

基于 Tag 的知识主题网络构建与 Web 知识推送研究*

易 明 曹高辉 毛 进 邓卫华

摘 要 采用移植借鉴的研究方法,从认知心理学和认知语言学的视角论证了具有类似 Tag 使用模式的用户在兴趣、思考与表达方式上的相似性,以此为基础提出了基于 Tag 的知识主题网络的构建方法,并建立了基于知识主题网络的 Web 知识推送模型。实验结果表明了本文理论与方法的科学性。然而,由于受到 Tag 使用量的分布规律呈现出明显的“幂律分布”特征的负面影响,Web 知识推送的覆盖度相对较低,需要在今后的研究中进一步解决。图 7。参考文献 17。

关键词 知识主题网络 Web 知识推送 社会化标签 Tag

分类号 G250.73

ABSTRACT By using the research methods of transplant and reference, this article expounds the similarities of interests and ways of thinking and speaking among users who have similar usage of tags from the perspective of cognitive psychology and cognitive linguistics; it describes then the method of knowledge topic network construction and discusses finally the Web knowledge push model based on knowledge topic network. The experimental results support the theory and method proposed in this paper. However, the coverage of the Web knowledge push model is relatively narrow due to the influence of the tag usage amount with Power Law degree distribution, which needs further research in the future. 7 figs. 17 refs.

KEY WORDS Knowledge topic network. Web knowledge push. Social tag. Tag.

CLASS NUMBER G250.73

1 问题的提出

知识推送的概念,最早由美国《信息周刊》主编 Ricadela 在介绍微软公司的知识推送时提出^[1],旨在依据用户的知识需求主动将合适的知识提供给用户。在 Web2.0 环境下,互联网已成为全球最大的知识库,它在给人类的生活和工作带来革命性变化的同时,也引发了“知识泛滥”、“知识迷航”等问题。由此,Web2.0 环境下的知识推送问题成为学者们关注的热点之一。同时,Web2.0 时代的互联网已经成为一个非线性、自组织的社会化网络,其最大特色就是充分

利用集体智慧,当更多的用户加入到这个集体时,个体属性的重要性将逐渐淡化,个体间的互动行为和个体间的关系将发挥更重要的作用^[2]。有鉴于此,部分学者将社会网络引入 Web 知识推送中,形成了两种不同的研究思路。首先,依据社会网络的结构特征进行 Web 知识推送。比如,Palau 等人将推送系统中的用户关系表示为社会网络的形式,并通过社会网络中的密度、规模、中心性、派系等特征变量为当前节点,寻找能够提供知识推送服务的节点^[3];丁连红利用员工的电子邮件、聊天记录、手机短信

* 本文系国家社会科学基金项目“潜在社会网络发现与互联网知识传播研究”(批准号:10CTQ025)、中国博士后科学基金面上资助项目“基于 SNA 的标签网络分析与个性化信息推荐研究”(批准号:20100471156)、教育部人文社会科学青年基金项目“虚拟社区中基于 Tag 的知识协同机理与应用研究”(批准号:09YJC870008)的研究成果之一。

等信息流建立社会网络,然后应用扩展的 GN 算法、加权 PageRank 算法从社会网络中寻找特定节点为当前员工推送科技文档^[4];Wang Ziqi 等人综合利用用户的标注行为、评价数据和关系的相似性建立社会网络,然后应用 RWR(Random Walk with Restarts)算法进行个性化推送^[5]。其次,依托社会网络借鉴协作过滤思想进行 Web 知识推送。比如,陈君针对 Blog 发布的信息,综合考虑用户节点的点出度、点入度,提出了 Web 社会网络扩展算法、相邻用户生成算法,并建立了协作过滤模型^[6];Carmagnola 等人借鉴协作过滤思想,将关系强度作为综合判断 Web 资源推送权重的主要依据,提出了 SoNARS 推送算法^[7];Esslimani 等人依据用户浏览行为的相似性建立社会网络模型,并利用关系传导性来发现用户之间新的关联,以此为基础进行协作过滤推送^[8]。

