ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
E-mail:fcst@public2.bta..net.cn
http://www.ceaj.org
Tel: +86-10-51616056

# C-Rank: 一种 Deep Web 数据记录可信度评估方法*

艾静， 王仲远， 孟小峰+

中国人民大学 信息学院，北京 100872

# C-Rank: A Credibility Evaluation Method for Deep Web Records

AI Jing, WANG Zhong-yuan, MENG Xiao-feng

School of Information, Renmin University of China, Beijing 100872, China

+ Corresponding author: E-mail: xfmeng@ruc.edu.cn

**Abstract:** How to identify and evaluate information credibility ranking has become an increasing important problem. To address the issue, an effective credibility evaluation method called C-Rank to compute trust values of records in Deep Web databases is proposed, which constructs an S-R Credibility Graph for each record. The graph contains 2 types of vertices and 3 types of edges. Firstly, each vertex's local trust value is computed by using out-degrees based on the idea of trust propagation. Then, the weight of Record vertex is computed by using its in-degree and adjacent Site vertices' local trust values. Lastly the global trust value of the S-R Credibility Graph is computed, which denotes the record's credibility in the whole Web. Experiment results show C-Rank can evaluate credibility rankings of records appropriately and discriminate false information effectively. This method is generally applicable to all domains of Deep Web.

**Key words:** Deep Web; Web Information Credibility; S-R Credibility Graph; Trust Propagation

**摘 要**：针对 Web 信息可信度问题，提出了一种为 Deep Web 数据记录计算可信度的有效方法 C-Rank。该方法为每一条记录构造一个 S-R 可信度网络，包含两种类型顶点及三种类型边。首先基于可信度传播的思想，利用顶点出度为每一个顶点计算其局部可信度值；再利用 Record 顶点入度及相邻 Site 顶点的可信度值，为该 Record 顶点计算权值；继而求得整个 S-R 网络的全局可信度值。实验证明，C-Rank 方法能够合理而有效地评价数据记录的可信度，从而达到甄别虚假信息，为用户推荐可信数据记录的目的。该方法普遍适用于 Deep Web 的各个领域。

**关键词**： 深层网络; Web 信息可信度; S-R 可信度网络; 可信度传播

**文献标识码**：A　　**中国分类号**：TP391

## 1 引 言

随着网络与通信技术的迅速发展，Web 上的信息呈现爆炸性增长，互联网上存储了海量信息。其中，面向领域的 Deep Web[1]数据源是网络上一种非常重要的数据源，包含极为丰富的结构化数据，成为人们获取信息的一个重要途径。

以工作信息领域为例，越来越多的用户开始依赖互联网来获取招聘信息，寻找理想的工作。但是，面对纷繁复杂的数据源（如招聘网站、博客、论坛等），以及每日成千上万条新发布的招聘信息，用户无法对所有招聘信息的可信度进行正确辨别。而这些信息的可信度可能会涉及到如下两个重要问题：

（1）用户的隐私泄露问题

通常应聘者的简历上包含许多个人信息。因此，一些不法分子通过发布虚假的招聘信息，收集用户的简历，从而获取他的个人信息，并将其转让给第三方，

*the grants from the Natural Science Foundation of China under Grant No.60833005, 60573091(国家自然科学基金项目); National High-Tech Research and Development Plan of China under Grant No.2007AA01Z155，2009AA011904(国家863高技术研究发展计划); the Ph.D. Programs Foundation of Ministry of Education of China No.200800020002(教育部博士点基金项目)
Received 2009-07, Accepted 2009-

从而导致用户隐私泄露问题。更有甚者，利用简历中的个人信息去填写信用卡申请表，然后刷卡或提现，给应聘者造成巨大的损失。

因此，这是一个不容忽视的问题，也是为 Deep Web 数据记录计算可信度的迫切需求。

（2）数据记录最优选择问题

在择业时，招聘公司的数量远远超过一名应聘者所能够了解的能力范围。面对一家未知的公司，应聘者常常需要花费大量的时间和精力，来查询这家公司是否是一家与其描述相符的公司。例如，其有可能是一家刚进入中国不久的世界 500 强企业，也可能是一个随时都会倒闭的“皮包公司”。用户只有了解了这条招聘信息的真实可信度，才能够根据自己的实际情况进行正确的选择。

在其它领域，如网上购书、预订机票、转让二手物品等，其数据记录的可信度问题也是关系到用户切身利益的一个重要问题。因此，数据记录的可信度问题在 Deep Web 的许多领域里都广泛存在。

