因特网上舆情传播的预测建模和仿真研究

方 薇[1,3] 何留进[2] 宋良图[1]
(中国科学院合肥智能机械研究所 合肥 230031)[1]
(安徽省淮南市人民政府信息化工作办公室 淮南 232000)[2]
(中国科学技术大学计算机科学与技术学院 合肥 230027)[3]

摘 要 网络舆情一定程度上表达出社会公众意愿,它虽然具有一般社会的舆论共性,但由于因特网的影响范围及传播速度,使其在虚拟社会中具有复杂系统的基本特征,故其传播倾向及发展方向受到重视。首先研究整体情感(正、负面)传播的预测模型及其算法;然后通过仿真找出影响其增长和消亡的规律。建模的出发点是将舆情传播看作一个时间序列的马尔科夫链;再利用哈肯协同理论提供的协同概率作为马尔科夫链的状态一步转移概率,构成一个协同-马尔科夫模型。在仿真实验中改变协同概率的各个参变量,以获取舆情随时间传播的不同演化过程的曲线簇,其结果可为虚拟社会管理提供参考。

关键词 网络舆情传播,马尔科夫模型,协同学,预测建模,传播模型仿真
中图法分类号 TP391.7 **文献标识码** A

Predictive Modeling & Simulation for Propagation of Internet Public Opinion

FANG Wei[1,3] HE Liu-jin[2] SONG Liang-tu[1]
(Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China)[1]
(The HuaiNan Office of Information Application, Huainan 232000, China)[2]
(College of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)[3]

Abstract The Internet Public Opinion(IPO) usually expresses the voice of masses, although it hdue to the common attribute of the society opinion, but due to the large scope of influence and fast propagation on Internet, it has been attracted much attention to its tendency of propagation. The paper firstly studied the computer model and its algorithm to predict the propagation of IPO, i. e. the distributed probability of positive or negative opinion. Then the simulation gave the propagation order for growing or descending. In order to modeling, the propagation of IPO was observed as a random Markov time sequence chain, and its one step transition probability between states in the Markov chain was given by a synergistic probability which developeds Dr. H. Haken in his book of synergistic theory. A Synergistic-Markov model was proposed in the paper. By the simulation, the cluster of curves which expresses different evolving states of IPO propagation with time was depicted when we changed the variable parameters at the formula of synergistic probability. The result could provide reference for the management of virtual society.

Keywords Propagation of internet public opinion, Markov model, Synergistic theory, Predictive modeling, Simulation of propagation modeling

1 引言

20 世纪 90 年代出现的因特网改变了世界。由于网络世界具有广泛性、虚拟性、开放性、平等性和实时性的特点,舆情在网上的传播不再像传统媒体那样容易受时间、空间的限制和控制。通过网络,人们能够迅速地了解相对全局性的舆论,所以传统社会舆论系统中的局部和全局的关系正在发生变化[1]。

在因特网舆情(IPO)研究中,计算机建模方法日益受到重视。如文本倾向性分析方面,通常采用相似性判别[2,3]或隐马尔科夫模型,对采集到的大量数据分别用相似度距离或样本训练构成分类模型来进行计算。在舆情传播方面有基于主题相似性的传播预测模型[4]、基于隐马尔科夫的因特网舆情状态及其发表文章数和增量趋势的建模[5],模型参数均通过机器学习获取。文献[6]提出宏观的舆论模型正逐渐被基于局部个体空间相互作用的微观离散动力学模型所代替,并

到稿日期:2011-03-23 返修日期:2011-05-16 本文受国家 863 计划(2006AA10Z23702),2009 年度安徽省信息产业发展专项基金(财建[2009]722 号)资助。

方 薇(1977—),女,硕士生,助研,主要研究方向为计算机信息处理、数据挖掘,E-mail:wfang@iim.ac.cn;何留进(1962—),男,高级工程师,主要研究方向为信息化管理、建设;宋良图(1963—),男,博士,研究员,主要研究方向为基于 Web 的信息管理、信息获取技术、复杂系统仿真与控制。