不难发现,学者们在这一领域的研究还处于起步阶段,缺乏具有高度和深度的研究成果,而且在研究思路上也存在一些问题。第一,Web 知识推送需要依据用户的特定知识需求,在一般的社会网络中,用户间“关系的互动”并不意味着在特定知识情境下必然能产生“知识的互动”。无论是依据社会网络的结构特征进行 Web 知识推送,还是依托社会网络借鉴协作过滤思想进行 Web 知识推送,都会面临这一问题的严峻挑战。第二,现有研究都局限在用户已经形成的社会网络中寻找节点进行知识推送。然而,Web2.0 网站链接了无数用户节点,可能存在很多有价值的知识节点,用户还没有与之建立联系。如何发现这些潜在知识节点,并利用它们提供知识推送服务,也是需要研究的重点。针对这些问题,本文将利用 Tag(社会化标签)来展开探索。从社会学来看,Tag 可以帮助用户寻找、建立知识互动型人际关系,因为类似的 Tag 使用模式暗示了这些用户拥有相同的兴趣、相似的思考与表达方式。即使这些用户彼此并不认识,但如果利用某种机制(如知识主题)将他们聚合在一起,彼此之间产生知识互动

的概率会相对较高,有利于 Web 知识推送。

2 用户使用 Tag 的认知科学视角分析

1975 年,由于美国著名的斯隆基金的投入,美国学者将哲学、心理学、语言学、人类学、计算机科学和神经科学整合在一起,研究“在认识过程中信息是如何传递的”,这个研究计划的结果产生了认知科学^[9]。一般认为,认知科学是研究人类认知和智力的本质与规律的科学^[10],研究的发展形成了心智哲学、认知心理学、认知语言学、认知人类学、人工智能、认知神经科学等学科分支。显然,Tag 体现的是用户的思维认知,它是用户自由发挥主体能动性的产物。本文将应用认知科学中认知心理学和认知语言学的相关理论对用户使用 Tag 的过程进行剖析,论证具有类似 Tag 使用模式的用户在兴趣、思考与表达方式上的相似性。

2.1 认知心理学视角

认知心理学认为人是一种信息加工系统,人的心理活动是一种主动寻找信息、接受信息,进行信息编码,在一定的信息结构中进行加工的过程,而且在此过程中特别强调认知中的结构优势效应,即原有的认知结构对当前认知活动的影响^[11]。显然,用户使用 Tag 的行为过程可以从认知心理学的角度进行分析。由于用户使用 Tag 的信息动机不同,最后所导致的信息加工模式也不一样。一般来说,用户使用 Tag 的信息动机主要有记录关注点、知识组织、知识共享三种。首先,每个用户可能有着自己独特的关注点,Tag 可以帮助用户记录这些关注点;其次,针对所关注的知识,用户也会添加 Tag 作为将来检索利用的入口;最后,用户希望他人能够分享自己的知识,由此会添加相关的 Tag,从而为他人检索利用该知识提供入口。显然,受到不同信息动机的驱使,用户可能会选择不同的 Tag,即如果用户的信息动机是记录关注点,那么关

注点就可以看作是用户即将改变自身认知结构的突破点;如果信息动机是知识组织,那么用户主要是依据自己的知识结构或者是记忆特征来选择 Tag;如果信息动机是知识共享,那么用户则是遵循逆向思维,依据他对相关用户的认知结构特征的理解来选择 Tag。无论是哪种信息动机,Tag 最终都体现了用户的认知结构特征,具有类似 Tag 使用模式的用户就很有可能拥有相同的兴趣、相似的思考与表达方式。

2.2 认知语言学视角

人类思维的结晶是语言,语言是人类表达观念和思想的方式之一,也是认知系统的重要组成部分。认知语言学一方面运用认知科学的理论和方法来探讨语言现象,另一方面又通过语言现象来揭示人的认知能力。与其他相关学科不同,认知语言学揭示了人类认识世界的另外一幅路线图“现实—认知—语言”,即:现实是认知和语言的基础,现实决定认知,认知决定语言;认知是现实与语言的中介,认知反映现实,依靠语言凝化;语言是现实与认知的结果,语言反映和影响认知,认知反映和影响现实^[12]。在解释人类的认知机制时,认知语言学采用了与认知模型相似的一些概念(如框架、脚本、方案、典型事件模型等)进行阐述^[13]。认知模型就是包含某一认知事件范围内所有存储在头脑中的有关语境(认知描述)的抽象化或对复现经验的一般化形式。认知语言学认为,当人们面对一个新情景时,几乎总会找到合适的认知模型作为参照;在找不到现有认知模型作为参照的情况下,人们也会尽力在记忆中搜寻相似经验来创造一个认知模型。显然,用户使用 Tag 的过程也是一个“现实—认知—语言”的转换过程,即:用户依据合适的认知模型对知识进行理解,并以 Tag 作为最终语言表达。由此,用户最终选择的 Tag 便暗示了用户的认知模型特征,而具有类似 Tag 使用模式的用户很有可能拥有相同的兴趣、相似的思考与表达方式。