为了解决数据记录的可信度问题，通过对工作领域数据记录的观察我们发现，Web 上的招聘信息，其可信度具有如下两个重要特点：

（1）某条招聘信息所发布的网站的可信度越高，通常这条招聘信息的可信度就越高。如图 1 所示，是一条 Google 招聘工程师的信息，它在多个网站上发布。由于这条招聘信息出现在了 Google 的网站上，以及高校招生就业网等高可信度的网站，所以这条招聘信息的可信度也比较高。

（2）某条招聘信息被转载的次数越多，通常这条招聘信息的可信度就越高。这是由于用户在转载这条招聘信息的时候，通常会进行一些判断。也就是说，在转载的时候，这条信息就包含了一定的人类的知识在其中。在图 1 中，这条 Google 的招聘信息，被转载到了论坛、博客等网站。

忽视上述任何一点，都无法完全断定一条数据记录的可信度。例如，一则招聘信息虽然有可能只发布在一家著名公司的网站上，但是这条招聘信息的可信度就极高；而如果一个恶意的信息发布者，即使他在许多论坛上重复发布、转载同一条招聘信息，这条信息仍然是不可信的。因此，为了真实展现一条 Deep Web 上的招聘记录的可信度，只有综合上述两个特点，才能够计算得出正确的可信度值。

通过如上两点在工作信息领域的观察，本文提出了一种为 Deep Web 中的数据记录自动构建一个可信度网络，通过计算局部可信度值以及全局可信度值，为用户评价一条 Deep Web 数据记录可信度的方法。


**Fig.1** A recruitment record of Google on the Web

**图 1** Web 上关于 Google 的一条招聘信息

这种方法，通过设置传播率参数以及采取适当的全局可信度计算方法，易于扩展到 Deep Web 的各个领域，具有较强的适用性。

## 2 相关工作

关于 Web 上的信息可信性研究，是目前的一个研究热点问题。文献[2]是早期的工作之一，提出了一种将名声、信誉、口碑等信任管理的社会机制引入计算中网络中，辨别信息可信度的方法。

在 Web 上的信息可信性问题上，研究者们通常比较关注新闻网站[3,4]以及有评论系统网站[5,6]上的信息可信度。这些工作一般都是以网页为基本单元，通过各种方法将该网页上信息的可信度计算出来。这些网页，绝大多数是 Surface Web 页面。而 Deep Web（即 Web 数据库）中的记录的可信性问题，就较少被研究者关注。而本文研究的正是工作信息领域中的招聘记录的可信性问题。

Deep Web 数据库中的记录之间并不是完全孤立隔离开的。根据记录与网站之间的链入和链出关系，以及不同的 Deep Web 数据库中记录重复出现的关系，可以形成网状结构，该网络的节点不再是通常的网页，而是一条条的记录。而在网状结构上辨别和维护节点的可信度，最常用的方法是信任值传播机制。在文献[7]提出的方法中，每个节点只需维护一个与它有直接关联的邻居节点的可信度值向量，网络中其他节点的可信度值均可根据节点信任向量的共享机制以及传播法得到。文献[8,9]也是传播法和信任值传递模型的典型代

表。然而，在由 Deep Web 记录组成的网络中，并不是所有的节点之间都是相异的。关于同一领域的不同网站的 Deep Web 数据库之间经常会有重叠部分，即同一条记录可能会在网上重复出现多次。因此网络中可能有多个节点表示同一条记录。这是我们的研究问题和上述文献中的问题的不同之处，在目前已有的可信性研究相关文献中均未涉及到如何在这种情况下进行信任值的传播与计算。

在这种网络上做节点可信度计算和维护的研究，还需要解决表示同一条记录节点之间的信任值相互影响的问题，以及它们最终合并的问题。在这方面，Deep Web 上的实体识别[10]是一个值得借鉴的工作，该文献提出了一种在两个Deep Web数据库之间识别出相同记录的方法，对于我们辨别哪些记录节点之间有这种关系以及后续的信任值影响研究有很大的帮助。

## 3 问题分析

Deep Web 数据库中记录的可信度，与它所在的网站的可信度值有关，也与链接指向它的网站的可信度值有关。

首先，设想这样一个例子：如果 Google 在它的网站上发布了一条“招聘工程师”的记录，显然地，这条记录的可信度一定比较高。如图 2 所示，这条记录发布在 google.com 网站上，而 google.com 本身是一个可信度值非常高的网站，因此它发布的信息的可信度也比较高。

