基于属性坐标的文本信息检索模型*

The Text Information Retrieval Model Based on Attribute Coordinates

李广原[1]，**冯嘉礼**[2]

Li Guangyuan[1]，Feng Jiali[2]

（1. 广西师范学院信息技术系，广西南宁 530001；2. 上海海事大学信息工程学院，上海 200135）

（1. Info. Tech. Dept.，Guangxi Teachers Edu. Univ.，Nanning，Guangxi，530001，China；2. Info. Engi. Coll.，Shanghai Maritime Univ.，Shanghai，200135，China）

摘要：文本和用户查询用属性坐标表示，以交点与查询重心点的距离确定为文本与查询间的相似度进行计算，利用相关性反馈技术调整检索策略，得到一个基于属性坐标的文本信息检索模型. 实验表明，该模型的检索方法可行，检索效果较好.

关键词：文本信息　检索　属性坐标系　相似度　相关反馈

中图法分类号：TP391.1　　**文献标识码**：A　　**文章编号**：1002-7378(2005)04-0225-03

Abstract：A kind of text information retrieval model based on attribute coordinates is presented. In this model，the texts and the queries are represented using attribute coordinates. The distance of the point of intersection to the barycenter of query is represented as the similarity between the text and the query，and be calculated. With the use of the relevant feedbacks，the retrieval tactichas been changed. The efficient of information retrieval is improved.

Key words：text information，retrieval，attribute coordinate，similarity，relevance feedback

随着网上信息的大量涌入，信息的检索研究面临着新的发展机遇与挑战. 信息检索中，文本信息占据着重要的地位，据不完全统计，超过90%的信息检索属文本信息检索. 传统的信息检索方法主要有布尔检索、向量空间检索和概率检索等方法，为提高检索效率，人们又提出了不少新的检索方法，如Salton提出扩展布尔检索[1]，W. 旺提出广义向量空间模型[1]. 除此之外，还有基于神经网络的检索方法、基于推理网络的检索方法、基于贝叶斯网络的检索方法，以及模糊检索方法等. 本文提出文本和用户查询用属性坐标表示，依据相关理论与方法进行文本与用户查询间的相似度计算，利用相关反馈信息调整检索策略，给出一个基于属性坐标表示与计算的文本信息检索模型. 实验表明，基于此模型的检索方法不仅可行，而且检索效果较好.

收稿日期：2005-06-09

作者简介：李广原(1969-)，男，广西上林人，硕士，讲师，主要从事信息检索和数据挖掘研究。

* 广西师范学院青年科研基金资助项目。

1 相关定义

定义1[2]　设$M(X)$，$N(X)$分别表示事物X的不同属性，用$\wedge$表示合取算子，则用$M(X)\wedge N(X)$表示属性的合取过程，令$S(X)=M(X)\wedge N(X)$，则$S(X)$称为合属性，而$M(X)$和$N(X)$称为素属性，以合属性代表事物X.

设事物X的属性集$P(X)=\{E_0(X),E_1(X),\cdots,E_n(X)\}$，则有：

定义2[2]　设n维单纯形$K=(E_0,E_1,\cdots,E_n)$，其顶点为属性集$P(X)$中的第$n+1$个属性，则K为属性多面体. 在K的第一次重心剖分$K^{(1)}$中，$r+1$个属性的整合属性$E_{i0}\wedge E_{i1}\wedge\cdots\wedge E_{ir}$置放在由这$r+1$个属性所构成的$r$维单纯形的重心剖分点上，记为$P(S_{ir})$，且$P(S_{ir})=E_{i0}\wedge E_{i1}\wedge\cdots\wedge E_{ir}$. 依次类推，这样的模型称之为属性重心剖分模型，或称属性坐标系.

设事物$P(X)=(T_1,T_2,T_3)=(0.8,0.7,0.3)$其中，$T_1,T_2,T_3$为$P(X)$的素属性，则用属性重心剖分模型可表示如图1所示，$T_1,T_2,T_3$构成了属性

坐标系的 3 个坐标轴，三角形 $\triangle ABC$ 的重心代表了事物 $P(X)$.