3 基于 Tag 的知识主题网络构建方法

既然具有类似 Tag 使用模式的用户在兴趣、思考与表达方式等方面具有较大的相似性,那么基于 Tag 的用户聚类便能将那些具有相同的兴趣、相似的思考与表达方式的用户聚集,他们之间产生知识互动的概率相对较大,从而可以弥补一般社会网络的不足。如果按照传统的分析思路,基于 Tag 的用户聚类只需要比较用户使用 Tag 的相似性便可以发现那些具有相似 Tag 使用模式的用户群体。然而,由于用户兴趣的多元性,每个兴趣领域对应的 Tag 可能存在较大的差异,如果直接依据用户使用的所有 Tag 进行相似性分析,必然会大幅降低用户聚类的精度。由此,本文首先将单个用户所拥有的知识进行聚类,以区分单个用户的不同兴趣领域,然后以不同兴趣领域对应的标签云为依据对所有用户的知识进行聚类,进而形成不同的知识主题。显然,知识主题不仅仅是知识的聚类,同时也是拥有这些知识的用户的集合。利用知识主题不但可以建立用户之间的知识互动关系,从而形成知识主题网络(见图 1),而且依托知识主题网络提供各种知识服务的效果也必然会得到有效提升。

3.1 基于 Tag 的知识聚类

定义 1:Tag 系统模型 $D = (U, K, T, A)$ 。其中, $U = \{u_1, u_2, \dots, u_k\}$ 为用户集合, $K = \{k_1, k_2, \dots, k_m\}$ 为知识集合, $T = \{t_1, t_2, \dots, t_n\}$ 为 Tag 集合, $A = \{a = (u, k, t) \mid u \in U, k \in K, t \in T\}$ 为用户、知识和 Tag 之间的连接关系。

定义 2:用户 u_i 的 Tag 系统模型 $Du_i = (Ku_i, Tu_i, Au_i)$ 。其中, $Ku_i = \{ku_{i1}, ku_{i2}, \dots, ku_{in}\}$ 为用户 u_i 标注的知识集合, $Tu_i = \{tu_{i1}, tu_{i2}, \dots, tu_{in}\}$ 为用户 u_i 使用的 Tag 集合, $Au_i = \{a = (ku, tu) \mid ku \in Ku_i, tu \in Tu_i\}$ 为用户 u_i 标注的知识和使用的 Tag 之间的连接关系。

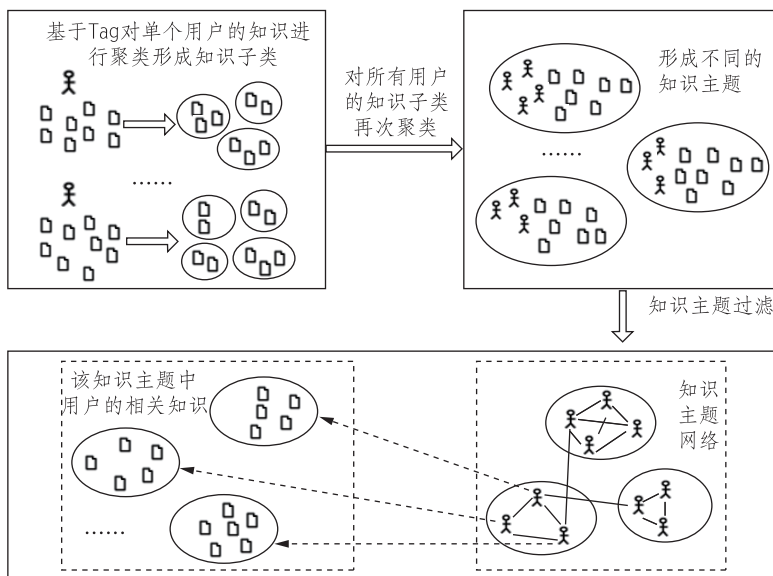


图1 基于 Tag 的知识主题网络构建方法

(1)对单个用户 u_i 标注的知识 $Ku_i = \{ku_{i1}, ku_{i2}, \dots, ku_{in}\}$ 进行聚类。聚类的依据是各个知识对应的 Tag 之间的相似性,聚类的结果就是将单个用户 u_i 的知识划分为 P 个知识子类 $Ku_i = \{cku_{i1}, cku_{i2}, \dots, cku_{ip}\}$,每个知识子类都有对应的标签云。虽然单个用户所使用的 Tag 的数量级别不是很大,每个知识所涉及的 Tag 可能只有几个,如果直接借鉴 TF·IDF 方法进行基于 Tag 的聚类,必然会面临严重的“数据稀疏性”问题。由此,可以考虑利用单个用户使用 Tag 的共现频次建立用户 Tag 网络,然后依据社会网络中的凝聚子群分析对 Tag 进行聚类,进而将每个 Tag 聚类映射到相应的知识集合以完成对单个用户的知识聚类。

