doi:10.3969/j.issn.1005-3026.2016.02.004

基于可信度的投票列表合并算法

杨红果,申德荣,寇 月,于 戈
(东北大学 计算机科学与工程学院,辽宁 沈阳 110819)

摘 要:在投票系统中,每个投票人按照自己对候选人的认可程度对候选人进行排名,从而得到大量的有序投票列表.为了从这些列表中得到一个综合投票结果,需要找到一种合理有效的列表合并算法,综合分析列表数据并将它们合并为一个综合列表.本文提出一种基于可信度的投票列表合并算法,其基本思路是:通过综合分析投票列表中蕴含的众多排名信息,度量出每个列表中每条排名信息可被采信的程度,简称为可信度,然后基于已经得到的可信度,让那些高可信度的排名信息在综合排名中发挥更大的作用,从而得到一个更好的综合排名结果.实验结果充分表明,本文提出的算法能够更有效地挖掘出排名信息的可信度,从而得到准确度更高的合并结果.

关 键 词:列表;投票系统;列表合并;可信度;综合排名
中图分类号:TP 301 **文献标志码**:A **文章编号**:1005-3026(2016)02-0165-05

Credibility-based Algorithm for Merging Vote Lists

YANG Hong-guo, *SHEN De-rong*, *KOU Yue*, *YU Ge*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: SHEN De-rong, E-mail: shenderong@ise.neu.edu.cn)

Abstract: In a voting system, each voter makes a preferential list about candidates, thus a large amount of ordered lists are obtained. To get a comprehensive voting result from these lists, an effective lists merging algorithm is required, which can analyze these lists data and output a comprehensive list. A merging algorithm based on credibility is proposed. Through analyzing the data of lists, numerous ranking messages are extracted, then the credibility of them is formulated and measured, with which the final comprehensive list is computed such that those ranking messages with high credibility could play a more influential role in the final ranking result. Experimental results fully indicate that the algorithm proposed can dig out the credibility about ranking information more effectively, thus attaining the merging results more accurately.

Key words: list; voting system; lists merging; credibility; comprehensive ranking

在投票系统中,每个投票人按照自己对候选人的认可程度,对候选人按照从高到低的顺序进行排名,从而得到大量的有序列表.为了生成一个综合投票结果,需要找到一种合理有效的列表合并算法,综合分析这些列表数据,将其合并为一个综合列表.本文提出一种基于可信度的列表合并算法.在现实中,不同投票人的可信度一般不相同,因此不同投票人所提供的排名信息的可信度也不同;同时,同一个投票人对不同候选人的了解程度也是不相同的,导致由同一个投票人提供的关于不同候选人的排名信息的可信度也不相同.基于这些启发式规则,本文提出一种基于可信度的列表合并算法.该算法首先通过综合分析众多投票列表,度量出投票人的可信度以及对候选人的了知度;然后综合以上二者信息,进一步计算出每个列表中每条排名信息的可信度;最后基于已经得到的排名信息的可信度值,综合多个投票列表中关于相同候选人的排名信息,让那些高可信度的排名信息在综合排名中发挥更大的作用,从而得到候选人的综合排名结果.

收稿日期:2014-11-06
基金项目:国家重点基础研究发展计划项目(2012CB316201);国家自然科学基金资助项目(61033007, 61472070).
作者简介:杨红果(1983-),女,河南邓州人,东北大学博士研究生;申德荣(1964-),女,辽宁铁岭人,东北大学教授,博士生导师;于 戈(1962-),男,辽宁大连人,东北大学教授,博士生导师.

现有的投票算法大致可以归为两类:一类是单议席投票算法[1-3],如波达投票法 BC(Borda count)[2-3] 等;与之对应的是多议席投票算法[4-6],如单一可转移票制 STV(single transferable voting)[4-5]等;第二类是随机投票算法[7-8].这些算法均没有考虑信息的可信度因素,使得最终得到的投票结果不是很理想.基于可信度的算法在其他领域(如实体识别[9]、元搜索[10]等)也有涉及,但迄今为止,投票分析系统中关于投票人可信度方面的研究尚未见报道.本文提出一种新的适用于投票系统的计算可信度的方法,用以计算每个投票人以及投票信息的可信度,并将之应用到投票结果的综合分析中.

