一种基于粗集的文本数据特征信息的挖掘方法

A Mining Method for Characteristic Information in Text Data Based on Rough Sets

易树鸿　张为群

(遵义师范学院计算机科学系　贵州遵义563002)(西南师范大学计算机与信息科学学院　重庆400715)

Abstract In this paper, we provide a new mining method on text data. Aiming at the mining for characteristic information in text data, we have the aid of knowledge representation based on rough sets to make concept frame, and advance quantitative index.

Keywords Text data, Characteristic information, Mining method, Rough sets, Concept frame

1. 引言

随着 Internet 的飞速发展，人们的信息交流越来越多地依赖于网络，人们在网上发表自己的意见和见解、相互讨论各种问题、交流情感和思想。在网上传输的这些数据中，大量涉及到的是文本数据，网络应用的普及使得文本数据呈现出高速膨胀的态势，面对浩瀚的文本大海，人们迫切需要快速、准确地从需要的文本数据中了解其观点、思想或热点问题等等。例如，在远程教育中，教师可能面对的是成百上千个学生，教师希望快速地从学生的讨论和交谈中寻找学生集中关心的问题，以便及时回答和调整教学。又如，出于国家安全的考虑，需要对类似于 BBS 的公众论坛的文档进行鉴别，以便进行有效地监督和管理。以上问题所涉及的都需要高效、快捷地对文本数据进行特定的信息挖掘。

2. 方法概述

对于文本数据的挖掘，人们已经提出了多种方法，有基于关键词匹配的，有基于扩展短语描述的[6]，有基于潜在语义分析的[7]。本文将另辟途径，给出一种新的挖掘方法，为叙述方便，我们把文本数据中所反映的观点、思想、热点问题等统称为文本数据的特征信息，上面所叙述的问题就是挖掘文本数据中的特征信息。我们提出的方法是，构建一个概念框架，用它来对需要的文本数据进行表示，从中筛选出所包含的特征信息。现在的关键是，用什么方式来表示文本数据，使得能够清晰地呈现其特征信息？近年来，随着粗集理论的不断发展，人们提出了知识的粗集表示方法[4,5]，利用粗集的特性，可以把知识的概念特征加以明确的体现，这为我们的研究提出了全新的方法。有关粗集的知识，感兴趣的读者可查阅相关的文献[1~3]。

3. 概念框架的建立

概念框架用来对采集的文本数据进行表示，它是文本数据挖掘的工具，我们对它既有属性方面的要求，又有量方面的刻画，整个框架由三个层面构成。

3.1 鉴别词料库(或短语库)

首先我们需要建立相应的词料库(或短语库)，它是相应领域的专业词汇的的集合，称为鉴别词料库，记为 U。

鉴别词料库是概念框架的基础，为挖掘工作划定范围和界限。

3.2 特征概念空间

有了鉴别词料库，我们需要根据所感兴趣的特征和明确的特征关系来对鉴别词料库做结构化处理。

定义1 设 $U=\{u_1,u_2,\cdots,u_n\}$ 是鉴别词料库，$T=\{\tau_1,\tau_2,\cdots,\tau_m\}$ 是特征集，称 $[\tau_i]=\{u\in U;(u,\tau_i)\in\Gamma\}$ 为 τ_i 属性集，其中 Γ 是特征关系。对 m 个属性集 $[\tau_1],[\tau_2],\cdots,[\tau_m]$，构造如下的集合：

$\alpha_0=\overline{[\tau_1]}\overline{[\tau_2]}\Lambda\overline{[\tau_m]}$，

$\alpha_1=[\tau_1]\overline{[\tau_2]}\Lambda\overline{[\tau_m]}$，

$\cdots\cdots$

$\alpha_N=[\tau_1][\tau_2]\Lambda[\tau_m]$，

其中，$\overline{[\tau_i]}=U-[\tau_i]$，$N=2^m-1$。

定义2 称 $\alpha_0,\alpha_1,\cdots,\alpha_N$ 为特征概念，所有特征概念的集合称为特征概念空间，记为 C，即 $C=\{\alpha_0,\alpha_1,\cdots,\alpha_N\}$。

显然，C 构成了 U 的一个划分，也就是说，我们可以依据所关心的特征问题，将鉴别词料库 U 划分成一些等价类，使得 U 具有一定的结构。这样，就可以用它来对需要的文本进行知识匹配处理，以寻找其规律性的知识，这就是特征信息挖掘的工作。