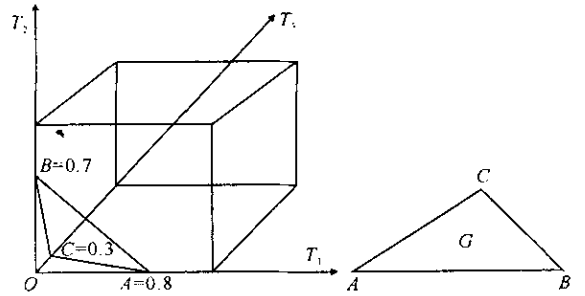

图 1 属性重心剖分模型

2 基于属性坐标的文本检索模型

2.1 文本和查询的属性坐标表示

基于属性坐标系的检索模型中，文本和查询用属性坐标表示，设文本集 D 中共使用了 n 个标引词标引其中的文本，记 n 个标引词为 $t_1, t_2, \cdots, t_n$. 文本集中每一文本和用户每一次查询均可用等长的向量表示，即

$$d_i = (w_{i1}, w_{i2}, \cdots, w_{in}).$$

同理，$q_j = (w_{j1}, w_{j2}, \cdots, w_{jn})$，

其中，w_{ik} 表示第 k 个词在文本 d_i 中的权值，则

$$w_{ik} = \frac{n_k \times N}{M \times S_K},$$

其中，n_k 为词 t_k 在文本中出现的次数；M 为所有标引词在 d_i 中出现的次数的最大值；N 为文本集中文本的数目；S_K 为包含词 t_k 的文本个数. 根据上式，w_{ik} 可能大于 1，为便于在属性坐标系上表示文本与查询，可对权值作归一化处理.

定义 3[2] 文本向量 $d_i = (w_{i1}, w_{i2}, \cdots, w_{in})$ 所确定的多面体的重心称为文本重心 G_{di}，$G_{di} = (\frac{w_{i1}}{n}, \frac{w_{i2}}{n}, \cdots, \frac{w_{in}}{n})$. 同理，查询向量 $q_j = (w_{j1}, w_{j2}, \cdots, w_{jn})$ 所确定的多面体的重心称为查询重心 G_{qj}，$G_{qj} = (\frac{w_{j1}}{n}, \frac{w_{j2}}{n}, \cdots, \frac{w_{jn}}{n})$，则可用文本重心向量和查询向量分别代表文本和查询.

2.2 相似度计算

基于属性坐标的表示法中，连结坐标原点至各文本重心点所形成的直线与查询向量所确定的平面相交，设交点为 K，交点与查询重心点的距离可确定为文本与查询间的相似度. 设查询向量为 $(w_{j1}, w_{j2}, \cdots, w_{jn})$，文本 d_i 的重心点坐标为 $G_{di} = (w_{di1}, w_{di2}, \cdots, w_{din})$，则文本重心线方程为：

$$\frac{x_1}{w_{di1}} = \frac{x_2}{w_{di2}} = \cdots = \frac{x_n}{w_{din}}.$$

因此，交点 K 的坐标为

$$\begin{cases} \dfrac{x_1}{w_{di1}} = \dfrac{x_2}{w_{di2}} = \cdots = \dfrac{x_n}{w_{din}}, \\ \dfrac{x_1}{w_{j1}} + \dfrac{x_2}{w_{j2}} + \cdots + \dfrac{x_n}{w_{jn}} = 1 \end{cases}$$

的解. 这样，求解相似度的匹配函数为：

$$f: sim(d_i, q) = 1 - \frac{r(G_q, k_{di})}{R},$$

其中，G_q 为查询重心点；k_{di} 为文本与查询向量所确定的平面的交点；$r(G_q, k_{di})$ 为查询重心点与交点的距离，不妨称其为查询距离；R 为平面上至查询重心点的距离，用于控制输出. 当 $r(G_q, k_{di})$ 大于 R 时，则不输出 d_i，R 的第一次取值可取在查询平面上的点至查询重心点的最大距离，以后每次取值，从上一次用户选择的文档集中取最小的查询距离作为本次检索的 R 值.

2.3 相关性反馈检索

相关性反馈检索即利用上次的查询结果与用户的选择来指导本次查询. 利用查询结果相关信息的反馈进行检索是现代信息检索常采用的技术与策略，采用上述模型进行文本信息检索时，为提高检索的查准率，本文利用相关信息的反馈来进行检索. 从检索的结果来分析，还存在这样的文本，尽管它们的查询距离小于，但仍不符合用户需要，例如设有查询 $q = (0.8, 0.7, 0.6)$，如有文本 $d = (0.3, 0.2, 0.2)$，则该文本的查询距离很小，但它极有可能由于关键词词权值太小而导致的.

符合用户需要的文本应满足的条件：一是其查询距离小于 R，二是相关词的权值应满足一定的条件. 下面讨论第二个条件.

