基于粗糙集的K均值聚类算法在案例检索中的应用

陈 千 向 阳 郭 鑫 王 栋
(同济大学电子与信息工程学院计算机科学与技术系 上海 201804)

摘 要 在基于本体的案例检索系统中,由于数据库中的案例数量随着时间的推移而成倍增加,案例检索的效率不断降低,因此如何有效地提高案例检索系统的效率是个亟待解决的问题。提出一种基于粗糙集的 K-means 聚类算法,在用户检索之前对案例库中成千上万的案例进行有效聚类,从中定义基于粗糙集的聚类中心和上下近似以及边界。实验证明,该方法在系统检索时不必对每个案例都进行相似度的计算,从而大大提高了检索性能。

关键词 粗糙集,K 均值聚类,本体,案例检索
中图法分类号 TP391 **文献标识码** A

Rough-set-based *K*-means Clustering Algorithm in Case Retrieval

CHEN Qian XIANG Yang GUO Xin WANG Dong
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

Abstract The retrieval efficiency will fall down due to the increasing numbers of cases in Database in marketing Ontology based case retrieval system. How to effectively improve the efficiency of case retrieval system is a serious problem. A K-means clustering algorithm based on rough set was proposed which can do the clustering work on thousands of marketing cases in Case Retrieval System in order to figure out the center of each cluster. So there is no need to do the similarity work in terms of each case. It is improved through experiment that the methods can largely enhance the performance of case retrieval system.

Keywords Rough set, K-means clustering, Ontology, Case retrieval

1 引言

案例的获取对一个企业的决策支持起着基础性作用。通过对特定的案例进行检索,企业的决策者能够更方便地获取最相似的案例,从而指导决策。现代信息检索技术的飞速发展,为案例检索提供了技术上的指导。由于案例数据具有不可确定性、不完整性以及不精确性,因此针对这种不完备的信息可采用模糊智能计算的相关理论研究成果实施有效的数据聚类和检索。

基于营销本体的案例检索系统目前采用了概念匹配和概念相似度的方法,但是数据库中的营销案例的数量随着时间的推移会逐渐呈指数级增长,使得系统在用户提交的样本案例与数据库中的案例进行逐一本体概念匹配的过程中效率大大降低,最终难以满足终端用户的需求。作为数据挖掘中比较活跃的领域,聚类分析技术近年来发展迅速,根据应用背景的不同聚类方法主要有以下几种:划分方法、层次方法、基于密度的方法、基于网格的方法和基于模型的方法[2]。现有的划分方法存在着执行代价高、计算复杂等缺点,有待改进,而粗糙集理论为空间数据属性的分析另辟一条蹊径[3]。粗糙集的理论最早由波兰科学家 Pawlak[1] 提出,用来描述某个集合中的特定对象的确切属性,可能没有精确的定义。由于它能有效分析不精确、不完整和不一致等不完备的信息,因此这种智能计算理论的优点可能应用在聚类算法中。

随着数据库中案例的不断扩充,其信息呈现爆炸式增长。如何从大量案例中正确、高效地检索出最相似的案例,已成为案例检索领域研究的热点之一。本文介绍了相关的理论背景知识和关键概念,提出了一种基于粗糙集的 K-means 聚类算法,并对已有的系统进行了改进。采用本文聚类算法能有效区分营销案例库中的不同类型,用户在检索时,系统不必对案例库中的每个案例进行相似度计算,从而大大提高了系统的检索效率。实验证明,本文方法大大提高了系统的检索性能。

2 基于粗糙集的 K-means 聚类算法

2.1 粗糙集理论

一般说来,Zadeh 的模糊集理论以及后来的直觉模糊集合(Intuitionistic Fuzzy Sets, IFS)[4] 对经典集合进行了有效扩充,是基于集合的模糊定义,可以扩展描述外延不分明的模糊概念。它增加了非隶属度函数属性参数,细腻地刻画了客观

到稿日期:2010-01-15 返修日期:2010-04-06 本文受国家自然科学基金项目(70371054,70771077),国家高技术 863 研究发展计划(2008AA04Z106),上海市科委制造业信息化专项基金(08DZ1122303)资助。