1 标记符号和投票系统

本文算法的基本标记符号如下:m 代表候选人数量;n 代表投票人数量;e_i 指代第 i 个候选人;v_k 指代第 k 个投票人;由投票人 v_k 给出的排名列表用 L^k 表示;投票人 v_k 的可信度为 c^k;第 i 个候选人 e_i 在列表 L^k 中的排名用符号 r_i^k 表示;符号 f_i^k 用来表示投票人 v_k 对候选人 e_i 的排名 r_i^k 的"了知度";而符号 c_i^k 则代表排名信息 r_i^k 的可信度;d_{ij}^k代表 e_i 到 e_j 在列表 L^k 中的排名距离,其值为 $d_{ij}^k = r_j^k - r_i^k (1 \leqslant i < j \leqslant m)$;与符号 f_i^k 类似,本文中用 f_{ij}^k代表投票人 v_k 对差距信息 d_{ij}^k 的"了知度";同时符号 c_{ij}^k代表排名信息 d_{ij}^k的可信度;最后符号 $s_i^k = m - r_i^k$ 代表候选人 e_i 在列表中的得分.

一个投票系统一般包含候选人、投票人、投票列表以及采用的统计算法.其目的是根据该统计算法和相关投票列表数据,得到一个综合投票结果列表.表 1 中给出一个全文通用的例子.假设有 3 个候选人,5 个投票人,一共得到 5 个有序投票列表,投票结果如表 1 所示.

表 1 投票列表

Table 1 Voting lists

e_i	L^1	L^2	L^3	L^4	L^5
e_1	2	3	1	1	2
e_2	3	2	3	2	3
e_3	1	1	2	3	1

由表 1 可知,候选人数 $m = 3$,投票人数 $n = 5$.此时需要将表 1 中给出的多个投票列表,合并为一个综合的排名结果.现有的投票列表合并算法很多,典型的有波达投票法 BC 和单一可转移票制 STV.

2 BC 算法和 STV 算法

在 BC 算法中,先算出每个候选人的综合得分.候选人 e_i 的综合得分是该候选人在所有列表中的得分 $s_i^k = m - r_i^k$ 的总和,即 $s_i = s_i^1 + s_i^2 + \cdots + s_i^n$,然后按照 s_i 的大小顺序得到最终的综合排名.BC 算法本质上是根据多数规则提出的,即获得投票人支持最多的候选人获胜.

STV 算法按照以下规则得到综合列表:首先统计所有列表中排在第一位的候选人的得票数,并将得票最多的候选人排在第一位.如表 1 中,排在第一的候选人 e_1 得票数是 2,e_3 得票数是 3,则将 e_3 排在综合列表中的第一位(如果排在第一位的有多个候选人,则通过一定的机制处理).其次将第一轮中落选的候选人的一票转移到排在第二位的候选人身上.如在列表 L^3 中,将落选的 e_1 的那一票加到 e_3 上,e_3 为 2 票,L^4 中 e_2 为 2 票.然后统计所有列表中排在第二位的候选人的得票数,可知这一轮 e_1 得 2 票,e_2 得 3 票,e_3 得 2 票(但不参与竞争),e_2 胜出.如此迭代直至第 m 轮,得到最终的综合列表.STA 算法本质上遵从的是最大满意度原则,即尽量使最终结果让大多数投票人感到满意.

现有的投票列表合并算法均没有考虑排名信息的可信度问题.

3 基于可信度的列表合并算法

本文提出两个基于可信度的算法:基于距离的算法和基于排名的算法.

3.1 基于距离可信度的列表合并算法

由表 1 可以得到任意两个候选人之间的排名距离 $d_{ij}^k (d_{ij}^k = r_j^k - r_i^k)$,如表 2 所示.由 d_{ij}^k的定义可知,d_{ij}^k和 d_{ji}^k是反对称的,因此只需考察其中的一种情况,这里只考虑 $i < j$ 的情况,m 个候选人一共有 $m(m-1)/2$ 对组合.

