基于概念格标尺的广义匹配研究*)

智东杰[1] 智慧来[2] 刘宗田[2]
(河南理工大学计算机科学与技术学院 焦作 454150)[1]
(上海大学计算机工程与科学学院 上海 200072)[2]

摘 要 为了提高计算机对自然语言的理解能力,实现概念的语义匹配,提出了基于概念格标尺的广义匹配。对现有的模式匹配的概念进行拓宽,利用概念格特有的层次关系建立语义标尺,通过标尺的同构来实现语义的匹配,同时分析了标尺刻度不一致时的刻度转换。研究表明该方法符合人们对现实世界中事物的认识,能够提高计算机对自然语言处理的智能。

关键词 概念格,标尺,模式匹配,映射,语义距离

Augmented Pattern Matching Based on Concept Lattice Scale

ZHI Dong-jie[1] ZHI Hui-lai[2] LIU Zong-tian[2]
(Dept. of Computer Science and Technology, Polytechnic University, Jiaozuo 454150, China)[1]
(Dept. of Computer Engineering and Science, Shanghai University, Shanghai 200072, China)[2]

Abstract In order to improve computers' understanding ability of natural language, semantic pattern matching based on concept lattice scale was promoted. This method is seen as an augmentation of traditional pattern matching. It first utilized concept lattice's own hierarchical structure establishing semantic scale, then used isomorphs between two scales to realize semantic pattern matching of two different concepts. Graduation conversion between different scales was also studied. The research shows that our method confirms the way people look at the world, and can be realized by the computer to improve natural language processing ability.

Keywords Concept lattice, Scale, Pattern matching, Mapping, Semantic distance

文献[1]是关于模式匹配问题的一篇非常重要的综述性文章。在该文中,作者介绍了模式匹配问题研究的起源和应用领域,并且详细综述了现有的实现方法和各种原型系统。文献[2]提出基于泛代数理论的观点,模式是一类有限的结构,模式匹配能够被形式化为寻找两个结构之间保持的映射关系,即为两个结构之间的同态。

语义匹配可以看作是一类拓宽了的模式匹配。在现实世界里有许多语义匹配的例子,两个概念分别属于不同的领域,有着完全不同的属性,但是在语义下它们是匹配的。例如,工作的表现分为很好、好、一般和差等,对应的就有不同的薪水和奖励。再如,随着气温的变化,人们的着装也随着改变,从衬衣、茄克、毛衣到厚厚的棉衣。再例如,在安静的环境里人们的动作也相对轻柔,在马路上走路的动作和声音要大些,在运动场上的动作就最为激烈。工作表现的优良与薪水的高低,温度和着装,以及环境的安静程度和走路动作的轻重都分别属于不同的领域,描述不同的事或物,但它们之间却存在着自然的"匹配"。

1 基于标尺的广义匹配原理

从现实的大量例子可以看出,两个事物或概念分别属于不同的领域,有着完全不同的意义和性质,但是事物或概念自身在语义程度上却呈现出从高到低、由强到弱的层次关系,正是在这个层次关系上,不同的事物和概念实现了匹配。

这里引入记号 $*$,设Σ是一个有穷字母表,用Σ^*表示Σ上的所有字的全体,空字 ε 也包括在其中。称Σ^*是Σ的闭包。

定义 1 如果 $c_1,c_2,c_3,\cdots,c_n$ 是一组概念,这组概念形式上可以表示为$(+)^* c$,那么这组概念 $c_1,c_2,c_3,\cdots,c_n$ 是语义相关的。$+$代表语义的增强,c 代表基本概念。

定义 2 如果 c_1,c_2 可以表示为 $c_2=(+)^* c_1$,那么 c_2 的语义强过 c_1 的语义,记作 $c_2>c_1$。

定义 3 如果一组概念语义相关的概念 $C=\{c_1,c_2,c_3,\cdots,c_n\}$包含描述最小概念和最大概念,按照语义的递减排序后的序列 $D=\{d_1,d_2,d_3,\cdots,d_n\}$,且任意两个相临的概念差值不大于一个$+$,则称此序列是完备的。

