持续时间的最小值与最大值的比值来表征其相似度. 当 $T_P(T_1\sim T_N)=T_Q(T_1\sim T_N)=0$ 时, 即二者的兴趣度都不持续, 此时兴趣度持续时间为 0. 公式表达为:

$$Sim_F(P,Q)=\begin{cases}\dfrac{\min\{T_P(T_1\sim T_N),T_Q(T_1\sim T_N)\}}{\max(\{T_P(T_1\sim T_N),T_Q(T_1\sim T_N)\}} & ,\max(\{T_P(T_1\sim T_N),T_Q(T_1\sim T_N)\}\neq 0\\ 0 & ,T_P(T_1\sim T_N)=T_Q(T_1\sim T_N)=0\end{cases}\quad(6)$$

2.3 兴趣集序列相似度 $Sim_G(P,Q)$

兴趣集序列的相似度表征一段时间内用户兴趣集变化趋势的相似程度, 考虑到兴趣集序列是兴趣度序列的集合, 先计算兴趣度序列相似度, 然后计算兴趣集的相似度.

2.3.1 兴趣度序列相似度 $Sim_{g_i}(P,Q)$

兴趣度 g_i 的时间序列记为 $S=\{<x_1=(f_1,T_1),x_2=(f_2,T_2),\cdots,x_n=(f_n,T_N)\}>$, 其中, f_i 是 T_i 时

刻该兴趣度的频数. 将不同时刻所对应的频数用平滑的曲线连接起来, 则该曲线的意义是用户在此时间段内对此兴趣度感兴趣程序的变化趋势.

考虑到曲线可能存在拉伸, 压缩, 噪音干扰, 本文通过比较序列的形状来实现兴趣度序列相似度的比较. 首先选取曲线上的特征点, 描述如下:

在 $S=\{<x_1=(f_1,T_1),x_2=(f_2,T_2),\cdots,x_n=(f_n,T_N)\}>$ 中, 当 x_m 满足: 存在常量 R, i 和 j 且 $1\le i<m<j\le n$, 使得: 1) f_m 是 $f_1,\cdots,f_j$ 中的最大值; 2) $f_m/f_i\ge R$ 且 $f_m/f_j\ge R$ 成立. 则称 $x_m\ (1<m<n)$ 是一个极大特征点. 同理, 当 x_m 满足: 1) f_m 是 $f_1,\cdots,f_j$ 中的最小值; 2) $f_i/f_m\ge R$ 且 $f_j/f_m\ge R$ 成立. 则称 $x_m\ (1<m<n)$ 是一个极小特征点.

用线段连接相邻特征点, 通过斜率比较函数与时间轴长度比较函数的乘积来简化曲线相似度的计算.

线段斜率 $\rho_i=(f_{i+1}-f_i)/(T_{i+1}-T_i)$, 时间轴上的长度用 l_i 表示, 序列表示为 $(<\rho_1,l_1>,\cdots,<\rho_m,l_m>)$. 假设两用户的序列 S_1 和 S_2 分段后的线段的斜率存放于数组 $U_1(1,\cdots,n)$ 和 $U_2(1,\cdots,n)$ 中, $s(i)$ 表示斜率比较函数,

若 $U_1(i)=0$;

$$s(i)=\begin{cases}1 & ,\ U_2[i]=U_1[i]\\ 0 & ,\ U_2[i]\ne U_1[i]\end{cases} \tag{7-1}$$

若 $U_1(i)\ne 0$,

$$s(i)=\begin{cases}1 & ,\ U_2[i]/U_1[i]\ge 0\\ 0 & ,\ U_2[i]/U_1[i]<0\end{cases} \tag{7-2}$$

时间轴上的长度存放于数组 $V_1(1,\cdots,n)$ 和 $V_2(1,\cdots,n)$ 中, $w(i)$ 表示时间轴长度比较函数.