(2)对所有用户 $U = \{u_1, u_2, \dots, u_k\}$ 的知识子类进行聚类。聚类的依据是知识子类对应标签云之间的相似性,聚类的结果是形成不同的知识主题。需要说明的是,Tag 使用量的分布规律呈现出明显的“幂律分布”特征,排序在前几位的 Tag 具有较大的使用量,而大量的 Tag 都处于“长尾”区域。由此,在基于 Tag 的聚类数据空间中,数据点的分布可能非常不规则,而且存在着大量的“孤立点”。当前,学者们应用的聚

类方法大多都采用欧氏距离、均值、方差等概念^[14-16],只能发现球状的聚类,不能有效发现任意形状的聚类,而且对“噪声”数据特别敏感。考虑到在聚类方法中,基于密度的聚类可以发现任意形状的子类且能够较好地处理含有“噪声”的数据,所以可以采用基于密度的聚类方法进行探索。

3.2 知识主题过滤

基于 Tag 的知识聚类完成后,Tag 系统中将产生各种知识主题。但是,并不是每个知识主题都有意义,还需要分析知识主题的基本属性(如包含的用户数量、知识数量等)以评价知识主题的相对成熟度^[17],从而过滤掉那些相对成熟度较低的知识主题,进而保证最终构建的知识主题网络能够有效反映用户之间的知识互动关系。

3.3 知识主题网络构建

知识主题网络的构建主要包括知识主题内用户之间关系强度的识别、知识主题间用户之间关系强度的识别。

定义3:知识主题网络 $FG = (FU, FK, FV,$

FA)。其中, FU 为知识主题网络的节点(用户)集合, FK 为知识主题网络中的知识集合, $FV = \{v_{x,y} = (u_x, u_y) | u \in FU\}$ 为知识主题网络中节点(用户) u_x 和节点(用户) u_y 之间的知识互动关系强度, $FA = \{a_{x,i} = (u_x, k_i) | u_x \in FU, k_i \in FK\}$ 为用户 u_x 和知识 k_i 之间的连接关系(如果用户 u_x 标注了知识 k_i , $a_{x,i} = 1$; 否则, $a_{x,i} = 0$)。

(1) 起初, 同一知识主题内用户之间的关系强度可以依据用户所拥有的知识子类之间的相似性来计算; 随后, 如果同一知识主题的用户之间产生了知识互动行为(如知识推送), 那么可以依据知识互动行为的效果来修正彼此的关系强度。

(2) 由于用户兴趣的多元性, 同一用户可能会出现在多个知识主题中, 这就为跨知识主题的用户间知识互动关系的建立提供了线索。例

如, 如果用户 A 同时出现在多个知识主题中, 那么用户 A 就扮演着“知识桥”的角色, 从而将若干知识主题联系在一起。

4 基于知识主题网络的 Web 知识推送模型

与一般的社会网络不同, 知识主题网络中用户之间的关系体现了彼此的知识互动性, 非常利于 Web 知识推送。同时, Web 知识推送本身就是用户之间的知识交互行为, Web 知识推送效果将直接影响彼此的关系强度, 又可以用于知识主题网络优化, 进而实现“知识主题网络 → 知识推送 → 知识反馈 → 知识主题网络优化 → 知识推送……”的良性循环(见图 2)。