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

设计了一个元胞自动机模型。之后，不少学者采用各类元胞自动机模型来研究舆情传播[7-9]。

从复杂系统理论出发，认为元胞自动机的研究如果只根据少数服从多数的原则，将会忽视偏好和环境影响；而机器学习对于时间序列状态迁移概率如何变化和影响的机理考虑不足。因此，需要寻求更好的方法。本文重点研究一个论坛中网民对于某个议题的正面或反面意见人数的演化过程。从而展示其发展的不同阶段和特点。在研究预测模型时，将网络舆情传播看作是一个随机过程。从长期来看，尽管舆情有可能从渐变走向突变，但是从时间序列的每两次前后关系来看，后次舆情受到前次舆情的强烈影响，因而可以采用马尔科夫模型对后次予以预测。但本研究与训练学习方法获得的马尔科夫模型[2-5]参数不同，在研究网络舆情整体正反意见人数概率分布时，充分考虑偏好、环境等参变量对状态预测的影响，采用哈肯[11]协同理论来计算马尔科夫模型中的状态转移概率；然后，基于马尔科夫链获得 IPO 时间序列的状态（正面、负面）的概率分布，形成一个协同-马尔科夫模型，并设计了算法和进行不同参变量影响下的仿真实验。该模型和仿真为 IPO 传播趋势预测带来有意义的启示。

本文第 1 节为引言；第 2 节简要介绍协同理论及其在舆情传播中的应用；第 3 节描述协同-马尔科夫模型和算法；第 4 节给出仿真，对不同参变量的协同影响进行了讨论；最后为结语。

2 舆情传播的协同作用

协同是一个跨学科的研究领域，由德国理论物理学家赫尔曼·哈肯（H·Hacken）[11]于 1969 年创立。协同擅长处理复杂系统，研究系统中各组件、元素间相互影响，并展示出自组织的空间、时间或功能结构。所以，协同能用于解决自组织的动力学问题。在香农信息论中，信息熵 $H=-k\sum p_i\ln p_i$，当 $\sum p_i=1$，即消息集合为一种等概率分布时，信息熵 H 为最大值，信息概率在 0－1 区间分布完全相等，信息失去流动，系统处于平衡态。一个孤立系统达到平衡态时最无序、最混乱，此时其熵最大。于是，孤立系统熵的增加是不可逆的。在复杂系统中，人们关注的重点是非平衡态，因为在平衡态下，系统内部不存在物理量的宏观流动，状态参量也不再随时间变化，处于定态状况。这样的封闭系统内部是无序的。而一个开放的系统，能通过与外部交换信息与能量，从无序走向有序，进入非平衡态，社会之所以能从低级到高级不断进化，其关键是开放。显然，网络舆情是一个开放系统。网络舆情正是不断地从外界获得新的信息，才能成为耗散结构[12]，通过自组织，由不平衡趋向平衡，从低级走向高级，起到了推动社会发展的作用。研究网络舆情传播就是研究舆情从非平衡态如何趋向平衡态，当达到新的平衡时，网民共同认识趋向一个更高的水平。

2.1 定义

定义 1（舆情主体） 参与舆情讨论的网民，设主体总数为 N。

定义 2（舆情客体） 舆情传播中主体讨论的公共事务。

定义 3（舆情本体） 主体表达对公共事务的意愿、情绪和态度，本文将正向本体归为＋1，负向本体为归为－1；N_+，N_- 分别代表本体为＋1 或－1 的主体数。

定义 4（舆情不平衡度 q） 主体中 N_+，N_- 所占比例为：

$$q=(N_+-N_-)/2N,\ q\in[-0.5,+0.5] \tag{1}$$

定义 5（初始自转率 v） 舆情发生的初始阶段，在舆情空间即硬空间（如时间、空间、传播媒体等）和软空间（如文化与道德、现代与传统价值观、民主与法治等）的影响下，主体从当前本体状态（如赞成/反对）向相反本体状态（如反对/赞成）转移的可能性。

$v\in[0,+1]$

当不能确定时，一般设 $v=0.5$。

定义 6（环境适应度 k） 表示主体受舆情多数意见和社会压力的心理倾向程度。当 $k\geqslant 0$ 时，k 越大受环境影响越大，从众心理越严重，当 $k=0$ 时，表示不受环境影响。