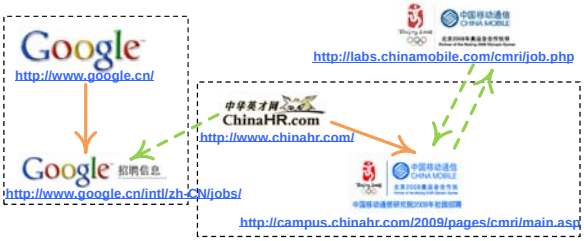


**Fig.2** Relationships among recruitment records on the Deep Web

**图 2** Deep Web 上招聘信息链接关系

而与此对应的是，专门的招聘信息发布网站，如51job，智联招聘等网站，包含成千上万条招聘信息。与上面例子中 Google 发布招聘信息不同，专门的招聘信息发布网站包含的记录并不是由这个网站发布的，网站只是给发布招聘信息的公司提供一个平台。即使51job 网站本身的可信度值较高（确实是一家发布工作信息的专门网站），也并不能说明它上面所有招聘记录的可信度值都比较高。

还有一些公司，如中国移动研究院，并不把招聘的详细信息放在自己的网站上，而是外包给 51job 这样的平台，在 51job 上发布完整的招聘记录。而这样的记录，一般都会带有超链接，指向中国移动研究院的网站。而中国移动研究院的网站上也有指向这条招聘记录的链接，如图 2 所示。这种记录的可信度值也是比较高的。

此外，还有一些数据记录会被不断转载，如图 1 所示，Google 招聘工程师的记录被好几个高校的就业网转载，也被 51job、智联招聘、中华英才网等工作信息网站转载，而每一次转载，都相当于经过了转载者（或网站的建设者）的一次对其可信度的鉴别，所以如果一条记录被多个网站所指向，或重复出现在多个网站中，也能说明它的可信度较高。这相当于是综合了多个转载用户的集体智慧和判断。

通过以上分析，我们可以得出在 Deep Web 中为数据记录计算数据可信度需要考虑如下特点：

(1) 可信度值越高的网站，其发布的数据记录的可信度值也越高；

(2) Web2.0 信息共享平台、Deep Web 数据发布平台需要与专业机构等传统的网站平台区分开；

(3) 不同数据源之间的可信度值可以通过链接相互传递；

(4) 同一数据记录在不同数据源出现次数越多，其可信度值越高。

综上所述，一条招聘记录的可信度是多种因素共同作用的结果。基于工作信息领域的大量观察，我们设计了一个算法，引入这些影响因素，为 Deep Web 中的数据记录计算可信度。本文将在下一章节进行重点介绍。

## 4 S-R可信度网络构建与计算

### 4.1 面向Deep Web的S-R可信度网络

在 Deep Web 中，每一个领域都存在着大量的数据源。信息在数据源中以记录的形式存在。通常情况下，同一个领域的不同数据源之间是相互隔绝的，不存在直接的联系。然而，同一条记录可能会重复出现在多个不同的数据源中，因此数据源之间可能会有内容重叠。而且，不同数据源中的记录之间虽然没有直接的关联，但是通过网站之间的关联关系，也可以为记录建立间接的联系。根据数据源之间的重叠以及记录与网站之间的关系，本文提出了一种针对某一条记录构建 S-R 可信度网络的方法。

**定义 1（S-R 可信度网络）**针对 Deep Web 中某一条记录而构造的一个包含两种类型顶点、三种类型边的网络。

在 S-R 可信度网络中，顶点包含两类：

(1)**Site 顶点($v^s$)**：含有数据记录的网站。例如，在招聘信息领域，这类顶点指的就是一个个网站，包括专业工作信息发布网站，例如中华英才网、51job 等；公司或机构的网站；论坛、博客、bbs 以及各大高校的就业网站等。

(2)**Record 顶点($v^r$)**：各个网站上的数据记录。在一个 S-R 可信度网络中的所有 Record 顶点，代表的都是同一内容的数据，只是它们可能散布在不同的数据源里。例如，在招聘信息领域，这类顶点指的就是不同网站的同一条招聘记录。

S-R 可信度网络的边包含三类：

(1)**内部链接边($e^i$)**：如果一个记录是属于某一个数据源的，那么在 S-R 网络中，这条记录对应的 Record 顶点与这个数据源对应的 Site 顶点之间存在一条有向边，边的方向是从 Site 顶点指向 Record 顶点。这种有向边表示网站与它包含的记录之间的关系，因此被称为内部链接边。

(2)**外部链接边($e^o$)**：记录与记录，以及记录与外部数据源之间的链接关系用外部链接边表示。如果一条记录中包含一个指向外部数据源的链接，或某个数据源里有一个链接指向其他数据库中的一条记录，那么在 S-R 网络中，其对应的顶点间存在一条有向边，边的方向与链接指向的方向相同；

(3)**实体识别边($e^r$)**：如果属于不同数据源的两条记录经过实体识别技术[10]验证，被认为是表示同一实体，则这两个 Record 顶点之间存在一条无向边。