序列的完备性是基于以下的考虑:其一,现实世界中的事物和现象是连续的,且在语义上具有最大概念和最小的概念;其二,概念均匀分布且具有最大值和最小值,那么建立的相应的标尺上的刻度是均匀的;其三,标尺上的刻度的均匀特性使得不同的标尺可以按照一定的比例方便地进行转换。

基本原理 一组概念语义相关完备序列 $C=\{c_1,c_2,c_3,\cdots,c_n\}$,按照语义的递减排序,排序后记为 $D=\{d_1,d_2,d_3,\cdots,d_n\}$;另一组概念语义相关完备序列 $C'=\{c_1',c_2',c_3',\cdots,c_{n'}'\}$,按照语义的递减排序,排序后记为 $D'=\{d_1',d_2',d_3',\cdots,d_{n'}'\}$。

如果 $n=n'$,那么概念 d_i 就和 d_i'在语义程度上匹配;

*) 基金项目:国家自然科学基金(项目批准号 60575035)。**智东杰** 工程师,主要研究领域为人工智能和符号计算;**智慧来** 博士生,主要研究领域为信息处理、概念格、本体;**刘宗田** 教授,博士生导师,主要研究领域为人工智能、软件工程和形式概念分析。

如果 $n\neq n'$，那么概念 d_i 就和区间 $d_{[(i-1)*n'/n]'}\sim d_{[i*n'/n]'}$ 中的任意一个概念在语义程度上匹配。

2 基于概念格顺序标尺的广义匹配实现方法

在实际应用基本原理进行匹配时，必须事前建立概念语义相关完备序列模型。例如要在 $C'=\{c_1',c_2',c_3',\cdots,c_{n'}'\}$ 寻找 c_i 和相匹配的概念，就必须对 c_i 和 c_j' 建立相应的序列模型，在这里引入了概念格来实现。

R. Wille[3] 提出的形式概念分析是以序理论和完备格理论为基础，依据数据库中提供的基本信息建立起的一种刻画对象与属性之间关系的数学结构。这种概念及概念层次的数学化使形式概念分析成为数据挖掘与知识发现的重要方法，并广泛应用于许多领域[4-7]。

定义 4 设 $K=(U,A,I)$ 是一个形式背景，$X\subseteq U,B\subseteq A$。如果 X,B 满足条件 $X'=B,B'=X$，则称序对 (X,B) 为形式背景 K 的一个概念。X 称为概念 (X,B) 的外延，B 称为概念 (X,B) 的内涵[3]。$L(U,A,I)$ 或 $L(K)$ 表示 K 中所有概念全体构成的集合，即

$$L(U,A,I)=\{(X\times B)\in U\times A;X'=B,B'=X\}$$

定义 5 设 $K=(U,A,I)$ 是一个形式背景，$(X_1,B_1),(X_2,B_2)\in L(K)$，如果 $X_1\subseteq X_2$ 或 $(B_2\subseteq B_1)$，称 (X_1,B_1) 是 (X_2,B_2) 的子概念，记为 $(X_1,B_1)\leqslant(X_2,B_2)$。显然 $L(K)$ 关于“$\leqslant$”构成一个格，称为概念格[3]。

2.1 建立概念的语义标尺

概念的语义程度可以用标尺来度量，选择术语“标尺”是为了强调和数学理论中度量的联系[8]。一组概念语义相关完备序列 $C=\{c_1,c_2,c_3,\cdots,c_n\}$ 经过量化表示为 $C=\{(+)^*c\}$，这样，序列 C 就与标尺建立起了入射关系。概念的属性值形成一个外延链，每个属性描述一个层，所有的概念形成一个概念标尺[9]。

例如，一组描述安静程度的概念{noisy, loud, very loud, extremely loud}，形成了具有 4 个层次的形式背景，如表 1 所示，对应着顺序标尺 04。

表 1 顺序标尺 04 的背景

	1	2	3	4
1	*	*	*	*
2		*	*	*
3			*	*
4				*

如果两个概念的差值小于设定值 δ，则映射到同一个概念格标尺节点，映射到同一概念格节点的多个概念形成一个等价关系。

一组语义相关概念到标尺的映射可以表示为图 1，左侧是语义相关概念的集合，右侧是概念格标尺，每一个概念映射到唯一的一个节点。c_i 和 c_j 对应相同的节点，表明 c_i 和 c_j 的语义是接近的，两者的距离小于一个刻度单位“+”，它们形成一个等价关系。