$k\in[0,10]$

定义 7（偏好 h） 主体偏向正向本体或反向本体的程度。

$h\in[-1,+1]$

h 值的正负分别表示主体向正向本体或负向本体状态偏好的程度，其绝对值越大，偏好越明显。

定义 8（磁化率 M） 磁化率 M 表示主体的本体状态趋向＋1（正极）或－1（负极）的程度。

$$M=2P_{+1}-1 \tag{2}$$

式中，P_+ 是本体状态为正的主体概率。

2.2 协同转移概率

网络舆情传播遵循复杂系统自组织原理，需要关注它是如何从非平衡态走向平衡态的。依据哈肯提出的序参量作用，舆情传播概率计算公式同文献[11]：

$$p_{(+1->-1)}=v\cdot e^{(-(kq+h))} \tag{3}$$

$$p_{(-1->+1)}=v\cdot e^{(kq+h)} \tag{4}$$

式中，P 定义为舆情本体由正向（＋1）或负向（－1）状态朝负向（－1）或正向（＋1）本体状态的转移概率。v,q,k,h 定义同上。v,k,h 为序参量，序参量的大小体现协同作用的强弱。

3 协同-马尔科夫模型

3.1 舆情传播马尔科夫链

舆情传播不是一个确定性的过程，它具有随机性。该过程的将来值均与所有过去值是相互独立的，因此其只与现在值有关，存在所谓的过程无后效性，即马尔科夫性。

舆情传播时间序列可以用马尔科夫链来表达，如果 $t_1<t_2\cdots<t_n\subset\Gamma$，有

$$\begin{aligned}&p[X(t_n)=X_n\,|\,X(t_{n-1})=X_{n-1},\cdots,X(t_1)=X_1]\\&\quad=p[X(t_n)=X_n\,|\,X(t_{n-1})=X_{n-1}]\end{aligned} \tag{5}$$

将舆情传播主体倾向，即本体看作是一个两状态的马尔科夫链，图 1 用来描述马尔科夫链本体状态转移图。图中箭头表示状态转移的方向，边上的数字表示转移概率。若此刻本体状态为＋1（赞同），下一时刻也为＋1 的概率为 α；同样，此刻状态为－1（反对），下一时刻本体为＋1 的概率为 β，于是推出此时刻＋1 转向下一时刻－1 的概率为 $1-\alpha$，而此时刻为－1 转向下一时刻也为－1 的概率为 $1-\beta$。马尔科夫链的状态图描述了所有状态转移的可能性及其概率，用矩阵表示如下：

$$p_{ij}=\begin{bmatrix}\alpha & 1-\alpha\\ \beta & 1-\beta\end{bmatrix}=\begin{bmatrix}p_{(+1->+1)} & p_{(+1->-1)}\\ p_{(-1->+1)} & p_{(-1->-1)}\end{bmatrix} \tag{6}$$

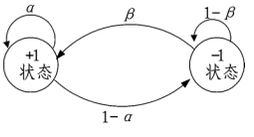

图 1 马尔科夫链本体状态转移图

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

现在问题转化成在给定初始状态概率后，如何求出其他各时刻状态的概率。

初始状态用初始的概率行向量来表示，即：

$$p(t_0)=\begin{bmatrix}p_{+1}(t_0) & p_{-1}(t_0)\end{bmatrix} \tag{7}$$

下一时刻状态的概率为：

$$p(t_1)=p(t_0)p_{ij}=\begin{bmatrix}p_{+1}(t_0) & p_{-1}(t_0)\end{bmatrix}\begin{bmatrix}p_{(+1->+1)} & p_{(+1->-1)}\\ p_{(-1->+1)} & p_{(-1->-1)}\end{bmatrix} \tag{8}$$

式中，p_{ij} 为状态 $i\to j$ 的转移概率，它与所研究对象时序的机理或过程的统计特性有关。考虑到网络舆情的协同效应，引入哈肯提出的序参量作用下的舆情传播概率为马尔科夫链的状态转移概率。

