虚拟医疗社区中用户相似度计算方法*

方 安 李亚子

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 用户相似度计算在协同过滤系统、用户推荐系统以及社交网络中有着非常重要的作用。在对虚拟医疗社区用户关系进行分析的基础上，提出一种基于医学主题词表的用户相似度计算方法。在虚拟医疗社区的平台上用户可以进行信息咨询以及相互交流，从而有效分配医疗资源。

〔关键词〕 虚拟医院；虚拟医疗社区；用户相似度；医学主题词表

Computing Algorithm of User Similarity in Virtual Medical Communities *FANG An, LI Ya-zi, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, China*

〔**Abstract**〕 User similarity computing plays a very important role in collaborative filtering systems, user recommendation systems as well as social network services. Based on the analysis of user relationships in the virtual medical communities, the paper proposes a user similarity computing algorithm based on Medical Subject Headings (MeSH). On the virtual medical communities' platform, users can get information consult and mutual communication, so as to effectively allocate medical care resources.

〔**Keywords**〕 Virtual hospital; Virtual medical community; User similarity; Medical Subject (MeSH)

1 引言

目前我国医疗资源分布很不合理，高端医疗资源被过度使用，造成效率低下，而低端医疗资源没有得以充分利用，其后果就是资源浪费，并造成“看病难、看病贵”。近几十年来，随着计算机技术与性能的不断提高，它在医学、生物学等各个领域的应用已越来越受到研究者的重视，如何应用计算机技术进行辅助医疗就成为当前的研究热点问题。“虚拟社区”是网络技术发展的结果，在现有的技术条件下，网上的“虚拟社区”[1-2]不断涌现，虚拟医疗社区为合理分配医疗资源创造了有利的条件[3]。对此提出了一种既兼顾医疗规模，又能很好地兼容各级医疗机构的方案：虚拟医疗社区平台。该平台是一个充分信息化的智能系统，既能使医疗机构之间信息充分共享，还能够帮助医生、病人科学决策，也能为社会减少资源消耗、提高资源利用率。本文主要内容围绕虚拟医疗社区中用户相似度的计算展开。

〔收稿日期〕 2011-01-11

〔作者简介〕 方安，硕士，馆员，副主任，发表论文10篇。

〔基金项目〕 国家科技支撑计划课题“公众健康知识整合技术研究与应用”（项目编号：2009BAI76B04）；中国医学科学院医学信息研究所中央级公益性基本科研业务费专项“网络环境下公众健康信息服务模式研究”（项目编号：09R0213）。

2 虚拟医疗社区平台与用户关系计算

2.1 虚拟医疗社区平台架构

虚拟医院是随着计算机网络的发展，在数字化医院的基础上产生的[4]，虚拟医院的实时性、互动

性不强，成员之间的交流很少，而虚拟社区在这方面做的很好。把虚拟社区和虚拟医院结合起来，构造虚拟医疗社区可以很大程度上解决医疗资源的分配问题。目前，国内外对虚拟医疗社区的研究还比较少，成型的产品也很少[5]。在美国医疗信息与管理系统协会（Healthcare Information Management and Systems Society，HIMSS）2008 年会上，IBM 推出了在其 Second Life 在线虚拟世界上发布的虚拟医疗社区（Virtual Healthcare Island）[6]。

虚拟医疗社区平台应该包含的功能：平台门户、信息咨询、信息检索、就诊指南、个性化信息推荐、智能问答与智能服务、成员交流等。虚拟医疗社区具有论坛的所有功能，也具有虚拟医院的所有功能，在社区内用户可以实现所有和医疗相关的信息检索、咨询、留言、问题解答、预约就诊等。

搭建一个虚拟医疗社区平台，就少不了用户的参与和交流，社区成员主要有两种用户：医疗个体（患者、医生）和医疗群体（医疗机构）。成员需要进行注册才能成为正式用户，而在社区进行交流的主要就是医疗个体用户（患者、医生）。用户初始注册时需要选择和填写相关的信息，对患者用户来说需要填写诸如相关病症、反应、用药、时间等信息，医生用户需要填写自己擅长治疗的病种、典型病例用药情况等。用户以后在社区的每次活动都会进行记录，然后计算该用户的个人信息，确定用户相似关系模型。这是虚拟医疗社区和虚拟医院、医疗相关论坛的主要不同之处。