陈 千(1983—),男,博士生,主要研究方向为语义网与本体论、数据挖掘、图像视频检索,E-mail:521mike@tongji.edu.cn;向 阳(1962—),男,博士,教授,博士生导师,主要研究方向为决策支持系统、人工智能;郭 鑫(1982—),女,博士生,主要研究方向为数据仓库、数据挖掘;王 栋(1981—),男,博士生,主要研究方向为语义网、人工智能、数据挖掘。

世界的模糊性本质。但是模糊集往往需要模糊隶属函数，而这个先验知识一般很难得到。粗糙集是建立在分类的基础上，不需要附加信息或先验知识，这一点是其它方法无法做到的。它将分类理解为在特定空间上的等价关系，而等价关系则构成对该空间的分类。粗糙集将知识理解为对数据的划分，将机器学习中的学习理解为对数据的分类，且每个分类后的集合称为概念。知识库可形式地表示成 $K=(U,R)$，其中 U 为有限的非空集合，即论域；R 为一组等价关系，对于 $\forall M\in R$，由 M 所产生的不可分辨关系，或称之为等价关系即为 $IND(M)$，$IND(M)=\bigcap_{R\in M}R$，它表达了系统利用知识库中的一部分知识 M 所能达到的最高认识程度。

集合的上下近似和边界：从外延的角度考查一个概念，近似空间 $apr=(U,R)$ 提供了可分辨、可表达的空间，其中等价类 $[u]_R$ 构成了最小的分辨单位。用颗粒状的基本集去表达一个集合或概念，则形成了在近似空间中集合的上下近似和边界概念。从拓扑的角度看，集合的上近似是包含了该集的最小精确集，是个闭包；而下近似则为包含于该集的最大精确集，为该集的开核。下近似由必然属于该集的元素构成，而上近似由可能属于该集的元素构成。边界中的元素则不能判断是否属于该集[3]。

2.2 基于粗糙集的 K-means 聚类算法

案例检索中最大的问题是，在实际应用领域中，案例库中的案例数量会随着时间的推移和应用不断增长，这时就构成了检索效率的瓶颈，大大延长了系统的响应时间。本文利用基于粗糙集理论的 K-mean 聚类算法将案例库划分成若干个小的案例集合，使得在同一个案例集合中的两个案例相比较于其他的案例集合更加相似。将聚类后的案例集合的聚类中心作为该案例集合的索引，从而建立海量案例库索引，这样不仅能大大提高案例的相似度的计算，而且能大大缩短检索的响应时间。

聚类分析算法是统计学的一个分支，属于非监督的机器学习范畴。典型的聚类算法，如 KNN 算法是属于一种懒惰的机器学习算法。由于 K-means 聚类算法简单易懂，且基于欧氏距离，参照 Lingras 等人在 Rough K-means 算法的基础上提出了适用于文本案例检索的基于粗糙集 K-means 聚类算法。在这之前先做一下定义，如图 1 所示。

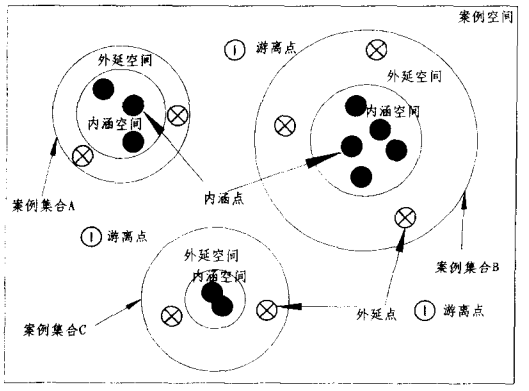

图 1 案例空间示意图

定义 1 整个案例所处的空间称为案例空间。

定义 2 案例空间中经过聚类后的案例库有 K 个聚类，每个聚类中的案例彼此间是最相似的，称聚类为案例集合。

定义 3 案例集合中的中心叫案例集合中心。

定义 4 案例集合中与中心的距离小于某个阈值 w 的集合称为该案例集合的内涵空间，反之大于 w 的集合称为该案例集合的外延空间。而所有在该集合的内涵空间的案例称为内涵点，所有该集合的外延空间的案例称为外延点，这个阈值 w 就称为凝聚系数。