3.2 初始概率和一步协同转移概率

假设网络舆情主体总数为 N，持赞同和反对的初始人数比例为 k/m。

赞同者初始概率为：

$$p_{+1}(t_0)=k/(k+m) \tag{9}$$

反对者初始概率为：

$$p_{-1}(t_0)=1-p_{+1}(t_0) \tag{10}$$

采用协同式(3)、式(4)，求取主体 N 中持本体正向与本体反向的状态转移概率 p_{ij}。

为了在 $k\in[0,10]$ 范围内使式(3)、式(4)的计算结果落在[0,1]范围内，文献[10]对式(3)、式(4)做如下变换，因为将它们各自同时乘上一个常数，这不会改变公式中两者概率的比例关系。为了容易进行仿真，将整体状态的协同转移概率改写为：

$$p_{(+1->-1)}=v\cdot e^{(-(kq+h))}/e^{(k/2)} \tag{11}$$

$$p_{(-1->+1)}=v\cdot e^{(kq+h)}/e^{(k/2)} \tag{12}$$

由于 IPO 马尔科夫过程不符合稳定性假设，因此状态的一步转移概率并非固定不变。随着时间序列的推演，每次需要重新计算协同转移概率。

3.3 算法

```
if confidence
{
  N=361 //假定参与论坛舆情传播的主体个数
  t 为传播循环次数
  t=0 为初始分布时刻
  N 中赞同者与反对者初始分布百分比为 0.70∶0.30 或反之 0.30∶0.70
}
init {
      Set v0=0.5 //初始自转率
      set q(t0)=(N+−N−)/2N
          //初始网络舆情不平衡程度
      Set M(t0)=2p+1(t0)−1
          //初始磁化率
      }
```

仿真实验条件(序参量的变化)如下。

1)初始分布不同情况下，指定偏好度 $h=0$ 不变，改变环境适应度 $k=0,1,2,4,8$。

2)初始分布不同情况下，指定环境适应度 $k=0$ 不变，改变偏好度 $h=0,0.2,0.4,0.6,0.8,1.0$。

3) 初始分布不同情况下，指定环境适应度 $k=0$ 不变，改变偏好度 $h=0,-0.2,-0.4,-0.6,-0.8,-1.0$。

仿真算法：

```
Model1 Algorithm
{
  Do(t=1 to tn) //进入传播循环
  {
    Do calculate P+1->-1
    //调用式(9)计算 N 从本体状态+1 转换到−1 的协同转移概率
    Do calculate P-1->+1
    //调用式(10)计算 N 从本体状态−1 转换到+1 的协同转移概率
    Do
```

$$P(t)=\begin{bmatrix}p_{+1}(t-1) & p_{-1}(t-1)\end{bmatrix}\begin{bmatrix}1-p_{(+1->-1)} & p_{(+1->-1)}\\ p_{(-1->+1)} & 1-p_{(-1->+1)}\end{bmatrix}=\begin{bmatrix}p_{+1}(t) & p_{-1}(t)\end{bmatrix} \tag{13}$$

```
    //计算主体 N 下一时刻本体状态概率
    Let
    N+(t)=361*p+1(t)
    N−(t)=361−N+(t) //计算新分布个数
    Do
  q(t)=(N+(t)−N−(t))/2N //计算新 q
    Do
  M(t)=2p+(t)−1 //计算新 M
  }while(M(t)=M(t−1)≤0.0005)
} //end
```