这样，按照如上定义，我们就可以为 Deep Web 中的某一条记录，通过链接关系以及实体识别技术，为其构造出一个 S-R 可信度网络，如图 3 所示。

在一个 S-R 网络中，所有 Record 顶点（用椭圆形表示）在理论上应该都是表示同一条记录。因此每一条记录都有一个它自己的 S-R 网络图。目前，两两数据源之间的实体识别能够达到较高的准确率，而大规模数据源之间的实体识别工作仍未取得有效成果，因此，本文所构造的 S-R 网络中，所有 Record 顶点之间并不是一个完全子图。

S-R 网络中的多个 Record 顶点表示同一条记录在整个 Web 上重复出现在不同数据源中的情况。矩形节点表示与这些记录有关联（内部链接边和外部链接边）的网站。

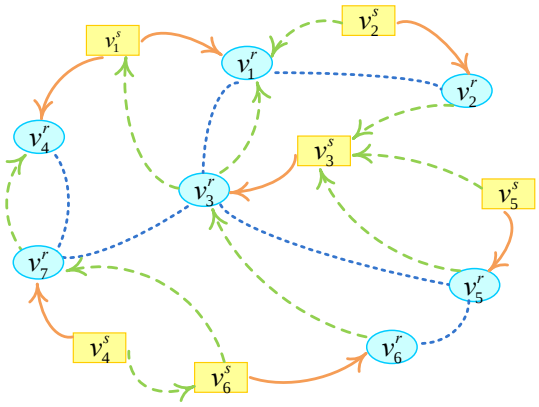


**Fig.3** S-R credibility network

**图 3** S-R 可信度网络

在开始计算这条记录的可信度值之前，要为网络中的每一个顶点赋予一个初始可信度值。构造初始可信度值的方法有很多种（比如 Site 顶点取 PR 值，Record 顶点为 0），但是这不是本文所关注的重点，本文的重点在于可信度值传播算法，因此这部分不展开叙述。关于顶点初始可信度值的设置我们会在实验中有进一步的说明。

本文用传播迭代法为 S-R 网络中的每一个节点计算可信度值。一个顶点的可信度值可以通过有向边以及无向边传递给其相邻顶点。这种传播行为在整个 S-R 网络的所有节点之间进行。然后，将此 S-R 网络中所有顶点的可信度值以合适的方式合并起来，计算出整个网络的全局可信度值，就是这条 Deep Web 中的记录的可信度值。

S-R 可信度网络可以应用于 Deep Web 中的各个领域。只是在不同的领域中，可信度传播方法以及全局可信度值的计算可能会有所不同。

以下，将着重讨论一种在 Deep Web 中，利用 S-R 可信度网络来计算某一条数据记录的可信度的方法：C-Rank。

### 4.2 S-R可信度网络局部可信度值计算

在构造 S-R 可信度网络之后，本文利用可信度传播的思想，利用顶点的出度，在 S-R 可信度网络中迭代计算可信度传播率，从而得到 Record 顶点的局部可信度值。

**定义 2（局部可信度值）**在 S-R 可信度网络中，每一个顶点的可信度值称为局部可信度值。

不同于以往的网络，本文所提出的 S-R 网络包含两类顶点、三类边。因此不同类型边的可信度的传播率也不相同。

如图 4 所示，图中的黄色实线箭头表示内部链接边 $e^i$，绿色虚线箭头表示外部链接边 $e^o$，蓝色虚线边表示实体识别边 $e^r$。

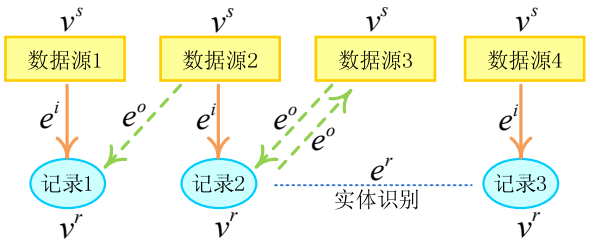


**Fig.4** Credibility propagation graph

**图 4** 可信度传播数据图

三种类型的边拥有不同的含义，因而拥有不同的可信度传播率。我们为这三种类型的边设置三种传播率类型：

(1)内部链接边的传播率类型为 $\alpha(e_G^i)$；

(2)外部链接边的传播率类型为 $\beta(e_G^o)$；

(3)实体识别边的传播率类型为 $\gamma(e_G^r)$。

对于 S-R 图中的每一条边，首先判断它是属于哪种类型的边，根据传播率类型和相应的边的出度，再计算这条边的实际传播率。例如，如果边 $e_k^i=(u\to v)$ 是 S-R 图中的第 k 条边，它属于 $e^i$ 边，其传播率类型为 $\alpha(e_G^i)$，那么这条边的传播率为：

$$\alpha(e_k^i)=\begin{cases}\dfrac{\alpha(e_G^i)}{OutDeg(u,e_G^i)}, & OutDeg(u,e_G^i)>0\\ 0, & OutDeg(u,e_G^i)=0\end{cases} \quad (1)$$