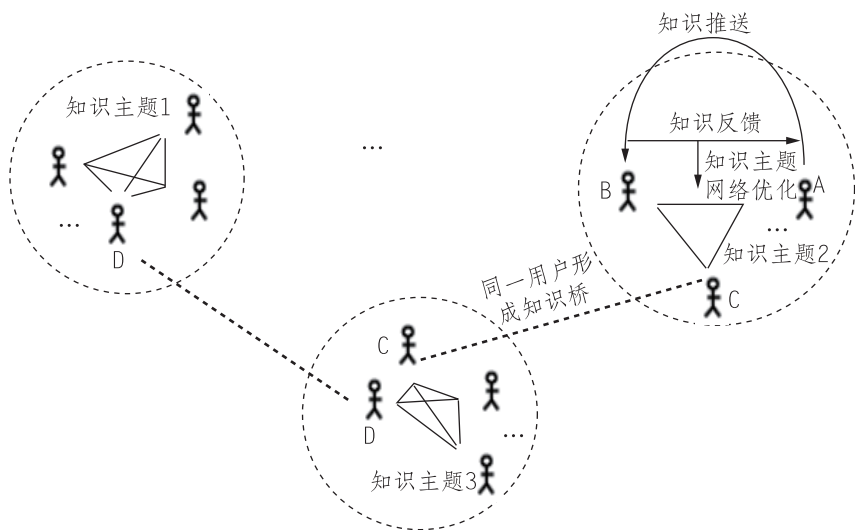


图2 基于知识主题网络的 Web 知识推送模型

4.1 知识主题内基于协作过滤的 Web 知识推送

在每个知识主题内部, 都可以借鉴协作过滤思想实现 Web 知识推送。以图 2 中知识主题 2 为例, 如果用户 A 和用户 B 之间的知识互动关系强度较高, 可以依据用户 B 标注的属于知识主题 2 的知识来预测用户 A 的知识需求, 进而产生针对用户 A 的知识推送对象。

定义 4: $FG = \{Fg_1, Fg_2, \dots, Fg_n\}$ 为知识主

题网络中的知识主题集合, Fg 为知识主题。

定义 5: $Fg = (U_{Fg}, K_{Fg}, V_{Fg}, A_{Fg})$ 。其中, $U_{Fg} = \{u_{Fg1}, u_{Fg2}, \dots, u_{Fgn}\}$ 为知识主题 Fg 中的节点(用户)集合, $K_{Fg} = \{k_{Fg1}, k_{Fg2}, \dots, k_{Fgn}\}$ 为知识主题网络 Fg 中的知识集合, $V_{Fg} = \{v_{i,j} = (u_{Fgi}, u_{Fgj}) | u_{Fgi} \in U_{Fg}\}$ 为节点(用户) u_{Fgi} 和节点(用户) u_{Fgj} 之间的知识互动关系强度, $A_{Fg} = \{a_{i,i} = (u_{Fgi}, k_{Fgi}) | u_{Fgi} \in U_{Fg}, k_{Fgi} \in K_{Fg}\}$ 为知识主题网络 Fg 中用户 u_{Fgi} 和知识 K_{Fgi} 之间的连接关系

(如果用户 $u_{F_{gt}}$ 标注了知识 $K_{F_{gt}}$, 则 $a_{i,t} = 1$; 否则, $a_{i,t} = 0$)。

依据知识主题 F_g 中的节点(用户) $u_{F_{gi}}$, 可以预测节点(用户) $u_{F_{gt}}$ 对每个知识 $k_{F_{gt}} \in K_{F_g}$ 的偏好值 $V(k_{F_{gt}}, u_{F_{gi}})$:

$$V(k_{F_{gt}}, u_{F_{gi}}) = \begin{cases} 0, & \text{if } a_{i,t} = 1 \\ v_{i,j} \times a_{j,t}, & \text{otherwise} \end{cases}$$

同样, 依据知识主题 F_g 中的其他节点(用户), 可以预测节点(用户) $u_{F_{gi}}$ 对每个知识 $k_{F_{gt}} \in K_{F_g}$ 的偏好值, 进而得出面向用户 $u_{F_{gi}}$ 的知识推送对象集 $PK(u_{F_{gi}})$:

$$PK(u_{F_{gi}}) = \{ \langle k_{F_g}, V(k_{F_{gt}}, u_{F_{gi}}) \rangle \mid V(k_{F_{gt}}, u_{F_{gi}}) \geq \phi \}$$

对于依据不同节点(用户)计算得到的同一知识 k_{F_g} 的偏好值, 取最大值; 那些偏好值低于一定阈值 ϕ 的知识将被过滤掉。

4.2 知识主题间基于“知识桥”的 Web 知识推送

显然, 知识主题内基于协作过滤的知识推送主要以知识主题为分析对象, 推送的知识仅限于该知识主题所包含的知识, 所以存在一定的局限性。由此, 可以考虑利用“知识桥”实现跨知识主题的知识推送, 以弥补知识主题内知识推送的不足。扮演着“知识桥”角色的就是那些同时出现在不同知识主题的用户。在图 2 中, 用户 C 同时出现在知识主题 2 和知识主题 3 中, 用户 C 便构建了连接知识主题 2 和知识主题 3 的“知识桥”。由此, 通过比较知识主题 2 中用户 C 和知识主题 3 中用户 D 标注的知识列表, 可以将知识主题 2 中用户 C 标注的知识推送给用户 D, 进而实现知识从知识主题 2 到知识主题 3 的跨界流动。显然, 这种跨知识主题流动的知识进入了知识主题 3 之后, 可以通过基于协作过滤的方法推送给知识主题 3 中的其他用户。