4 仿真及讨论

4.1 仿真实验结果

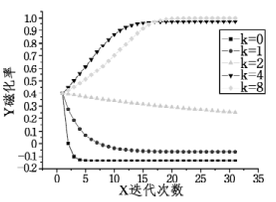

图 2 $h=0$，取 $K=0,1,2,4,8$

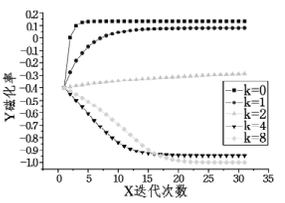

图 3 $h=0$，取 $K=0,1,2,4,8$

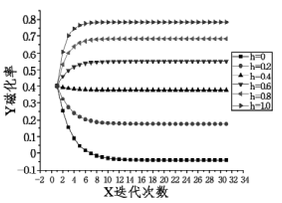

图 4 $K=0$，取 $h=0,0.2,0.4,0.6,0.8,1$

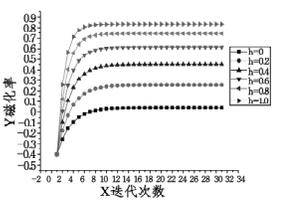

图 5 $K=0$，取 $h=0,0.2,0.4,0.6,0.8,1$

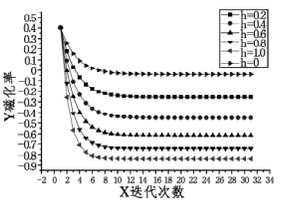

图 6 $K=0$，取 $h=-0.2,-0.4,-0.6,-0.8,-1$

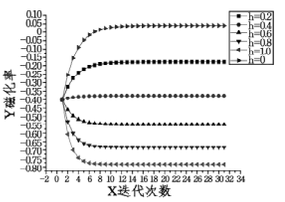

图 7 $K=0$，取 $h=-0.2,-0.4,-0.6,-0.8,-1$

初始分布 $p_{-1}=30\%$，$p_{+1}=70\%$，序参量 h,k 影响下的舆情传播曲线分别如图 2－图 7 所示。其中，X 轴为迭代次数，用 t 表示，Y 轴为磁化率，用 M 表示。

(下转第 235 页)

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其放入统计框架中，以进一步改进新词检测效果，为基于海量语料的机器翻译和舆情热点发现提供支持。

参 考 文 献

[1] 张海军，史树敏，朱朝勇，等. 中文新词识别技术综述[J]. 计算机科学，2010，37(3)：6-10

[2] 刘挺，吴岩，王开铸. 串频统计和词形匹配相结合的汉语自动分词系统[J]. 中文信息学报，1998，12(1)：17-25

[3] 郑家恒，李文花. 基于构词法的网络新词自动识别初探[J]. 山西大学学报：自然科学版，2002，25(2)：115-119

[4] 邹纲，刘洋，刘群，等. 面向 Internet 的中文新词语检测[J]. 中文信息学报，2004，18(6)：1-9

[5] 崔世起，刘群，孟遥，等. 基于大规模语料库的新词检测[J]. 计算机研究与发展，2006，43(5)：927-932

[6] Luo S，Sun M. Two-character Chinese word extraction based on hybrid of internal and contextual measures[C]// Proceedings of the Second SIGHAN Workshop on Chinese Language. Sapporo，Japan，2003：24-30

[7] 罗智勇，宋柔. 基于多特征的自适应新词识别[J]. 北京工业大学学报，2007，33(7)：718-725

[8] 贺敏，龚才春，张华平，等. 一种基于大规模语料的新词识别方法[J]. 计算机工程与应用，2007，43(21)：157-159

[9] 贺敏. 面向互联网的中文有意义串挖掘[D]. 北京：中国科学院研究生院，2007

[10] Sutton C，McCallum A. An Introduction to Conditional Random Fields for Relational Learning[M]. Cambridge MA：MIT Press，2006

[11] CRF＋＋：Yet Another CRF toolkit[EB/OL]. http://chasen.org/～taku/software/CRF＋＋，2009-05-01

(上接第 205 页)

4.2 讨论

1)收敛性：实验表明，无论参变量如何变化，和初始分布不同，算法都能收敛。即最终磁化率大多在经过短时间(t＝6)后，其波动幅度明显减少，逐渐趋向一个稳定值。这与舆情传播的趋势相仿，表明系统能从非平衡态趋向平衡态。仿真表明舆情传播起始 4 个周期内变化梯度较大，其逐步递减，第 4 至第 6 周期变化趋势平缓。

2) K 的影响(见图 2、图 3)：当 h＝0，即不考虑偏好度的影响时，无论正反向初始分布有何不同，k 值在[0，＋1]范围，最终磁化趋向零，说明环境影响较小时，舆情总体分布“正向”“负向”比例相近，但当 K 值继续增大到[2，8]时，环境影响度使得舆情状态向初始分布中多数人意见方向转变。

3) h 的影响(见图 4－图 7)：偏好度 h 为正(见图 4、图 5)代表网络舆情主体对“正向”本体的偏好程度。无论初始分布“正向”概率大或是“负向”概率大，随着时间增加，磁化率迅速从初始值(0.40 或－0.40)分别上升到与 h 值大致相等的稳定值，说明 h 对舆情传播起着主导作用，最终稳定值与初始分布无关。偏好度 h 为负值(见图 6、图 7)代表网络舆情主体对“负向”本体态度的偏好程度，与上述相仿，最终磁化率稳定在 h 值附近。