设 $D = \{d_1, d_2, \cdots, d_n\}$ 为一次检索后系统输出的文本集；$D' = \{d_1, d_2, \cdots, d_m\}$ 为一次检索后用户选择的文本集；$\overline{D} = \{d_1, d_2, \cdots, d_l\}$ 为用户排除的文本集；显然 $D = D' \cup \overline{D}$，且 $D' \cap \overline{D} = \Phi$. 对于 D' 中的某一文本 d_i，有 $d_i = (w_{i1}, w_{i2}, \cdots, w_{in})$，其中，$w_{ik}$ 为关键词 k 在文本 $i (k \in 1, 2, \cdots, m)$ 中的权值. 对于 $\overline{D}$ 中的某一文本 d_j，有 $d_j = (\omega_{j1}, \omega_{j2}, \cdots, \omega_{jn})$，其中，$\omega_{ki}$ 为关键词 i 在文本 $k (k \in 1, 2, \cdots, l)$ 中的权值. 下面对第二个问题进行讨论，即相关词的权值应满足的条件. 首先给出一个词权值向量：

$$W = (\gamma_1, \gamma_2, \cdots, \gamma_n), \quad (1)$$

其中，取 γ_i 的值等于 $w_{ki} (k = 1, 2, \cdots, m; i = 1, 2, \cdots, n)$ 中的最小值. 这里从 (1) 式起，构建 $m (0 \leqslant m \leqslant n)$

组词权值向量：$(\alpha_1,\alpha_2,\cdots,\alpha_n)$，$(\beta_1,\beta_2,\cdots,\beta_n)$，…，$(\eta_1,\eta_2,\cdots,\eta_n)$，其中 $\alpha_i,\beta_i,\cdots,\eta_i$ 是对应于词 i 的权值. 这 m 组权值向量，称之为权值阈值向量，对于这 m 组向量中的任一组，文本集 D' 和 $\overline{D}$ 中的文本应满足以下条件：

D' 中不存在任一文本 d_i，使得

$$(w_{i1}\leqslant\alpha_1\wedge w_{i2}\leqslant\alpha_2\wedge\cdots\wedge w_{in}\leqslant\alpha_n)\vee(w_{i1}\leqslant\beta_1\wedge w_{i2}\leqslant\beta_2\wedge\cdots\wedge w_{in}\leqslant\beta_n)\vee\cdots\vee(w_{i1}\leqslant\eta_1\wedge w_{i2}\leqslant\eta_2\wedge\cdots\wedge w_{in}\leqslant\eta_n)\quad(2)$$

成立.

同时，$\overline{D}$ 中至少存在一个以上的文本，任取其中的一个文本 d_i，使得

$$(\omega_{i1}\leqslant\alpha_1\wedge\omega_{i2}\leqslant\alpha_2\wedge\cdots\wedge\omega_{in}\leqslant\alpha_n)\vee(\omega_{i1}\leqslant\beta_1\omega_{i2}\leqslant\beta_2\wedge\cdots\wedge\omega_{in}\leqslant\beta_n)\vee\cdots\vee(\omega_{i1}\leqslant\eta_1\wedge\omega_{i2}\leqslant\eta_2\wedge\cdots\wedge\omega_{in}\leqslant\eta_n))\quad(3)$$

成立.

实际上，对这 m 组向量中的任一组，都是在保留 D' 中文本的前提下，尽可能多地排除 $\overline{D}$ 中的文本. 根据(2)、(3)式可以得到，对于待检文本集中的任一文本 d_i，经过一次检索后，在下次检索中，采用相关信息的反馈来进行检索，如果符合下列两条件之一的，则被系统排除，不作输出.

(Ⅰ) 查询距离 $r_i>R$；

(Ⅱ) $\begin{cases}r_i\leqslant R\\ \delta_i\leqslant\alpha_i\end{cases}$ 或 $\begin{cases}r_i\leqslant R\\ \delta_i\leqslant\beta_i\end{cases}$ 或…或 $\begin{cases}r_i\leqslant R\\ \delta_i\leqslant\eta_i\end{cases}(i=1,2,\cdots)$，

其中，δ_i 代表文本集中某一文档中第 i 个关键词.

以第 i 组为例，从(1)式权值向量出发，寻找这 m 组权值阈值向量需经过以下步骤：

步骤 1：保持(1)式中 $n-i$ 个权值不变，每次分别提高 $i(m\geqslant i\geqslant 1)$ 个关键词的权值的大小，步幅可取 0.1；

步骤 2：如果有 D' 中的文本使(2)式不成立，或权值增大到超过 1 时，则停止；记录此时符合(3)式的 $\overline{D}$ 中的最大文本数目 T，$T\geqslant 0$，记录此时的权值向量，然后，把在步骤 1 中提高的权值恢复到(1)式开始时的值，再选择另外的 i 个(与上次 i 个关键词不完全相同) 关键词，转步骤 1.

步骤 3：当所有 i 个权值都完成上述两步后，选择取得 T 值最大时的权值向量，该权值向量即为该组的最终权值阈值向量.