其中，$OutDeg(u,e_G^i)$ 表示顶点 $u$ 发出的 $e^i$ 类型边的数量。类似的，我们可以求出属于 $e^o$ 类型边以及属于 $e^r$ 类型边的传播率。

需要注意的是，由于 $e^r=(u,v)$ 是一个无向边，我们将其看成是两个单向边，即 $u\to v$ 的边以及 $v\to u$ 的边，然后分别处理这两条边。

然后，我们利用 PageRank[11]以及 ObjectRank[12]的基本思想，根据 S-R 网络的特点加以改进，经过 n 次可信度值传播后的顶点可信度值，来迭代计算第 n+1 次传播后顶点的可信度值：

$$R_{loc}^{(n+1)}=dAR_{loc}^{(n)}+\frac{(1-d)}{|S|}s \quad (2)$$

其中，$R_{loc}^{(n)}$ 表示经过 n 次可信度值传播后的顶点可信度值向量，它其中的每一个元素 $r_{loc}^{(n)}(v_i)$ 表示经过 n 次可信度值传播后的顶点 $v_i$ 的局部可信度值；$A$ 是一个 $m\times m$ 的矩阵，$A$ 中的每一个元素 $A_{ij}=\alpha(e^i)$；d 为阻尼系数；$S$ 是所有顶点的一个任意子集；而 $s=(s_0,\ldots,s_i,\ldots,s_m)^T$ 为基本向量，如果一个顶点 $v_i\in S$，则 $s_i$ 等于 1，否则为 0。

根据文献[13]，这样的迭代计算必然会趋向收敛。设定一个任意小的向量 $\varepsilon$，当符合如下条件时，迭代停止：

$$|R_{loc}^{(n+1)}-R_{loc}^{(n)}|<\varepsilon \quad (3)$$

至此，我们为 S-R 可信度网络中的每一个顶点求得了它的局部可信度值。

### 4.3 S-R可信度网络全局可信度值计算

正如前文所提到的，整个 S-R 可信度网络实际上是关于一条记录而构建的一个网络，因此，当我们求得所有顶点的局部可信度值后，需要综合考虑所有 Record 顶点的局部可信度值，求得全局可信度值。

**定义 3（全局可信度值）**整个 S-R 可信度网络的可信度值，它代表了此 S-R 网络对应的招聘记录在 Web 上的总体可信度值。

为了计算整个网络的全局可信度值，我们采用几种不同的方法综合 S-R 网络中所有 Record 顶点的局部可信度值。

**方法一：求和法**。将所有 Record 顶点的局部可信度值求算术和。如公式（4）所示：

$$C\text{-}Rank(S\text{-}R)=\sum r_{loc}(v_i) \quad (4)$$

这种方法反映了一条记录重复出现次数越多，可信度值越高的情况。

**方法二：最大值法**。在所有的 Record 顶点中，取局部可信度值最大的作为这条记录的全局可信度值。如公式（5）所示：

$$C\text{-}Rank(S\text{-}R)=\max\{r_{loc}(v_i)\,|\,i=1,\ldots,m\} \quad (5)$$

最大值法的合理之处在于：同一条记录在不同网站上多次出现，如果有一次能够被证明可信度是非常高的，那么这条记录应该也是非常可信的。

**方法三：顶点加权法**。上述两种方法虽然在一定程度上可行，但是无法正确处理虚假信息恶意转载以及中小型公司的招聘信息可信度问题。

根据问题分析中所提到的关于信息可信度的观察，在 Deep Web 中一条记录的可信度与其发布的数据源的可信度以及重复出现的次数均相关。因此，在根据局部可信度值计算全局可信度值时，需要考虑不同的 Record 顶点具有不同的权值。

$$\omega(v_i)=\sum InDeg(v_i^{\{i,o,r\}})\times\{\overline{\alpha}(e_{ji}^{i}),\overline{\beta}(e_{ji}^{o}),\overline{\gamma}(e_{ji}^{r})\mid\exists e_{ji}=v_j\rightarrow v_i\}$$
$$+\sum r(v_j^{s})\times\{\alpha(e_{ji}^{i}),\beta(e_{ji}^{o}),\gamma(e_{ji}^{r})\mid\exists e_{ji}=v_j\rightarrow v_i\wedge v_j\in v^{s}\} \quad (6)$$