2.2 虚拟医疗社区用户关系

在虚拟医疗社区中，对于每一个具体的用户，他往往只关心与自己的社交圈比较接近的用户并与之交朋友。图 1 中顶点表示用户，边表示两用户之间的好友关系，可以看出该图非常清晰地分成了两个区域，用户 B 一般想与用户 A，C，E 建立起好友关系，而对与用户 G，H 等建立好友关系的兴趣并不大。所以向 B 推荐 G，H 等得到的效果并不好，而向 B 推荐 A，C，E 可能会得到 B 积极的回应。并且由于社交网络中用户数目太过庞大，针对某一个特定的用户虚拟医疗平台中的绝大多数用户都是不会成为好友的，在整个虚拟医疗平台中寻找潜在好友是不必要的。要确定用户之间是否有关系，关键就是看用户之间是否相似。

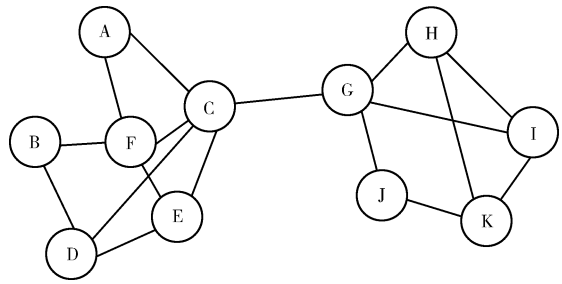

图 1 虚拟医疗社区中的用户关系

2.3 常用社区成员相似度计算方法

相似度已经在很多领域有了应用，如百度/谷歌网站排名、协作过滤等等。百度/谷歌利用相似度得出的值会为网站标识“重要性/等级”，从而提升搜索质量。协作过滤会利用用户对多种事物的评价（大多数是评分）来计算两个用户之间的相似度，然后再推荐用户可能喜欢的事物。推荐好友也需要利用用户之间的相似度产生推荐列表来提高推荐的精度。在同一社区中的用户虽然组成了一个社交圈，但是社交圈中人与人之间也是有亲疏关系的存在。比如即使是在同一个社交圈中，双方的共同好友数目是不一样的，从而造成了双方成为好友的可能性是不同的。在一个涉及用户关系的社区中，其关键问题是如何计算用户之间的相似性。比较常见的计算用户相似度的算法有余弦相似性、皮尔森系数、调整余弦相似性 3 种。

余弦相似性：把用户特征看作是 n 维线性空间上的向量，通过计算两个向量之间的夹角余弦来度量两个用户之间的相似性，其计算公式如下：

$$\cos(\alpha,\beta) = \frac{\alpha \cdot \beta}{|\alpha||\beta|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad (1)$$

皮尔森系数：又称相关相似性，通过皮尔森相关系数来度量两个用户的相似性。计算时首先找到两个用户共同评分过的项目集，然后计算这两个向量的相关系数，其计算公式如下：

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}\sqrt{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}} \quad (2)$$

调整余弦相似性：其本质上是对余弦相似度的一些改进，在此不再赘述。

3 虚拟医疗社区中用户相似度计算方法

3.1 医学主题词表系统

在虚拟医疗社区中采用医学主题词表可以更精确地表示用户的特征，为用户相似度的计算提供了强有力的支撑。医学主题词表借鉴本体理论，通过分析词间关系，抽象出更小粒度的关系基本信息、关系性质、关系的关联。通过组配的方式创建、描述新的词间关系。主题词间的关系主要有：属、分、用、代、参 5 种关系，从结构上来看是一种树型结构。如药品按功能来分，其主题词表结构，见图 2。