表 2 候选人之间的排名距离

Table 2 Rank distance between candidates

d_{ij}^k	d_{ij}^1	d_{ij}^2	d_{ij}^3	d_{ij}^4	d_{ij}^5
d_{12}^k	1	-1	2	1	1
d_{13}^k	-1	-2	1	2	-1
d_{23}^k	-2	-1	-1	1	-2

基于距离可信度的投票列表合并算法主要分

为两部分：首先对表 2 中的数据进行有效分析，挖掘并度量出投票人 v_k 对任意一对候选人之间的差距的了知度 $f_{ij}^k(1\leqslant i<j\leqslant m)$；然后基于已知的了知度的值，从一个全新的角度计算出距离信息 d_{ij}^k 可被采信的程度（即可信度 c_{ij}^k），并用于最终的综合排名计算中.

3.1.1　了知度的计算

对任意给定的一对候选人 e_i 和 $e_j(i<j)$，首先假定每个投票人 $v_k(1\leqslant k\leqslant n)$ 对他们之间差距的初始了知度是均等的，即 $f_{ij}^k=1/n(1\leqslant k\leqslant n)$. 然后依据式（1）计算每对候选人的平均距离 $\overline{d_{ij}}$ $(1\leqslant i<j\leqslant m)$，即以 f_{ij}^k 的概率来采信每个投票人 v_k 给出的距离信息 d_{ij}^k，得到期望值 $\overline{d_{ij}}$：

$$\overline{d_{ij}}=\sum_{k=1}^{n}f_{ij}^k\times d_{ij}^k. \tag{1}$$

这个期望距离 $\overline{d_{ij}}$ 近似反映了 e_i 到 e_j 的真实差距. 因此通过 $|d_{ij}^k-\overline{d_{ij}}|$ 的大小，可以看出投票人 v_k 对每对候选人之间的差距的了知度. 下面通过式（2）计算出新的了知度值：

$$f_{ij}^k=\exp\frac{-(d_{ij}^k-\overline{d_{ij}})^2}{\sigma}\Big/\sum_{l=1}^{n}\exp\frac{-(d_{ij}^l-\overline{d_{ij}})^2}{\sigma}. \tag{2}$$

其中参数 σ 起调节作用，当 σ 变小时，即使很小的 $|d_{ij}^k-\overline{d_{ij}}|$ 的变化也会导致较大的 f_{ij}^k 的变化，反之则变化平缓.

显然新的了知度值更加准确，因此再次利用式（1）就可以计算出更加合理的均值 $\overline{d_{ij}}$，然后再次用式（2）计算出更加准确的了知度值 f_{ij}^k. 如此迭代 t 次，用第 t 次迭代结果 f_{ij}^k 作为最终的了知度值.

3.1.2　距离可信度及综合排名

首先计算每条距离信息 d_{ij}^k 的可信度 c_{ij}^k，然后基于此进行综合排名. 通过 3.1.1 节，已经挖掘出了知度的值 f_{ij}^k，显然了知度的值越大，距离信息 d_{ij}^k 的可信度 c_{ij}^k 越高；但在本文中并不简单地设定 $c_{ij}^k=f_{ij}^k$，因为距离信息 d_{ij}^k 的可信度 c_{ij}^k 的大小，除了与投票人 v_k 对其的了知度 f_{ij}^k 有关外，还与该投票人自身的可信度 c^k 有关.

在现实生活中，对于一个低可信度的人，即使他对某些人很了解，人们也很难采信他所给出的信息；因此在本文的计算模型中，除了考虑了知度 f_{ij}^k 因素外，也考虑投票人 v_k 的可信度 c^k 因素，据此度量一条距离信息的最终可信度.