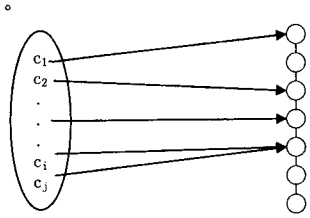

图 1 语义相关概念到概念格标尺的映射

2.2 标尺的同构

我们定义 $[x]\theta:=\{y\in V\mid x\theta y\}$，这是包含 x 的等价类。因子格具有序 $[x]\theta\leqslant[y]\theta:\Leftrightarrow x\theta(x\wedge y)(\Leftrightarrow(x\vee y)\theta y)$。

定理 1 如果 θ 是完全格 V 的一个完全同余关系，那么 $x\rightarrow[x]\theta$ 是从 V 到 V/θ 的一个完全同态。相逆地，如果 $\varphi:V_1\rightarrow V_2$ 是完全格之间的一个满射的完全同态，那么 $ker\varphi:=\{(x,y)\in V_1\times V_2\mid\varphi(x)=\varphi(y)\}$ 是 V_1 的一个完全同余关系；而且，$[x]ker\varphi\rightarrow\varphi(x)$ 是从 $V_1/ker\varphi$ 到 V_2 的一个同构[10]。

推论 1 $C=\{c_1,c_2,c_3,\cdots,c_n\}$ 映射到标尺 O_n，$C'=\{c'_1,c'_2,c'_3,\cdots,c'_{n'}\}$ 映射到标尺 $O_{n'}$，两个标尺的刻度可能相同，也可能不同，但它们两者是同构的。

这个推论是显而易见的，图 2 表现的就是两个标尺之间的同构。映射到同一节点的多个节点构成一个块。实际上标尺 2 是由标尺 1 上同余关系生成的因子格。

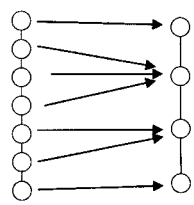

图 2 标尺的同构

2.3 概念的匹配

通过建立标尺，$C=\{c_1,c_2,c_3,\cdots,c_n\}$ 中的每一个概念 c_i 唯一地对应 $C'=\{c'_1,c'_2,c'_3,\cdots,c'_{n'}\}$ 中的一个块 $[c']\theta$，这个块可能只有一个概念，也可能包含多个概念。如果 $[c']\theta$ 只包含一个概念，那么 c_i 就和这个概念匹配；如果 $[c']\theta$ 包含多个概念，那么 c_i 就和这个块中的任意一个概念匹配。同理，$C'=\{c'_1,c'_2,c'_3,\cdots,c'_{n'}\}$ 中的每一个概念 c'_i 唯一地对应 $C=\{c_1,c_2,c_3,\cdots,c_n\}$ 中的一个块 $[c]\theta$。

3 实例

下面举一个简单的例子，说明基于语义匹配的实现。

例：这次考试小王很顺利，得了 80 多分，可以算得上()了。

A 优，B 良，C 中，D 差，E 很差

这个句子里有一个关键词“80 多分”，所有的分数按照分数差异，可以建立如表 2 所示的形式背景，这个背景对应一个双序标尺。

表 2 “分数”概念的形式背景

	>60	>80	>90	<59	<40
>60	*				
>80	*	*			
>90	*	*	*		
<59				*	
<40				*	*

同理，A 优，B 良，C 中，D 差，E 很差也对应一个形式背景如表 3 所示，这个背景对应一个双序标尺。

表 3 “评语”的形式背景

	C 中	B 良	A 优	D 差	E 很差
C 中	*				
B 良	*	*			
A 优	*	*	*		
D 差				*	
E 很差				*	*

(下转第 146 页)

lection、训练接口 ITrainer、分类接口 IClassifier、评估接口 IEvaluator 及 Instance 与 Instances 两个数据类为核心，其中 Instance 代表一篇文档，Instances 代表文档集。