在公式(6)中，$InDeg(v_i^{\{i,o,r\}})$表示指向顶点$v_i$的$e^i$、$e^o$以及$e^r$类型的边的数量(对于$e^r$边，将其看成两个单向边)；$\overline{\alpha}(e_{ji}^{i}),\overline{\beta}(e_{ji}^{o}),\overline{\gamma}(e_{ji}^{r})$表示顶点$v_i$属于$\alpha(e_G^i)$、$\beta(e_G^o)$和$\gamma(e_G^r)$传播率类型的边的平均传播率值；$r(v_j^s)$表示与顶点$v_i$相邻的 Site 顶点的可信度值；而$\alpha(e_{ji}^{i}),\beta(e_{ji}^{o}),\gamma(e_{ji}^{r})$表示的是 Record 顶点$v_i$与 Site 顶点$v_j$的边的传播率。

综合考虑了顶点的入度以及相邻 Site 顶点的可信度值，我们可以得到 S-R 图中的每一个 Record 顶点的权值，使用公式(7)将顶点的权值进行标准化：

$$\omega_{nor}(v_i)=\frac{\omega(v_i)}{\max\{\omega(v_k),k=0,\ldots,m\}} \quad (7)$$

在得到所有 Record 顶点的局部可信度值及其对应的权重后，我们通过加权计算，得到整个 S-R 网络的可信度值：

$$C\text{-}Rank(S\text{-}R)=\sum\omega_{nor}(v_i)r_{loc}(v_i) \quad (8)$$

顶点加权值法可以有效地防止恶意的信息发布者在多个网站上重复发布、转载同一条虚假招聘信息。如果恶意发布者想通过这种方法使记录的出现次数增多来达到骗取一定的可信度值的目的，也不可能获得较高的全局可信度值。同时，这种方法也能够为中小型公司的招聘信息计算出一个符合实际情况的可信度值。

## 5 实验

目前，在 Deep Web 上，没有为数据记录计算可信度值的相关工作，因此本文实验的主要目的在于验证 C-Rank 方法的有效性与合理性。在包含 1000 条招聘记录的数据集上，使用 C-Rank 方法为每一条记录进行了可信度值的计算，并邀请了用户为实验效果进行打分评价。在本节中给出了实验结果及用户反馈情况。

### 5.1 数据集

本实验的数据集需要从工作信息领域获取。我们使用 Jobtong[14]（一个工作信息领域的数据集成原型系统）从 Deep Web 数据源中爬取招聘信息记录。我们首先用 Jobtong 取到任意的 900 条不同的招聘记录，然后手工加入 100 条不可信的招聘记录。这 100 条招聘记录是我们从网上手动收集的，是由网友揭发出来的骗简历的虚假招聘信息，或完全名不符实的“皮包公司”。

该数据集的构造方法使它具有如下特点：第一，实验者知道哪些记录必定是不可信的，可以验证可信度值计算方法是否能够有效地将这些记录鉴别出来；第二，所有记录均从结构化数据源中获取，信息完备；第三，数据集中的所有招聘记录均为随机选取。基于这三个特点，该数据集可保证实验结果的客观性。

### 5.2 实验设置

在实验开始之前，使用 4.1 节中的方法为每一条招聘记录自动构造它的 S-R 可信度网络。Site 顶点的初始可信度值设为该网站的 PR 值。Record 顶点的初始可信度值为 0。如果某个网站是属于论坛、博客、信息集成系统这类允许自由发布信息的开放式平台，则将其初始可信度置为 PR 值与一个折扣系数的乘积。折扣系数可以由统计实验得到，也可以由人工根据经验设定。折扣系数是一个[0,1]区间内的值。

### 5.3 实验分析

我们使用 4.2 节中的方法，为这 1000 条记录的 S-R 网络中的每个顶点计算了局部可信度值，然后使用顶点加权法（经实验证明，求和法与最大值法的效果不如顶点加权法）计算记录的全局可信度值，并将其标准化为[0,1]区间内的值。实验结果如图 5 所示，横坐标表示每一条招聘记录的唯一标号，纵坐标表示这条记录对应的可信度值。