3 检索分析

为检验基于属性坐标的文本信息检索模型的可行性与检索效率，选择了国内计算机期刊中有关计算机技术的文献 45 篇，其中包括计算机网络，机器学习，算法设计与分析，数据挖掘，数据库设计等专业. 检索实验设计了若干个查询式，涵盖了 45 篇文献所包含的关键字，为了与传统的检索模型进行比较，选择与向量空间模型作为比较的对象，并选择若干查询表达式，这些查询表达式涵盖了上述计算机几个方向的相关关键词. 通过计算机检索实验，我们选择2 个和3 个权值作变化，求出相关的权值阈值向量，然后利用相关反馈检索，得到基于属性坐标的文本信息检索模型查全率和查准率(表 1).

表 1 计算机技术文献的相关查全率和查准率

查询表达式	向量空间模型(%)		相关性反馈检索(%)	
	查全率	查准率	查全率	查准率
计算机网络	81	75	75	82
算法设计与分析	82	74	74	78
机器学习	74	80	75	83
数据挖掘	76	76	71	85
数据库设计	75	78	70	81

从表 1 可以看出，无论是查全率还是查准率，采用相关性反馈检索后，检索的查全率基本相当，略有下降，但查准率有所提高；基于属性坐标的文本信息检索模型是可行的，且效果也不错. 事实上，基于此模型思想的一个检索方法已在实际检索系统中采用[2]，此模型不仅可应用于文本信息检索，还可应用到决策支持系统，如核事故应急评估与决策支持子系统[3]. 此外，此模型还可用于对事物进行分类、聚类及机器学习.

4 结束语

本文给出一个基于属性坐标表示与计算的文本信息检索模型，在此基础上，采取相关反馈技术，对检索策略进行了调整，取得较好的效果. 尽管如此，仍存在要改进的地方，如求权值阈值向量的计算量比较大，当关键词个数很高时，问题比较突出，为此，可对关键词进行筛选、合并与聚类，缩减关键词数量，从而减少计算工作量，另外，可考虑采用并行计算方法求权值阈值向量，以加快下一次检索的速度，这是下一步要研究的课题.

参考文献：

[1] Baeza Yates R, Ribirero Neto B. Moden Information Retrieval[M]. Addison Wesley: Longman Publishing, 1999.

(下转第 231 页)

精确度,达到了鉴定标准.

表1 运营和计算数据

日期	理论值(千元)	实际值(千元)	业务量理论值(万人)	业务量计算	设备预测精度	系统载荷预测精度
2004-10	1771.38	1800	29.52	30		
2004-11	1836.33	1860	30.60	31		
2004-12	1929.60	1920	32.16	32		
2005-01	2015.22	1980	33.58	33		
2005-02	2094.38	2040	34.90	34		
2005-03	2168.09	2100	36.13	35		
2005-04	2237.15	2160	37.28	36	0.636	0.745
2005-05	2302.25	2220	38.37	37	0.646	0.893
2005-06	2363.94	2280	39.39	38	0.712	1.108

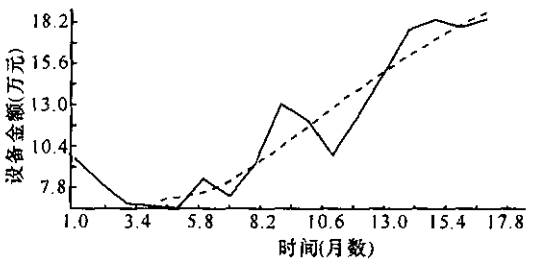

图1 运营曲线拟合

——:实际运营曲线;……:模型曲线

5 结束语

运营设备需求预测是IT企业控制成本的关键,许多大型IT企业为此提出了预算制度.本文在分析相关理论的基础上,提出了运营系统设备需求的鉴定标准,给出了相应的预测模型.从目前的3次迭代来看,准确率一直在提升,整个运营设备预算方案将会更为流畅,更工具化,这也是我们后续要继续推进的工作.

参考文献:

[1] 李懋和.概率论和数理统计[M].长春:吉林大学出版社,1999.

[2] 张 弛.6sigma实战[M].广州:广东经济出版社,2001.

[3] Hans Van Vliet. Software Engineering Principles and Practice[M]. Second Edition. New York: John Wiled & Sons, 2002.

(责任编辑:黎贞崇)

(上接第227)

[2] 潘谦红,王 炬,史忠植.基于属性论的文本相似度计算[J].计算机学报,1999,22(6):651-655.

[3] Feng J L. The research on decision supports system of nuclear accident emergency and its computer realization [D]. Beijing: Chinese Atomic Energy Institute, 2001, 97-118.

[4] 冯嘉礼.基于属性抽取与整合的感觉神经检测模型[J].计算机研究与发展,1997,34(7):481-486.

[5] Gerard Salton, Chris Buckley. Improving retrieval performance by relevance feedback [J]. J of the American Society for Information Science, 1990, 41(4):288-297.

[6] Salton G, Buckley C. Term-weighting approaches in automatic retrieval. Infor-mation [J]. Processing and Management, 1988, 24(5):513-523.

(责任编辑:黎贞崇)