首先通过式（3）评估投票人 v_k 的可信度 c^k，用该投票人对所有距离信息的了知度的平均值，即平均了知度，来度量该投票人的可信度；因为一个投票人对候选人之间的差距信息 d_{ij}^k 的了知度 f_{ij}^k 越高，则他的可信度就越高.

$$c^k=\sum_{1\leqslant i<j\leqslant m}f_{ij}^k/[m(m-1)/2]. \tag{3}$$

然后通过式（4）计算距离 d_{ij}^k 的可信度 c_{ij}^k. 即一个投票人 v_k 的可信度 c^k 越高，且对候选人 e_i 到 e_j 的相对差距的了知度 f_{ij}^k 越高，则 v_k 给出的距离信息 d_{ij}^k 的可信度 c_{ij}^k 也越大.

$$c_{ij}^k=\delta\times c^k+(1-\delta)\times f_{ij}^k,\quad 0\leqslant\delta\leqslant 1. \tag{4}$$

式中 δ 为投票人可信度在综合可信度中的权重.

显然可信度 c_{ij}^k 反映了距离信息 d_{ij}^k 可被采信的程度. 在 e_i 到 e_j 的最终距离 d_{ij} 的计算中，按照 c_{ij}^k 的大小采信距离信息 d_{ij}^k，使可信度高的距离信息发挥更大的作用. 距离 d_{ij} 的计算如式（5）所示：

$$d_{ij}=\sum_{k=1}^{n}c_{ij}^k\times d_{ij}^k. \tag{5}$$

最后根据 d_{ij} 的值及以下规则生成最终的综合列表：如果 $d_{ij}>0$，则将元素 e_i 排列到元素 e_j 之前，否则相反. 由定理 1 可知，基于距离 d_{ij} 的排名不会发生冲突情况.

定理 1　如果 $d_{ij}>0$ 并且 $d_{jz}>0$，那么 $d_{iz}>0$.

证明　$d_{iz}^k=r_z^k-r_i^k=(r_z^k-r_j^k)+(r_j^k-r_i^k)=d_{ij}^k+d_{jz}^k$，代入式（5）可得 $d_{iz}=d_{ij}+d_{jz}>0$. 证毕.

3.2　基于排名可信度的列表合并算法

基于排名可信度的算法，其大体的思想和方法和基于距离的基本一致. 在该算法中，首先通过相似的迭代方法计算出每条排名信息 r_i^k 的了知度 f_i^k，进而通过相似的方法，综合了知度 f_i^k 的值和投票人 v_k 的可信度 c^k 值求出 r_i^k 的可信度 c_i^k，然后通过式（6）得到每个候选人 e_i 的综合得分 s_i. 最后按照 s_i 的大小对所有候选人进行综合排名.

$$s_i=\sum_{k=1}^{n}c_i^k\times s_i^k=\sum_{k=1}^{n}c_i^k\times(m-r_i^k). \tag{6}$$

4　实　　验

本文的实验数据集分为两类. 一类是自动生成的：首先假定一个标准排名，然后模拟现实中人的不同判断力和了知度而自动生成投票列表数据，然后运用本文提出的算法，得到合并结果，最后通过计算结果向量和标准排名的 cosine 相似度来度量算法的准确度，比较不同算法的优劣. 模拟生成的实验数据集的规模为 1 000，即 1 000 种模拟投票数据. 另一类是一组真实的投票结果：给定 10 位明星，由实验室同学根据其知名度的高低各

自给出一个排名,一共得到 16 个排名结果,然后根据本文提出的算法,得到一个综合排名,最后将综合结果和中国福布斯名人排行榜中的这 10 个人的相对排名进行对比,用以度量算法的优劣.

本文分别提出了基于距离可信度的合并算法 DA 和基于排名可信度的合并算法 RA. 为了全面考察算法 RA 和 DA 的性能,从以下 4 个方面对算法进行分析:①在模拟数据集上对比算法 RA,DA,BC 和 STV;②在真实数据集上对比算法 RA,DA,BC 和 STV;③考察式(2)中参数 σ 对 RA,DA 算法的影响;④考察式(4)中参数 δ 对 RA,DA 算法的影响.