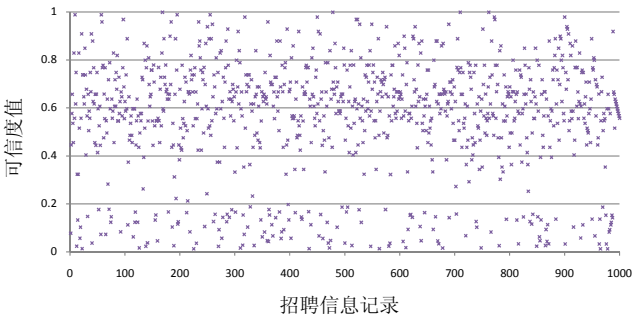


**Fig.5** Scatter diagram of records credibility values distribution

**图 5** 记录可信度分布散点图

经过确认，我们发现，在实验中人工加入的 100 条不可信记录全都分布在可信度值为[0, 0.2)的区间内。同时，从图 5 可以看出，可信度值的分布呈现一定的聚集效果。为了进一步量化这种聚集规律，我们将可信度值平均分为 5 个区间：[0, 0.2), [0.2, 0.4),……, [0.8, 1]。这 5 个区间代表 5 个可信度等级，等级越高越可信。

可信度等级 1～5 级中包含的记录数量如图 6 所示，分别为 156 条、45 条、294 条、390 条，115 条。

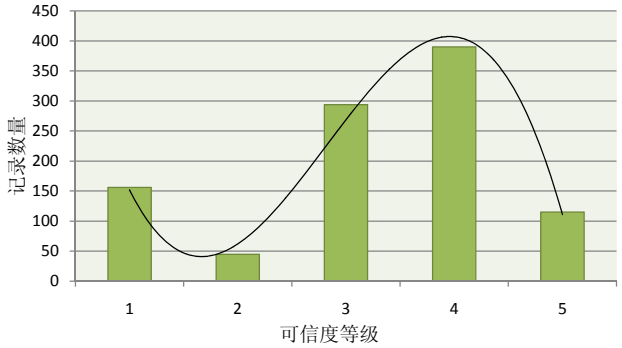


**Fig.6** Record numbers of different credibility levels

**图 6** 不同可信度等级的记录数

从图 6 中的趋势线可以看出，不可信记录全部被识别出来（经人工确认，第一等级[0, 0.2)区间中的 156 条记录完全包含了人工加入的 100 条不可信记录）。并且，第二等级的记录数相对较少，是整个趋势线的波谷，这说明了 C-Rank 方法能够有效地将可信记录与不可信记录区分开来。而第二等级到第五等级之间的记录数目呈现正态分布的趋势，这是符合真实的网络环境中招聘记录的可信度的分布情况的，这也说明了我们的方法能够对招聘记录的可信度给出合理而合适的评价值。

为了验证本文方法对于可信度的评价值是否准确，我们邀请了 10 名用户对实验结果进行评价。系统随机为每一名用户分配 100 条招聘记录（包括招聘职位、内容描述、记录来源等信息），以及用 C-Rank 方法为其计算的一个可信度值。用户根据人工判断，对每条记录的可信度值是否合理进行评价，评价的内容包括：合理、偏高、偏低、不合理。用户评价结果如图 7 所示。

从图 7 中的结果可以看出，用户认为使用 C-Rank 方法计算招聘信息可信度具有较高的合理性。10 名用户评价的平均合理率达到 94.2%，而认为偏高或偏低的记录只占 1.8%及 2.7%，认为不合理的记录仅占 1.3%。经过确认，用户认为偏高、偏低或不合理的记录多是容易导致主观性较强的一些招聘信息，例如一名用户以私人邮箱为公司招聘技术开发人员，但内容比较详细，所以用户仅凭内容分析无法做出正确判断。从这个意义上来说，本文提出的方法计算出的可信度值也能够为这些用户提供有价值的参考信息。

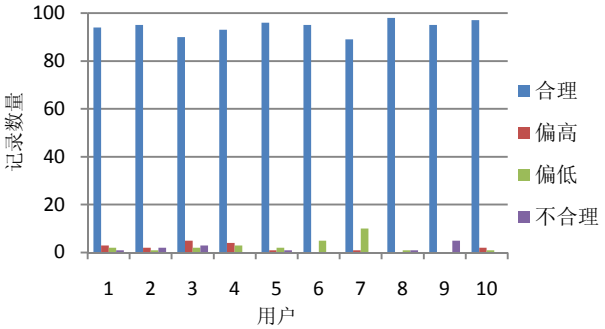


**Fig.7** Users' feedbacks for records' credibility values

**图 7** 用户对于记录可信度分值的评价结果

## 6 结束语

本文提出了一种基于传播机制的 Deep Web 数据记录可信度评估方法。该方法为每一条招聘记录构造一个 S-R 可信度网络，利用 S-R 网络中顶点的出度以及可信度传播机制为每一个 Record 顶点计算一个局部可信度值，然后使用入度及相邻 Site 顶点的可信度值为每一个 Record 顶点计算权重，求得关于这条记录的 S-R 网络的整体可信度值，作为该记录的全局可信度值。通过在 Deep Web 工作信息领域的实验证明，C-Rank 方法能够合理而有效地评价招聘记录的可信度，从而达到甄别虚假信息，为用户推荐可信的、更有价值的招聘记录的目的。根据用户对于计算结果的评价，证明了用 C-Rank 方法计算出的记录可信度值符合实际情况。