定义 5 还有一些案例与所有的案例集合的距离大于某个阈值 d，这些案例称为游离点，该阈值 d 称为游离系数。一般地，对于某个案例集合而言，$w<d$。

对于论域 U 而言，指的是案例空间，则对于 $U/P=\{X_1,X_2,\cdots,X_k\}$，可以看成是基于等价关系 P 的 U 的集合，其中 X_i 即为第 i 个案例集合，这里 $1\leqslant i\leqslant k$。由于案例缺乏足够的知识，因此不可能很精确地定义每个集合 X_i。引入上下近似集 $\overline{A}(X)$，$\underline{A}(X)$，分别对应案例集合 X 的外延空间 W 和内延空间 N。因此有 $\overline{A}(X)=(W)+(N)$，$\underline{A}(X)=(W)$。

2.3 算法步骤

基于粗糙集理论的 K-means 均值聚类算法的步骤如下：

Step 1 在案例空间中选取 K 个点，这些点代表着要聚类的案例，也是初始案例集合的中心。

Step 2 将案例空间中的每个案例同这 K 个案例集合的中心进行相似度计算，并把它们归并到 K 个集合中心最近的案例集合中。

Step 3 完成集合的赋值之后，重新计算赋值后的 K 个案例集合的中心。

Step 4 重复 Step 2，Step 3，直到中心不再变化和移动为止。

在这个算法中，中心其实也是一个 m 维向量，在结构上和案例 x_i 是一样的。此中心的计算方法如规则 1。

规则 1[3] 若 $\underline{A}(X)\neq\emptyset$，且 $\overline{A}(X)-\underline{A}(X)=\emptyset$，则对于每个案例向量 $x=(x_1,x_2,\cdots,x_j,\cdots,x_m)$，

$$x_j=\sum_{v_j\in\underline{A}()}v_j/|\underline{A}(x)|$$

否则，若 $\underline{A}(X)=\emptyset$ 且 $\overline{A}(X)-\underline{A}(X)\neq\emptyset$，则

$$x_j=\sum_{v_j\in(\overline{A}()-\underline{A}())}v_j/|\overline{A}(x)-\underline{A}(x)|$$

否则，$x_j=w_l\dfrac{\sum_{v_j\in(\underline{A}())}v_j}{|\underline{A}(x)|}+\dfrac{w_u\sum_{v_j\in(\overline{A}()-\underline{A}())}v_j}{|\overline{A}(x)-\underline{A}(x)|}$

式中，w_l 和 w_u 对应着下上近似的重要程度，且 $w_l+w_u=1$；$1\leqslant i\leqslant m$，m 为向量的维数。从另外一个角度说，若 $\overline{A}(X)-\underline{A}(X)=\emptyset$，也就是上近似等于下近似，则退化为普通的聚类算法。

3 基于营销本体的案例检索

本体在知识表达和语义处理方面发挥着重要的作用。它具有多种形式化定义，这里采用 OWL 的子语言 OWL Lite 就足够用来描述营销案例本体。本体的主要元素包括概念(Concept)、属性(Property)、关系(Relation)、实例(Instance)以及公理(axiom)。其中公理主要用于逻辑推理。营销案例表示的是案例检索的前提，也是案例的相似度计算的前提。我们首先将营销案例用向量空间模型来表达，接下来以知网(Hownet)的概念之间的相似度的计算接口为基础来进行概念相似度的计算。知网是以汉语和英语的词语所代表的概念为描述对象，是以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库参考，它着力反映了概念的共性和个性以及概念之间和概念的属性之间的各种关系[5]。基于这种网状的知识系统，Hownet 提供了具体的概念之间的相似度的计算接口。将营销案例用本体来表示，而概念之间的相似度采用知网中的相似度计算方法来计算。

3.1 基于营销本体的案例表示

以营销领域本体为前提，基于本体的案例表示方法将中文文本案例表示为独立的本体，将自然语言处理问题转为对本体的相似度计算，从而方便案例的相似度计算。假设营销领域本体已经构建完成，则案例的表示具体过程如下[6]：