图 1 比较了不同算法在模拟数据集和真实数据集(即由 16 位同学对 10 位明星按照知名度进行排序的投票结果集)上的准确度,其中纵坐标准确度的取值是在 1 000 组实验数据上所得结果的平均值. 可以看出,本文提出的算法 RA 和 DA 总体上优于现有算法 BC 和 STV,这主要是因为 RA 和 DA 算法不是简单地统计票数,而是通过数据分析和迭代计算挖掘出排名信息的可信度;另外 DA 算法总体上要优于 RA 算法,这主要是因为 DA 算法通过分析更加复杂的两两关系信息,通过相互制约,使得最终的投票结果的准确度可得到更高的保证. 从总体上看,真实数据集上的效果不如模拟数据集,一方面是由于各位同学总体上对明星不是很熟悉,另一方面是由于数据量少(投票人数少),使得算法的分析优势不能得到有效发挥,这也反映了 RA 和 DA 算法适合于数据量相对较大的投票场景中.

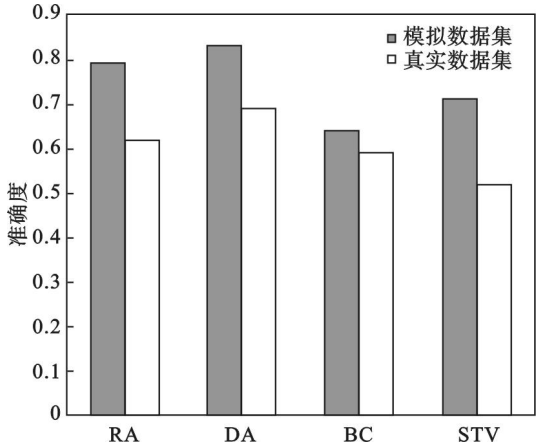

图 1 各算法在模拟数据集和真实数据集上的准确度

Fig. 1 Algorithm precision on simulated and real data sets

图 2 反映了算法中参数 σ,δ 对算法准确度的影响. 可以看出,当 σ 的取值在 1.1 左右时,算法的准确度最高,同时可以看出基于距离的合并算法 DA 的总体效果好于基于排名的合并算法 RA. 即在计算排名信息可信度时,那些偏离平均值远的排名信息,其可信度的取值不能太大也不能太小.

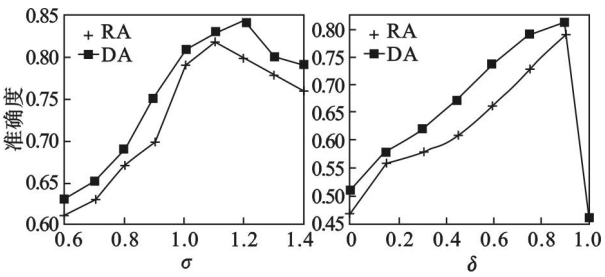

图 2 参数对算法准确度的影响

Fig. 2 Effect of parameters on algorithm precision

由图 2 可以看出,当 $\delta = 0.8 \sim 0.9$ 时,算法会取得比较好的计算效果. 也就是说,一条排名信息的可信度主要取决于其投票人的可信度,同时投票人对它的了知度虽然在可信度计算中的权重 $(1-\delta)$ 较低,但它有效区分了同一个列表中的不同排名信息的可信度(如,某个投票人的可信度虽然很低,但由于该投票人对某几个候选人很熟悉,那么该列表中这几个候选人的相对排名顺序的可信度很有可能比其他人高). 由图 2 也可以看出,当完全依赖于投票人可信度($\delta = 1$)时,算法效果急剧下降.

5 结 语

本文首次提出了基于可信度的投票算法. 首先算法通过对投票列表数据进行综合分析,挖掘出每条排名信息的可信度. 可信度的计算过程中不仅考虑了投票人的可信度因素,而且考虑了投票人对候选人的了知度,从而使得最终算得的排名信息的可信度值更加全面准确. 在最终的综合排名中,利用得到的可信度的值,加权计算出候选人的排名值,从而使得最终的排名结果更加合理有效.

今后的工作,希望将多人的兴趣列表融合为一个群体兴趣列表. 这就需要结合兴趣的具体特点提出一种新的算法,综合得到群体兴趣爱好,从而为更好的群体决策和服务提供依据和支撑.

参考文献:

[1] Kelly R, Lester P, Durkin M. Leadership elections: labour party[J]. *House of Commons Library Standard Note*, 2010, 26(5): 5 - 8.

(下转第 173 页)