$$w(i)=\frac{\min\{V_1[i],V_2[i]\}}{\max\{V_1[i],V_2[i]\}} \tag{8}$$

由以上的过程得出, 序列 s_1 和 s_2 的相似度函数的时间复杂度为 O(N^2).

$$Sim_{g_i}(S_1,S_2)=\sum_{i=1}^{n}\frac{1}{n}s(i)\bullet w(i) \tag{9}$$

设定一定的容忍限度 ε, 其中, $0<\varepsilon\le 1$. 当 $Sim_{F_i}(S_1,S_2)>\varepsilon$, 表示序列 s_1 和 s_2 相似. 用 1 来表示相似, 0 表示不相似, 其该兴趣度的相似性可表示为:

$$Sim_{g_i}(P,Q)=\begin{cases}1 & ,\ Sim_{g_i}(S_1,S_2)>\varepsilon\\ 0 & ,\ Sim_{g_i}(S_1,S_2)\le\varepsilon\end{cases} \tag{10}$$

2.3.2 兴趣集序列相似度 $Sim_G(P,Q)$

兴趣集序列的相似度表征一段时间内用户兴趣集变化趋势的相似程度, 前面已经得到了兴趣度序列的相似度计算方法, 那么兴趣集序列的相似度由兴趣度序列的相似度的平均值得出. 公式表达如下:

$$Sim_G(P,Q)=\sum_{i=1}^{n}\frac{1}{n}\bullet Sim_{g_i}(P,Q) \tag{11}$$

其中 $Sim_{g_i}(P,Q)$ 表示兴趣集中第 i 个元素, 即兴趣度 g_i, 在该时间段内的序列相似度, n 表示兴趣集所包含的兴趣度元素个数.

2.4 性格相似度 $Sim_{I_i}(P,Q)$

性格特征项为上述的兴趣度、兴趣度持续时间、兴趣集序列, 即 $I=\{E,F,G\}$. 性格相似度 $Sim_{I_i}(P,Q)$ 用来衡量用户 P 和 Q 的性格相似程度, 通过对各特征项相似度与该相似度在性格相似度中所占比例乘积求和得出,公式表达如下:

$$Sim(P,Q)=\sum_{i=1}^{3}P(I_i)\ Sim_{I_i}(P,Q) \tag{12}$$

其中, $Sim_{I_i}(P,Q)$ 表示用户 P 和 Q 特征项 I_i 的相似度, $P(I_i)$ 表示该特征项相似度在性格相似度中所占的比例.

3 实验及结果分析

选取社区网站中的 200 个用户在 3 个月内的发表和分享的日志, 经过中文分词和同义词近义词处理后得到兴趣元素及其频数二元组. 选择兴趣度频数 $f\ge 5$ 的兴趣元素, 组成该实验的实验数据. 假设 $P(I_E),P(I_F),P(I_G)$ 分别为 30%, 30%, 40%. 应用以上性格相似度计算模型和对应的参数, 得出用户的相似度 δ, 当 $\delta\ge\sigma$ 时, 满足好友推荐的条件, 成为一个好友对.

上图中, 横坐标表示参数,,, 的取值, 纵坐标表示好友对的数目. 图中的点所对应的参数分别为: (1)1,0.25,10,0.7; (2) 2,0.25,10,0.7; (3)1,0.3,10,0.7; (4)1,0.25,15,0.7; (5)1,0.25,10,0.8. 图 1 展示了不同参数下, 所对应的好友数量. 图 2 展示了对应参数下, 实验结果与实际结果中, 好友名单的一致性. 通过查询社区网站数据库, 好友的数目为 65. 实验显示, 当参数为 2/3,2/3, 10, 0.75 时, 对应的好友数目为 64, 该结果最接近于真实数据, 此时好友名单的一致性为 90%. 而此时分析用户注册时所提供的静态信息, 其一致性为 40%, 共同好友数量的为 5.