5 实验分析

此次实验利用 Zend Studio 和 Flex Builder 3 程序开发工具, 搭建“网上二手书 C2C 平台”(http://202.114.36.195:8088/CCNU_C2C), 在此基础上进行实验数据的收集和结果分析。在

该平台中, 利用 MySQL5.1 存储“用户—图书—Tag”数据, 用 Php 语言编程实现业务逻辑处理。在实验过程中, 要求用户登录平台标注自己感兴趣的图书。为了避免“数据稀疏性”等问题, 经过数据预处理, 最终采用 35 位用户的标注数据作为实验数据。

5.1 单个用户的图书聚类实验

借助于客户端编程语言 Flex 的 SpringGraph 组件实现用户 Tag 网络, 将 Tag 之间的共现关系以可视化方式呈现, 而 Tag 之间的关系数据则以 XML 为载体与后台程序交互。图 3 描述了针对用户“19”的标注数据所建立的用户 Tag 网络。

针对用户“19”的 Tag 网络, 利用 UCLINET6.0.0.232 工具进行凝聚子群分析。以 n -clique 为计算方法, 同时设定最大距离为 1、子群节点数最小为 6 的严格条件, 结果如图 4 所示。

结合用户“19”的标注数据, 可以得到每个凝聚子群对应的图书集合, 最终 7 个子类对应了 53 本图书(剩下 8 本图书没有归入任何一类)。例如, 第 1 个凝聚子群对应的图书包括:《电子商务理论与实务》、《数字图书馆理论、方法与技术》、《信息检索实验》、《真相幻觉之间: 信息传播中的卖拐喜剧》、《现代信息检索》、《知识工程与知识管理》、《信息管理学基础》、《图书馆学情报学经典著作选读》、《情报研究与创新》等。

5.2 所有用户的图书子类聚类实验

经过数据预处理之后, 针对 35 位用户的 94 个图书子类及其对应的标签云, 在 Java JDK1.6 环境下, 采用程序开发及运行工具——MyEclipse6.5, 计算每个 Tag 在每个图书子类中的权重, 并以 EXCEL 文档形式输出。然后, 采用公开的数据挖掘工作平台 WEKA3.7.0 的 DBSCAN 密度聚类算法进行实验。将参数 Eps 和 Minpts 分别设定为 0.9 和 5, 运行 DBSCAN 算法程序, 结果 49 个图书子类被划分为 8 个知识主题, 剩余 45 个图书子类作为“噪音”数据被过滤。