3.3 特征分布列

为了从量上把握特征概念在 U 上的分布情况，以便对挖掘的信息作出客观评价，我们引进下面的概念。

定义3 设 $U=\{u_1,u_2,\cdots,u_n\}$ 是鉴别词料库，$C=\{\alpha_0,\alpha_1,\cdots,\alpha_N\}$ 是定义在其上的特征概念空间，对任意 $\alpha_i\in C$，称 $f_i=|\alpha_i|/n$ 为 α_i 在 U 上的分布率。

定义4 称 F 为特征概念空间 C 在 U 上的分布列：

$$F=\begin{pmatrix}\alpha_0 & \alpha_1 & \Lambda & \alpha_N\\ f_0 & f_1 & \Lambda & f_N\end{pmatrix}$$

由特征概念空间 C 的构造和分布率的定义，不难发现如下事实。

定理1 设 $U=\{u_1,u_2,\cdots,u_n\}$ 是鉴别词料库，$C=\{\alpha_0,\alpha_1,\cdots,\alpha_N\}$ 是定义在其上的特征概念空间，F 为 C 在 U 上的分布列，则有：(1)$0\leqslant f_i\leqslant 1$，对任意 $\alpha_i\in C$；(2)$\sum_{\alpha_i\in C}f_i=1$。

定理告诉我们，分布列具有良好的数学性质，这为我们的挖掘工作奠定了坚实的基础。

3.4 概念框架

定义5 称三元组(U,C,F)为文本数据的概念框架。

至此，我们建立了文本数据特征信息的挖掘框架，我们可以根据需要，调整 U 和 C 的类别和属性，做成一个 Agent，让它在网络上帮助我们快速提取所关心的文本数据的特征信

息。

4. 文本数据特征信息的挖掘与评价

有了概念框架，怎样进行文本数据特征信息的提取呢?具体方法是，用知识表示来挖掘相关的特征信息，通过累计分布率来评价挖掘的有效性，下面分别叙述如下。

4.1 文本数据特征信息的挖掘

给定一个文本 A，为了挖掘 A 的特征信息，我们把 A 和特征概念空间进行匹配，即用特征概念空间 C 来表示 A。显然，一般情况下是不可能完全匹配的，只能是粗糙地匹配。为此，我们引进知识的粗集表示方法，详细内容可在文[4,5]中找到，这里我们直接引用表示结果。记 A_- 和 A^- 为 A 的下近似和上近似，即

$$A_- = \underset{\alpha_i \subset A}{Y} \alpha_i, A^- = \underset{\alpha_i \cap A \neq \Phi}{Y} \alpha_i$$

从以上表示可以看出，A_- 是文本 A 中所包含的特征概念，而 A^- 则是文本 A 所涉及到的特征概念。这正是我们要的结果，为此，我们定义如下。

定义6 在概念框架(U,C,F)下，称 $A_- = \underset{\alpha_i \subset A}{Y} \alpha_i$ 为文本 A 的 C 集中特征，$A^- = \underset{\alpha_i \cap A \neq \Phi}{Y} \alpha_i$ 为文本 A 的 C 关联特征。

集中特征和关联特征就是我们利用概念框架(U,C,F)所挖掘的文本数据的特征信息。对于这两种特征信息，在不同的应用场合往往有所侧重，有时候我们想确切地了解某个文本到底包含了多少种不同的观点，这时可以注重于集中特征，而有的时候我们只需大致了解该文本涉及到哪些方面的问题，那么用关联特征就比较适合。当然，一般场合应该把两者结合起来考虑，才能更全面透彻地了解所需要的特征信息。

4.2 特征信息的评价

通过特征概念空间 C，我们得到了某个文本 A 的集中特征 A_- 和关联特征 A^-，但是它们到底能在多大程度上反映 A 的 C 特征，我们还需要进一步引进数量指标来衡量。

定义7 称 $J(A) = \sum_{\alpha_i \in A} f_i$ 为 A 的 C 特征集中度，$G(A) = \sum_{\alpha_i \cap A \neq \Phi} f_i$ 为 A 的 C 特征关联度。