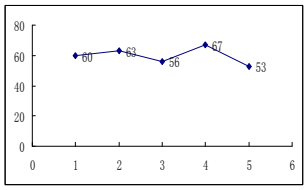

图 1 好友数量

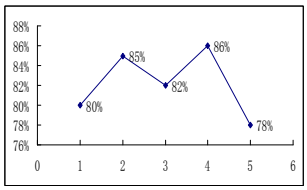

图 2 好友一致性

那么采用用户的静态信息一致性和共同好友数目的方法, 实验结果如何呢? 依据静态信息和共同好友数目构造相似度计算模型

, 该模型所涉及的静态信息为血型, 星座, 颜色喜好, 学历, 所在地, 所在院校, 根据这些参数以及共同好友数目的取值范围, 绘制不同参数情况下所对应的好友数目与好友一致性的图, 如图 3、图 4.

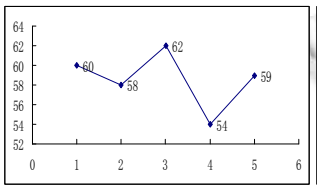

图 3 好友数量

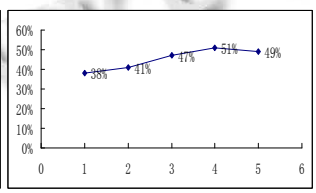

图 4 好友一致性

当实验结果接近于 65 时, 比较实验结果与上述的实际好友对, 其好友一致性为 53%.

由以上的实验结果和分析可知, 基于兴趣度变化的性格相似度模型能够更好的实现好友推荐, 为社区网站用户提供个性化服务. 同时, 该模型的参数选择对于实验结果极其重要. 当设置合适的参数时, 该模型能准确的为社区网站用户推荐好友.

4 结语

本文基于兴趣度变化建立性格模型, 然后通过用户性格的相似度实现用户的个性化服务, 其优点是基于动态信息, 即兴趣度变化建立性格模型, 改进了传统模型仅基于静态信息的不足. 今后还需要关注更多性格参数的选取, 以更精确的描述性格.

参考文献

1 陈晓旺.基于 SNS、电子商务、WiKi 结合的区域性社区网站的研究与设计.桂林理工大学,2010.3–5.

2 陈清华,李林锦,翁正秋.SNS 网站用户关系挖掘的设计与实现.计算机工程,2011,2(37):61–64.

3 童海妙.一种基于协同标签系统与用户建模的个性化好友推荐方法.杭州:浙江大学,2011.2–12.

4 钟梅.基于性格倾向性的雇主品牌类型偏好的实证研究.西南财经大学,2008.1–2.

5 洪晓,康松林,朱小娟,谢文彪.基于词频统计的中文分词的研究.软件学报,2005(7):67–68.

6 Sun TL, Liu YL, Yang LH, Li ZY, Liu ZH. An ambiguity discovery algorithm on Chinese word segmentation based dictionary. Proceedings of the 2009 2nd. Pacific-Asia Conference on Web Mining and Web-Based Application. 2009:39–42.

7 周法国,杨炳儒.句子相似度新方法及在问答系统中的应用.计算机工程与应用,2008,(11):165–167.

8 Rybinski H. Discovering synonyms based on frequent termsets. Lecture Notes in Computer Science.2007(4585: 516–525.

9 王微微,夏秀峰,李晓明.一种基于用户行为的兴趣度模型.计算机工程与应用,2012,(48):28–29.

10 Mohd LN, Shaharanee F, Dillon T. Interestingness. Measures for association rules based on statistical validity. Knowledge-Based Systems. 3, April 2011:386–392.

11 汤俊,熊前兴.基于时间序列相似度的离群模式检测模型.武汉大学学报,2006,39(3):112–113.

12 Sangwook. Efficient processing of similarity search under time warping in sequence databases: An index-based approach. Information Systems,2004,6(29):405–420.