8 个知识主题共包含了 31 位用户, 剩下 4 位用户由于标注数据相对较少而没有列入 8 个知识主题之中。同时, 考虑到知识主题 6 仅包含

2 位用户,知识主题成熟度相对较低,所以将其过滤。由此,最终建立的知识主题网络如图 5 所

示,大部分知识主题之间都有重叠节点,也就是所谓的“知识桥”。

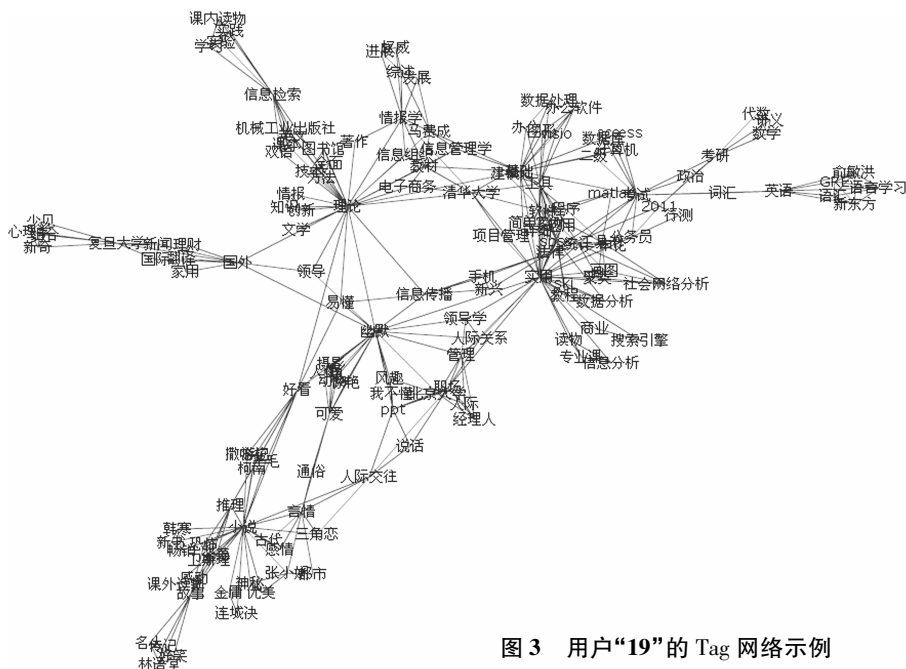


图 3 用户“19”的 Tag 网络示例

N-CLIQUE

```
Max Distance (n-):      1
Minimum Set Size:      6
Input dataset:         u19 (F:\科研项目\社会化标签研究\专业论文\图书标签数据集\u19)
```

WARNING: Valued graph. All values > 0 treated as 1

7 1-cliques found.

```
1: 7 53 60 63 103 139
2: 53 61 112 116 124 130
3: 2 12 47 89 128 136
4: 3 20 91 132 140 141
5: 13 31 39 62 71 79 107
6: 23 36 99 108 131 146
7: 34 57 68 77 83 114
```

图4 用户“19”的 Tag 网络凝聚子群分析

5.3 基于知识主题网络的 Web 知识推送实验

为了评价基于知识主题网络的 Web 知识推送模型的优劣,实验中采用平均绝对误差 MAE 作为度量的标准。MAE 通过计算预测的用户评分与实际的用户评分之间的偏差来度量预测的准确性:

MAE 值越小说明预测的质量越高。在实验过程中,以特定用户为对象,参照前文提出的 Web 知识推送模型,将小于阈值 ϕ 的兴趣度统一置为 0 (不感兴趣),其他兴趣度统一置为 1 (感兴趣)。然后,随机选择用户还未标注的那些兴趣度已置为 1 的

图书,提交给用户作出实际的评价(值为 1 或 0)。由此,实验中 MAE 的理论取值区间将是 $[0,1]$ 。

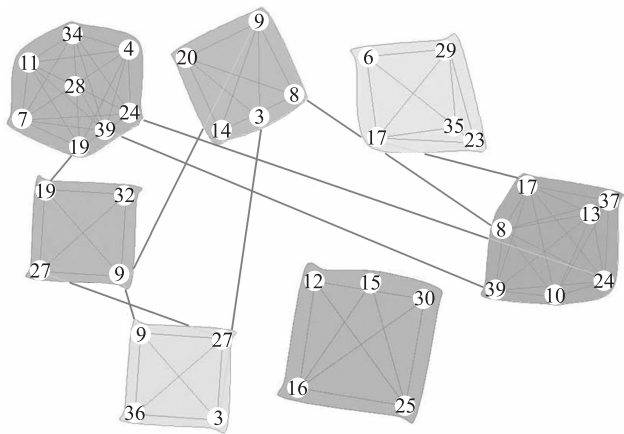


图 5 知识主题网络示意

图 6 描述了针对随机选择的 7 位用户(每个知识主题选 1 位)利用知识主题内基于协作过滤的 Web 知识推送方法的实验结果。其中,横轴为用户 ID 号,纵轴为 MAE 值。总体来看,MAE 值主要分布在 $[0.1, 0.4]$,一方面说明 Web 知识推送的准确度相对较高,同时也印证了本文知识主题网络建立的有效性以及类似的 Tag 使用模式暗示了这些用户拥有相同的兴趣、相似的思考与表达方式的科学性。

在知识主题间基于“知识桥”的 Web 知识推送实验中,分别利用用户“03”、用户“08”、用户“09”、用户“17”、用户“19”、用户“24”、用户“27”、用户“39”作为“知识桥”进行 Web 知识推送。图 7 描述了利用 8 个“知识桥”进行 Web 知识推送对应的 MAE 均值分布。其中,横轴为“知识桥”用户 ID 号,纵轴为 MAE 均值。虽然这种推送方法的准确度比前一种要低,但是绝大部分 MAE 均值还是分布在 0.5 以内,从而表明了利用“知识桥”进行 Web 知识推送的可行性。

6 结语

在一般的社会网络中,用户间“关系的互

动”并不意味着在特定知识情境下必然能产生“知识的互动”,以此为基础的知识推送并不能有效满足用户的知识需求。基于此,本文深度挖掘 Tag 的本质特性,构建知识主题网络,以期能够弥补一般社会网络的不足。实验结果表明,本文建立的知识主题网络具有科学性,以此为基础进行的 Web 知识推送准确度相对较高。然而,由于 Tag 使用量的分布规律呈现出明显的“幂律分布”特征,使得最终构建的知识主题网络无法包含所有的知识,从而会降低 Web 知识推送的覆盖度;而且,游离于知识主题网络之外的孤立用户也无法获得 Web 知识推送服务。这些问题都需要在今后的研究中进一步解决。