Deep Web 上的数据记录可信度问题具有很强的现实意义。本文所提出的解决办法具有实际操作性，且易于扩展到其它领域。

**References:**

[1] He B, Patel M, Zhang Z, et al. Accessing the Deep Web: A Survey[J]. Communications of the ACM (CACM), 2007, 50(5): 94-101.

[2] Alfarez A-R, Hailes S. Relying On Trust to Find Reliable Information[C]//Proceedings of International Symposium on Database, Web and Cooperative Systems (DWA-COS'99), 1999. [2009-04-25].http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.38.7934

[3] Lee R, Kitayama D, Sumiya K. Web-based Evidence Excavation to Explore the Authenticity of Local Events[C]//Proceedings of WICOW 2008, California: ACM, 2008: 63-66.

[4] Kawai Y, Fujita Y, Kumamoto T. Using a Sentiment Map for Visualizing Credibility of News Sites on the Web[C]//Proceedings of WICOW 2008, California: ACM, 2008:53-58.

[5] Wanas N, El-Saban M, Ashour H, et al. Automatic Scoring of Online Discussion Posts[C]//Proceedings of WICOW 2008, California: ACM, 2008:19-25.
[6] Staddon J, Chow R. Detecting Reviewer Bias through Web-Based Association Mining[C]//Proceedings of WICOW 2008, California: ACM, 2008:5-9.
[7] Kamvar S D, Schlosser M T, Garcia-Molina H. The EigenTrust Algorithm for Reputation Management in P2P Networks[C]//Proceedings of the 12th World Wide Web Conference, Budapest: ACM, 2003: 640-651.
[8] Ziegler C-N, Lausen G. Propagation Models for Trust and Distrust in Social Networks[J]. Information Systems Frontiers, 2005, 7:4/5: 337–358.
[9] Guha R, Kumar R, Raghavan P, et al. Propagation of Trust and Distrust[C]//Proceedings of the 13th World Wide Web Conference, New York: ACM, 2004: 403-412.
[10] Elmagarmid A K, Ipeirotis P G, Verykios V S. Duplicate Record Detection: A Survey[J]. IEEE Transactions on knowledge and Data Engineering (TKDE), 2007, 19(1): 1-16.
[11] Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[C]//Proceedings of the 7th World Wide Web Conference. Brisbane: ACM, 1998: 107-117.
[12] Balmin A, Hristidis V, Papakonstantinou Y. ObjectRank: Authority-Based Keyword Search in Databases[C]// Proceedings of the 30th International Conference on Very Large Data Bases(VLDB'04), Toronto: ACM, 2004: 564-575.
[13] Motwani R, Raghavan P. Randomized Algorithms[M]. United Kingdom: Cambridge University Press, 1995.
[14] Jobtong. [2009-4-25]. http://www.jobtong.cn

## 作者简介

**AI Jing** was born in 1985. She received the B.S. degree from Renmin University of China in 2007, and now is a M.S. candidate in Computer Application and Technology at Renmin University. Her research interest focuses on Web data management.

艾静(1985- )，女，北京人，2007 年于中国人民大学获计算机科学与技术专业工学学士学位，目前是中国人民大学计算机应用技术专业硕士研究生，主要研究领域为 Web 数据管理。

**WANG Zhong-yuan** was born in 1985. He received the B.S. degree from Renmin University of China in 2007, and now is a M.S. candidate in Computer Application and Technology at Renmin University. His research interest focuses on Web data management.

王仲远(1985- )，男，福建仙游人，2007 年于中国人民大学获计算机科学与技术专业工学学士学位，目前是中国人民大学计算机应用技术专业硕士研究生，主要研究领域为 Web 数据管理。

**MENG Xiao-feng** was born in 1964. He received the Ph.D degree in Computer Application and Technology from Institute of Computing Technology Chinese Academy of Sciences in 1999. Now he is a professor and doctoral supervisor at Renmin University. His research interests include web data management, native XML databases, mobile data management, etc.

孟小峰(1964- ), 男, 河北邯郸人, 1999 年于中国科学院获计算机应用技术专业工学博士学位, 目前是中国人民大学教授, 博士生导师, 主要研究领域为 Web 数据管理、XML 数据库、移动数据管理。