Step 1　使用中科院的 ICTCLAS 中文分词算法接口对案例文档进行切词。

Step 2　在 Step1 的基础上，依据营销本体的概念集，保留匹配的词汇，从而形成概念集合，并附上概念的词频构造该案例文档的向量。

Step 3　计算概念向量中每个分量对该案例的重要程度。

Step 4　根据阈值 μ，用营销本体对向量进行扩充。当 $\mu>5$ 时，则将该概念的父类、子孙类扩充到概念向量中；否则，不对概念进行扩充。

Step 5　扩充后的概念向量中的每个概念，参照营销本体中的属性、关系和实例等元素从中抽取出对应的属性、实例、关系和公理。

Step 6　标准化抽取出概念、属性、实例、关系和公理，以形成一个独立的本体。

根据以上 6 个步骤，一个营销案例 CT 就可以完全用基于本体的 $4\times m$ 的矩阵表示了，也即 $CT=\{C,I,R,P\}$。其中 $C=\{f_1,f_2,\cdots,f_m\}$ 为 CT 中概念 Ci 的词频向量；$I=\{i_1,i_2,\cdots,i_n\}$ 则为实例个数构成的实例向量；$R=\{r_1,r_2,\cdots,r_k\}$，$P=\{p_1,p_2,\cdots,p_j\}$ 分别为关系向量和属性向量。

3.2 相似度计算

概念、属性、实例和关系都是以概念为中心的。除了概念的相似度的计算之外，其他 3 个方面都是以 C 为依据，因而都是平等的。概念的相似度是通过 Hownet 提供的具体的概念之间的相似度的计算接口来进行计算的，针对属性、实例和关系这 3 个方面按照如下步骤进行计算：

对于向量 e_i 和 e_j，可以表示为 $e_i=(e_{i1},e_{i2},\cdots,e_{in})$ 和 $e_j=(e_{j1},e_{j2},\cdots,e_{jn})$ 的集合，则 e_i 和 e_j 的相似度计算方法如下：

1)把两个集合中的元素进行任意配对，计算出所有可能的配对元素的相似度。

2)取出相似度最大的一对，记为 $\mathrm{sim}^1(e_i,e_j)$。

3)在剩余的配对元素的相似度中，再取出相似度最大的一对，记为 $\mathrm{sim}^2(e_i,e_j)$。如此反复，直到所有元素完成配对。

4)采用式(1)计算最终的相似度：

$$\mathrm{Sim}(e_i,e_j)=\frac{\sum_{k=1}^{n}\mathrm{Sim}^k(e_i,e_j)}{n\sum_{k=1}^{n}\mathrm{Sim}^k(e_i,e_j)} \tag{1}$$

将案例用营销本体来表示，因此两个案例的相似度实际上就是两个本体的相似度。若 $\mathrm{sim}_C(D_i,D_j)$，$\mathrm{sim}_I(D_i,D_j)$，$\mathrm{sim}_R(D_i,D_j)$，$\mathrm{sim}_P(D_i,D_j)$ 分别表示案例文档 D_i 和 D_j 在概念、属性、实例和关系这 4 个方面的相似度，则 D_i 和 D_j 相似度计算方法如式(2)所示：

$$\mathrm{sim}(D_i,D_j)=w_1*\mathrm{sim}_C(D_i,D_j)+w_2*\mathrm{sim}_I(D_i,D_j)+w_3*\mathrm{sim}_R(D_i,D_j)+w_4*\mathrm{sim}_P(D_i,D_j) \tag{2}$$

式中，w_i 为权重，$w_i>0$，$w_1+w_2+w_3+w_4=1$。

3.3 案例检索系统

基于营销本体的案例检索系统依据一个相对完善的领域本体，将中文文本案例表示为小本体，并基于本体的案例相似度计算方法计算文本案例的相似度；将复杂的纯文本语义处理转换为本体相似度计算，回避了在中文文本语义处理中所带来的诸多歧义问题。通过充分考虑本体组成元素(概念、属性、实例)之间的相关性以及各组成元素本身的语义相似度，给出一种综合的本体相似度计算方法，从而达到计算两个案例相似度的目的。基于营销本体的案例检索系统框架图如图 2 所示。