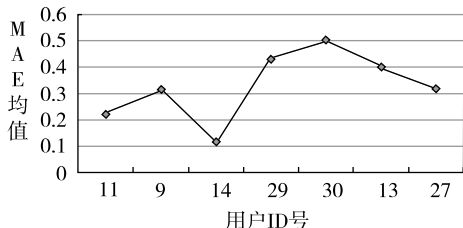


图 6 知识主题内基于协作过滤的 Web 知识推送实验对应的 MAE 均值分布

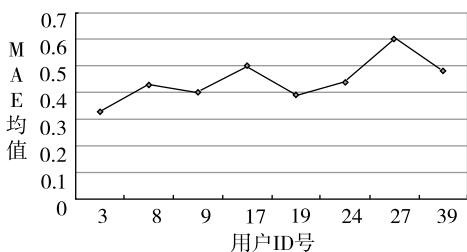


图7 知识主题之间基于“知识桥”的 Web 知识推送实验对应的 MAE 均值分布

参考文献:

- [1] Ricadela A. Microsoft's knowledge push [J]. Information - Week, 2000(805):151.
- [2] 陈定权, 武立斌. 社会网络视角下的信息推荐[J]. 情报杂志, 2007(11):37-40.
- [3] Palau J, Montaner M, López B, et al. Collaboration analysis in recommender systems using social networks [C]. Proceedings of the 8th International Workshop: Cooperative Information Agents VIII. Erfurt, Germany, 2004:137-151.
- [4] 丁连红. 基于信息流的个性化服务研究[D]. 北京: 中国科学院研究生院, 2006.
- [5] Wang Z Q, Tan Y W, Zhang M. Graph-based recommendation on social networks [C]. Proceedings of the 12th Asia-Pacific Web Conference. Busan, Korea, 2010:116-122.
- [6] 陈君. 基于 Web 社会网络的个性化 Web 信息推荐模型研究[J]. 计算机科学, 2006, 33(4):185-187, 193.
- [7] Carmagnola F, Vernero F, Grillo P. SoNARS: A social networks-based algorithm for social recommender systems [C]. Proceedings of UMAP2009, 2009:223-234.
- [8] Esslimani I, Brun A, Boyer A. From social networks to behavioral networks in recommender systems [C]. Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining. Athens, Greece, 2009:143-148.
- [9] 蔡曙山. 认知科学研究与相关学科的发展[J]. 江西社会科学, 2007(4):243-248.
- [10] 姜虹. 认知科学的兴起及其发展路径[J]. 学术交流, 2009(9):28-30.
- [11] 鲁欣, 周伟锋. 基于认知心理互动的网络信息组织的思考[J]. 图书馆学研究, 2008(5):30-33.
- [12] 张俊钰. 认知科学、语言与思维研究的结论与启示[J]. 学理论, 2009(11):174-175.
- [13] 刘辰诞. 结构和边界: 语言表达式的认知基础[D]. 开封: 河南大学, 2006.
- [14] Begelman G, Philipp K, Smadja F. Automated tag clustering: Improving search and exploration in the tag space [C]. Proceedings of the Collaborative Web Tagging Workshop at WWW'06, 2006:217-221.
- [15] 曹高辉, 焦玉英, 成全. 基于凝聚式层次聚类算法的标签聚类研究[J]. 现代图书情报技术, 2008(4):23-28.
- [16] Gemmell J, Shepitsen A, Mobasher B, et al. Personalizing navigation in folksonomies using hierarchical tag clustering [C]. Proceedings of the DaWaK 2008, 2008:196-205.
- [17] Bielenberg K, Zacher M. Groups in social software: Utilizing tagging to integrate individual contexts for social navigation [D]. Dissertation of Master Degree, Universität Bremen, 2005.

易明 华中师范大学信息管理系副教授, 武汉大学信息管理学院博士后。通讯地址: 武汉华中师范大学信息管理系。邮编: 430079。

曹高辉 华中师范大学信息管理系讲师。通讯地址同上。

毛进 华中师范大学信息管理系硕士研究生。通讯地址同上。

邓卫华 华中农业大学经济管理学院信息管理系讲师。通讯地址: 武汉华中农业大学经济管理学院。邮编: 430070。

(收稿日期: 2010-12-06)