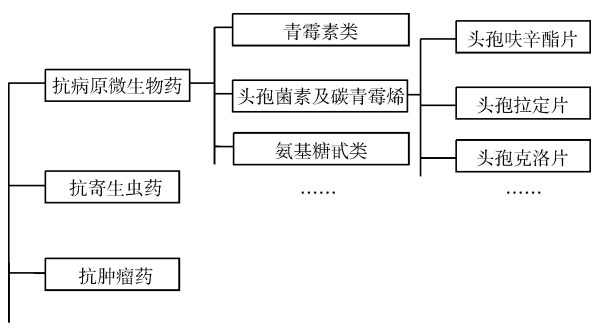

图 2 药品主题词表结构

在图 2 中，头孢呋辛酯片、头孢拉定片和头孢克洛片等都是抗病原微生物药下头孢菌素及碳青霉烯类的药品名称，它们具有相似的功能，虽然药品名称不一样，但它们之间具有一定的关系。同样头孢菌素及碳青霉烯类的药品和青霉素类的药品也有一定的相似关系，如头孢呋辛酯片和阿莫西林片都是抗病原微生物药，都具有消炎的功能，某些病种都可以使用。

还有一种就是药品化学名和商品名虽然看起来不一样，但却具有同一性，相当于同义词，如吗丁啉和多潘立酮片就是同一种药品，二者可以通用。同样也可以按科室与疾病来建立主题词表，见图 3。

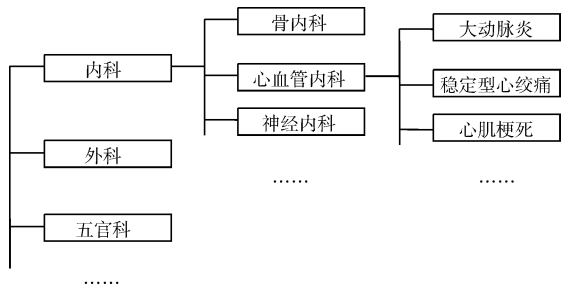

图 3 疾病主题词表结构

3.2 基于医学主题词表的社区成员相似度计算方法

社区中的成员具有和现实社会人员一样的主要特征，病人去哪个科室看哪种病，医生是哪个专科的医生等。在虚拟医疗社区中要为每个用户确定这些信息，就需要对用户进行建模。建模就是根据用户注册时填写的信息，以及其在社区的每次活动的信息记录进行计算。而这些信息最终都是由医学主题词表中的词来构成，每个用户的模型就是一个向量，该向量的每个分量就是主题词表中的词，这样就可以通过计算两个向量的夹角的余弦来判断两个用户的相似度。在数学上采用公式（3）来计算两个向量 α 和 β 夹角的余弦：

$$\cos(\alpha, \beta) = \frac{\alpha \cdot \beta}{|\alpha| \ |\beta|} \quad (3)$$

在计算两个向量夹角余弦的时候，两个向量的长度必须相同，否则无法计算。在给定的主题词表系统的基础上，由于不同主题词的部分相似性，不能照搬上述的计算公式，必须加以改进，给主题词之间进行加权，完全相同权值为 *1*，兄弟关系的词权值为 *0.5*，堂兄弟关系的词权值为 *0.2*，这个权重是根据经验并在使用过程中逐渐调整得到的。也可以根据系统中某些用户的信息，在主观确定用户相似度的基础上通过采用最小二乘法或相关的优化方法来确定主题词之间的权重系数[8]。在计算过程中对内积和向量长度的计算进行了改进，计算方法是设用户 A 的模型为 $\alpha = (a_1, a_2, \cdots, a_m)$，共包含 m 个主题词条，用户 B 的模型为 $\beta = (b_1, b_2, \cdots, b_n)$，共包含 n 个主题词条，则用户 A 和 B 的相似

度可以采用如下公式（4）来计算：

$$Sim(A,B) = \frac{\sum_{k=1}^{m}\left(\sum_{j=1}^{n} W_{a_k} W_{b_j} T_{kj}\right)}{\left(\left(\sum_{k=1}^{m}\left(\sum_{j=1}^{n}\left(W_{a_k} T_{a_{kj}}\right)^2 + W_{a_k}^2\right)\left(\sum_{j=1}^{n}\left(\sum_{k=1}^{m} W_{b_j} T_{b_{jk}}\right)^2 + W_{b_j}^2\right)\right)^{\frac{1}{2}}\right.} \quad (4)$$