4) 自转率 v 的影响：由式(11)和式(12)分析，当 k＝0，h＝0时，P＝v，即状态转移概率就是自转率，此时网络舆情空间状态概率仅与自转率和初始分布状态概率相关，v 值影响磁化率起始值的大小。

5) 初始分布的影响：当初始“正向”概率大于“负向”概率时，序参量 k 在[0，8]，h 在[0，1]范围时，磁化率始终在[0，1]的“正向”范围内变化，显示向多数意见靠拢的趋势。当初始时主体持“正向”本体概率小于“负向”概率时，k 在[0，1]的变化只引起磁化率在[0，－1]“负向”范围内变化，同样显示向“负向”多数意见靠拢的趋势。但 k＝0 时，无论 h 为正还是负，最终分布与初始分布无关，强烈趋向靠拢 h 值大小的分布概率。

结束语 本文研究网络舆情传播预测的计算机模型、算法及其仿真。首先，对 IPO 传播过程马尔科夫链引入状态一步协同转移概率，提出一个协同-马尔科夫模型，设计了算法并进行了计算机仿真。同时对仿真的结果做了详细的讨论。

1)当 k 值增大时，环境影响使得网络舆情主体从众心理加大，传播向初始分布的多数意见聚拢。

2)当 h 值增大时，偏好使网络舆情主体根据对本体状态偏好度的正或负分别向磁化率正或负聚拢。

3)网络舆情开始 1～4 次传播变化比较剧烈，需加以注意。同时密切关注偏好度影响起的“正向、负向”放大作用，此时不受初始多数分布的倾向影响。

4)该算法计算复杂度为 O(t)。

由于协同-马尔科夫模型对舆情空间整体状态进行处理，因此计算耗时较少，且因舆情空间整体协同影响强烈，故适合预测 IPO 的传播。未来将研究协同-马尔科夫模型与协同元胞自动机的不同比较，并研究参变量在现实 IPO 中的物理特性和数学表达，以便将仿真结果与 IPO 实际传播数据进行比对，进一步分析模型误差及其原因，提高预测精度。

参 考 文 献

[1] 刘建明. 舆论传播[M]. 北京：清华大学出版社，2001

[2] 张玉峰，王志芳. 基于内容相似性的论坛用户社会网络挖掘[J]. 情报杂志，2010，29(8)：125-130

[3] 章栋兵. 互联网舆情分析关键技术的研究和实现[D]. 武汉：武汉理工大学，2010

[4] Zeng J P，Zhang S Y，Wu C R，et al. Modelling Topic Propagation over the Internet[J]. Mathematical and Computer Modelling of Dynamic Systems，2009，15(1)：83-93

[5] Zeng J P，Zhang S Y，Wu C R，et al. Predictive Model for Internet Public Opinion[C]// Proceedings of Fourth Conference on Fuzzy System and Knowledge Discovery. 2007

[6] Alves S G，Oliveira Neto N M，Martins M L. Electoral Surveys' Influence on the Voting Processes：A Cellular Automata Model[J]. Physica A：Statistical Mechanics and Its Applications，2002，316(1-4)：601-614

[7] 刘慕仁，邓敏艺，孔令江. 舆论传播的元胞自动机模型(I)[J]. 广西师范大学学报：自然科学版，2002，20(2)：1-3

[8] 曾祥平，方勇，袁媛，等. 基于元胞自动机的网络舆论激励模型[J]. 计算机应用，2007，27(11)：2686-2688

[9] 方薇，何留进，等. 采用元胞自动机的网络舆情传播模型研究[J]. 计算机应用，2010，30(3)：751-755

[10] 曾显葵. 基于多数规则和协同规则的元胞自动机舆论传播模型研究[D]. 南宁：广西师范大学，2007

[11] 哈肯 H. 协同学导论[M]. 张纪岳，等译. 西安：西北大学出版社，1981

[12] 沈小峰，胡刚，江璐，等. 耗散结构论[M]. 上海：上海人民出版社，1987

© 1994-2012 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net