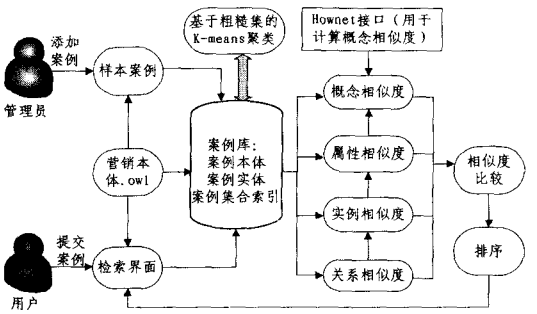

图 2　基于营销本体的案例检索框架

系统涉及到 3 种类型的案例，分别是样本案例，即管理员往案例库中将要添加的案例；任务案例，即用户提交的检索任务中的范文案例，用于表达用户的检索需求；标准案例，即已经添加到案例库中、供用户检索的大型案例库，它是整个本文论述的重点和中心。案例库由 3 个部分组成：案例(原始案例，包括样本案例、任务案例和标准案例)、案例本体(包括所有样本案例和提交案例以及案例库中所有案例的本体表示)。和例索引(本质上，是针对标准案例集合建立的语义索引文件)系统分为 3 个部分功能，分别为案例库聚类、案例添加和案例检索。通过采用基于粗糙集的 K-means 聚类算法，对案例库建立以 K 个案例中心为内容的索引，一并存在数据库中。系统是以 Hownet 提供的两个概念之间的相似度的计算接口为基础的。由于 Hownet 考虑到了上下位语义关系，因此检索的结果从一定程度上符合用户的要求。

4　系统架构设计

在用户进行案例检索之前，案例库管理员已经通过聚类算法对案例库中的案例进行了整理和分类。在实际检索中，用户提交样本案例后，系统将样本案例通过营销本体来表示，并写入到数据库中，这些案例在数据库中的结构如图 3 所示。

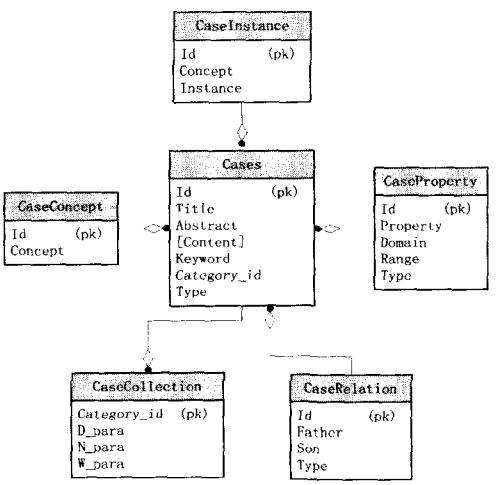

图 3　本体表示数据库关系图

当管理员往案例库中添加样本案例时，系统根据营销本

体将样本案例表示成本体的形式，并计算与案例库中的 k 个案例集合中心的距离，把该案例添加到相似度最大的一个案例集合中。当终端用户需要检索时，同样需要将用户提交的任务案例表示成本体，并且与案例库中的标准案例的案例集合中心进行相似度计算，并在返回的最相似度案例集合中继续进行单个案例匹配。匹配的相似度计算方法同上节中的一样涉及到以概念为中心的概念相似度、属性相似度、实例相似度和关系相似度。最终将案例按照相似度从大到小的顺序返回给用户。案例库中的标准案例在数据库表中全部使用本体来表示，其中 *CaseCollection* 表的数据就是 K 个聚类案例集合的详细信息，包括游离系数 d_para、凝聚系数n_para。

5 实验结果分析

本系统在内存为 1G 的戴尔 Inspiron6400 笔记本上运行，数据库服务器安装在 Dell 专用服务器上，并且通过局域网链接，采用的是 Eclipse3.5 的 Java 开发环境。我们采用公司里的营销案例库作为实验的样本案例库，将其中的 30%作为测试案例库。检索图如图 4 所示。由于将系统设置中的相似度限制为 10%，因此检索出来的案例与提交的范例案例相似度大于 0.10。