其中，W_{a_k}表示 用户 A 的主题词条 a_k 的权值，W_{b_j}表示用户 B 的主题词条 b_j 的权值，T_{kj}表示两个词条 a_k 和 b_j 之间的相似度，$T_{a_{kj}}$表示主题词条 a_k 和 a_j，$T_{b_{jk}}$表示主题词条 b_j 和 b_k 的相似度。

虚拟社区平台主要就是进行用户之间的交流。在进行了用户的相似度计算之后，就对每个用户进行相关信息的设定，当用户再次登录社区后就给该用户推荐和他比较相似的用户信息，这些用户即有作为病人的用户，也有作为专家医生的用户。以病人用户为例，该用户可以向与他相似的病人用户咨询医生情况、医院情况、疾病的治疗情况，交流医生信息、治疗过程中的经验和教训、医院医生的推荐等；也可以与专家用户进行交流，由专家介绍治疗期间应该注意的问题、同行的相关信息，真正做到医疗资源的合理利用。

3.3 用户相似度计算算法描述

输入：要计算相似度的两个用户 A 和 B；输出：用户 A 和用户 B 的相似度。

步骤如下：从社区中获取用户 A 和用户 B 的主题词信息，确定用户 A 和用户 B 的特征个数 m 和 n，确定主题词相似度的权重系数，根据公式（2）计算用户 A 和用户 B 的相似度，输出用户 A 和用户 B 的相似度。

与其他算法相比，该算法中的核心思想仍然是向量空间中计算向量的余弦，在计算相似度时还需要借助一部医学领域的主题词典。该算法具有以下特点：简单，所利用的信息仍是基于词的表层信息，只是这里的词是医学主题词。精确度更高，比其他算法更准确，计算的时候利用了主题词之间的相似性，而其他相似度算法一般是基于通用词汇的，但词汇之间的相似度很难确定。用户相似性质满足弱等价性质：自反性：用户 A 和用户 A 相似；对称性：若用户 A 和用户 B 相似，则用户 B 和用户 A 也相似；弱传递性：用户 A 和用户 B 相似，用户 B 和用户 C 相似，则用户 A 和用户 C 弱相似。

4 结论

本文在对虚拟医疗社区用户关系进行分析的基础上，提出了一种基于医学主题词表的用户相似度计算方法。在虚拟医疗社区的平台上，用户可以进行信息咨询以及相互交流，介绍与总结经验和教训。文中仅仅介绍了虚拟医疗社区中用户相似度的计算，今后将更关注于虚拟医疗社区的发展与具体建设以及用户之间的交流方式与效果的评价等。

参考文献

1 Howard Rheingold. The Virtual Community：finding connection in a computerized world〔M〕. Boston：Addison－Wesley Longman Publishing Co., Inc., 1993.

2 Howard Rheingold. The Virtual Community Homesteading on the Electronic Frontier（revised edition）〔M〕. Massachusetts：the MIT Press, 2000.

3 郭茂灿. 虚拟社区中的规则及其服从〔J〕. 社会学研究，2004，（2）：103－111.

4 张大平. 数字化医院与虚拟医院〔J〕. 中国医院管理，2003，23（7）：39－41.

5 黎亮，张君雁. 区域虚拟医疗体系的研究〔J〕. 现代医院管理，2009，（6）：9－11.

6 http://www－03.ibm.com/press/us/en/pressrelease/23580.wss〔EB/OL〕.〔2010－10－31〕.

7 张鹏，乔秀全，李晓峰. 基于社区划分和用户相似度的好友推荐〔EB/OL〕.〔2010－11－01〕. http://www.paper.edu.cn.

8 Faguo Zhou, Bingru Yang and Linna Li, etal. A New Method for Chinese Sentence Similarity Computing and Its Weighting Coefficients Determination〔C〕// Wuhan：Proceedings of International Conference on Chinese Computing 2007：143－146.