用它们可以定量地刻划概念空间 C 和文本 A 的匹配程度，从而给出用 C 来提取文本 A 的特征信息的数量评价。一般情况下，我们希望 J(A)和 G(A)都应达到一定的阈值。J(A)的值可以反映 A—作为文本 A 的 C 特征信息的显著程度，如果 J(A)太小，那么 A_- 并不能集中代表 A 的 C 特征信息。G(A)的大小则在一定程度上给出了挖掘工作是否有效的评价，下面是几种典型的评价：

(1)$G(A)<0.3$，称 A 和 C 不关联。从数量上讲，用 C 的概念来表示 A，其相关成分不足30%，这意味着 A 具有的 C 特征很少，因此挖掘的 C 特征没有多少实际价值；

(2)$0.3 \leqslant G(A)<0.55$，称 A 和 C 为低关联。此时 A 只具有50%左右的 C 特征，因此挖掘的特征信息只能作为参考；

(3)$0.55 \leqslant G(A)<0.85$，称 A 和 C 为基本关联。这种情况下，挖掘的 C 特征占到了60%左右，所以挖掘的特征信息有一定的参考价值；

(4)$0.85 \leqslant G(A)$，称 A 和 C 为高度关联。由于相关的 C 特征含量已在85%以上，故挖掘的特征信息是有效的，特别地，如果 $G(A)=1$，则称 A 和 C 完全关联，这是最理想的情况。

在实际问题中，除了整体评价特征信息外，有时还需要对涉及的特征概念进行比较分析，看看谁更重要，这时可以利用特征概念 α_i 的分布率 f_i 来实现，f_i 大的相对来说就更重要一点。特别地，我们经常需要知道某个文本中的核心概念，借助于分布率，可以给出如下定义。

定义8 记 $f_k = \max_{\alpha_i \subset A}\{f_i\}$，则称 α_k 为 A 的核心概念。

核心概念在实际问题中有不同的表现形式，可以是某篇文档的中心论点、也可能是某个讨论组的热点问题。

4.3 多框架挖掘

对于文本数据特征信息的挖掘通常有两种用意，一是考察某个文本是否有我们所关心的特征信息，二是从文本数据中分析并提取最能反映该文本特征的特征信息。对于第一种情形，通过计算 C 特征关联度就可以帮助我们做出判断，但是对于第二种情形，上面的评价结果却意义不大，因为即使是 C 特征关联度很小，那也只能说明文本 A 并不具有 C 特征，而这并不是我们所要的结果。为此，我们引进多框架挖掘的方法来解决第二种情形的问题。

对鉴别词料库 U，我们可以构造多个挖掘框架(U,C_1,F_1)、(U,C_2,F_2)、…，然后用它们来挖掘同一文本 A 的特征信息，得到 $A_{1\_}$、$A_{2\_}$、…、$A_{\bar{1}}$、$A_{\bar{2}}$、…，；接着计算 $J_1(A)$、$J_2(A)$、…，$G_1(A)$、$G_2(A)$、…，通过数量比较，可以找到 A 的最佳特征信息——即最能反映文本 A 的特征信息。

小结 本文通过构造概念框架，借助于粗集的方式，实现了对文本数据特征信息的挖掘。借助于我们提供的方法，不仅可以完成内容的处理，还给出了数量上的评价，使得我们对挖掘的信息做到"心中有数"。利用粗集的方法来进行文本数据的挖掘，其过程和评价比较客观和自然，希望引起同行的注意。

参 考 文 献

1 Pawlak Z. Rough set. International Journal of Information and Computer Science, 1982(11):341～356

2 Pawlak Z. Rough set-theoretical aspects of reasoning about data. Dordrecht:Kluwer Academic Publishers,1991

3 曾黄麟. 粗集理论及其应用. 重庆大学出版社,1998

4 胡涛,吕炳朝,等. 基于粗糙集的不确定知识表示方法. 计算机科学,2000,27(3):90～92

5 易树鸿,张为群. 基于粗集的不确定性知识表示的精确性和关联性的研究. 计算机科学,2001,28(9):91～94

6 张磊,杜小勇,等. 文本数据库中的扩展短语挖掘. 第18届全国数据库学术会议论文集,计算机科学,2001,28(8,增刊):154～158

7 刘昌钰,郭颖,等. 基于潜在语义分析与 Bayes 分类的 BBS 文档鉴别. 第18届全国数据库学术会议论文集,计算机科学,2001,28(8,增刊):179～183