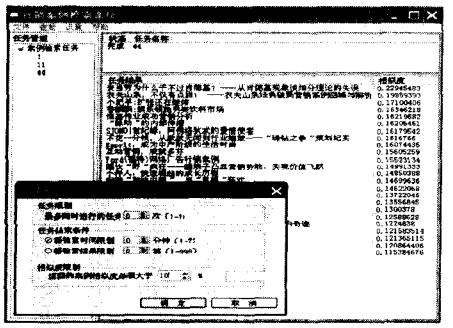

图 4 系统检索界面图

文本检索的两个评测因子通常为查准率 p 和查全率 r，见式(3)、式(4)[7]。

$$p=\frac{|\{rc\}\cap\{cr\}|}{|\{rc\}|} \tag{3}$$

$$r=\frac{|\{rc\}\cap\{cr\}|}{|\{cr\}|} \tag{4}$$

式中，$\{rc\}$为整个测试集中实际与某案例文档相匹配的文档的集合，$\{cr\}$为检索结果集。p 表示检索结果中实际与某案例文档相匹配的个数与整个测试集中实际与某案例文档相匹配的个数的比率，也即匹配的准确率(precision)。r 表示检索结果中实际与某案例文档相匹配的个数与检索结果中的案例文档个数的比率，也即查全率(recall)。本案例检索系统使用及不使用 K-means 聚类算法的检索效果如表 1 所列，其中 R. 表示系统响应时间，其单位为 s(秒)，Recall 表示查全率，Prec. 为准确率。

表 1 实验结果数据对比

	R.(s)	Recall	Prec.	R.(s)	Recall	Prec.
No	25.486	97.313	87.533	54.155	96.654	84.341
Yes	3.237	92.527	89.380	5.176	92.618	86.127
Scale	1000 级别			10000 级别		

从表 1 中可以看出，基于粗糙集的 K-means 在营销案例检索系统中以系统的检索响应时间、检索的准确率和查全率为检索目标，能够大大提高检索的效率，但会对查全率有一定的影响。不过，相比较于响应时间上的突破，这个代价还是值得的。

结束语 本文主要探讨了聚类方法在案例检索中的应用研究，详细地介绍了将聚类方法应用于案例检索系统中，并作了有益的探索，为案例库的检索提供了一条有效的途径。基于粗糙集的 K-means 在营销案例检索系统中以系统的检索响应时间、检索的准确率和查全率为检索目标，能够大大提升用户的满意度。但是 K 的取值和初始点的确定还是个难点，若取值不当反而会影响聚类的效果。因此，以后的研究主要集中在如何确定 K 值和初始点的选定问题上。

参 考 文 献

[1] Pawlak Z. Rough set theory and its application to data analysis [J]. Cybernetics and Systems, 1998, 29(9): 661-668

[2] Raymond T N, Han Jiawei. CLARANS: A Method for Clustering Objects for Spatial Data Mining [J]. IEEE Transaction on Knowledge and Data Engineering, 2002, 14(5): 1003-1016

[3] Pawan L. Applications of Rough Set Based K-Means, Kohonen SOM, GA Clustering [J]. Transactions on Rough Sets VII, 2007: 120-139

[4] Dubois D, Gottwald S, Hajek P, et al. Terminological Difficulties in Fuzzy Set Theory-The Case of Intuitionistic Fuzzy Sets[J]. Fuzzy Sets and Systems, 2005, 156: 485-491

[5] 董振东. 知网主页[OL]. http://www.keenage.com, 1999: 1-15

[6] Wang Dong, Xiang Yang, Zou Guobing, et al. Research on Ontology-Based Case Indexing in CBR[C]// International Conference on Artificial Intelligence and Computational Intelligence. US, Springer, 2009: 238-242

[7] Wikipedia Home page [OL]. http://en.wikipedia.org/wiki/Precision_and_recall, 2000

(上接第 113 页)

[4] Pande H, Landi W, Ryder B G. Interprocedural Def-Use Associations for C Systems with Single Level Pointers [J]. IEEE Trans. Software Engineering, 1994, 20(5): 385-403

[5] Harrold M J, Rothermel G. Performing Data Flow Testing on Classes[C]// Second ACM SIGSOFT Symposium on the Foundations of Software Engineering. Dec. 1994: 54-163

[6] Yang J, Huang J, Wang F, et al. An Objected-oriented Architecture Supporting Web Appliction Testing[C]// Proc. of Computer Software and Applications Conference. 1999

[7] Ricca F, Tonella P. Analysis and testing of Web application[J]. IEEE Computer, July 2001

[8] Liu Chien-hung, Kung D C, Hsia P. Structural testing of Web application [J]. IEEE Computer, March 2000

[9] Kung D C, Liu C H. An Object-oriented Web test model for testing Web application[J]. IEEE Computer, January 2002