4 实验结果与分析

文本分类从根本上说是一个映射过程，因此评价分类系统的标志是映射的准确程度和映射的速度。评价分类效果的标准很多，本系统采用准确率[8]和查全率[8]作为评价标准。

实验过程中，把海洋类信息网站上收集来的信息分为海洋调查与观测、区域海洋学、海洋基础科学、海洋资源与开发、海洋工程、海洋环境科学、潜水医学、军事海洋学 8 个大类，其中各大类又分成众多子类别。以海洋基础科学为例，分为：海洋水文学、海洋气象学、海洋物理学、海洋化学、海洋生物学、海洋地质学、海洋地貌学、海洋地球物理学。在实验初期，我们选择海洋基础科学信息这一大类共 240 篇文档，平均每类 30 篇文档。每一类选择训练集和测试集的方法如下：将这些分类好的数据平均分成 10 份，选择其中 1 份作为开放测试集，剩余的 9 份作为训练集和封闭测试集。经测试，封闭测试准确率平均达到 86.27%，查全率平均达 86.25%，具有较好的分类效果。表 1 给出有实际意义的开放测试结果。

表 1 分类结果

类别	人工归入	机器正确归入	机器实际归入	准确率 (%)	查全率 (%)
海洋水文学	30	26	29	89.66	86.67
海洋气象学	30	27	31	87.09	90.00
海洋物理学	30	25	30	83.33	83.33
海洋化学	30	28	31	90.32	93.33
海洋生物学	30	26	31	83.87	86.67
海洋地质学	30	23	28	82.14	76.67
海洋地貌学	30	28	33	84.85	93.33
海洋地球物理学	30	24	27	88.89	80.0

结束语 本文主要讨论基于海洋信息处理的文本分类系统的设计与实现，提出系统结构，详细介绍训练和分类的步骤及所使用的相关算法，提出一种改进 χ^2 特征选取算法，最后给出实验结果和分析。从实验结果看，系统具有较好的查全率和准确率。进一步的研究工作改进是分词和分类算法，尝试几种不同的分类算法，并比较其性能，提高分类准确度，完善分类评估体系。

参 考 文 献

[1] 刘宝银，张杰. 海洋科学的前沿——"数字海洋"[J]. 地球信息科学，2000(1)：8-10

[2] 苏新宁，杨建林，江念南，等. 数据仓库和数据挖掘[M]. 北京：清华大学出版社，2006

[3] 朱珣. 中文自动分词系统的研究[D]. 武汉：华中师范大学，2004

[4] 刘丽珍，宋瀚涛. 文本分类中的特征提取[J]. 计算机工程，2004(4)：14-16

[5] Yang Yiming, Liu Xin. A Re-Examination of Text Categorization Methods[J]// 22nd Annual International SIGIR. 1999：42-49

[6] Wu T-F, Lin C-J, Weng R C. Probability estimates for multi-class classification by pair wise coupling. J. of Machine Learning Research, 2004(5)：975-1005

[7] Zhang J, Yang Y. Robustness of regularized linear classification methods in text categorization// SIGIR'03. 2003：190-197

[8] 宋枫溪，高林. 文本分类器性能评估指标[J]. 计算机工程，2004(13)：107-110

(上接第 136 页)

两个背景对应的标尺如图 3。

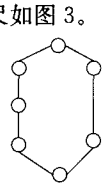

图 3 双序标尺

从标尺的同构关系可以直接得到"B 良"与关键词"80 多分"相匹配。

结束语 模式匹配不是一个刚性的构架，而是一个从实用主义出发考虑的思想体系，放宽模式匹配原有的某些约束是可以接受的，甚至是有利的。

基于概念格标尺的语义匹配是自然的，符合人们的思维方式，具有一定的理论基础，也便于计算机的实现，从而使计算机具有一定程度的智能，使之具有一定的解决实际的问题的能力。

需要进一步研究的内容和克服的困难在于如何量化多个语义相关概念间的距离，使得各个概念能较为精确地映射到一个标尺，从而使概念在语义上得到较好的匹配。

参 考 文 献

[1] Rahm E, Bernstein P A. A Survery of Approaches to Automatic Schema Matching. The International Journal on Very Large Data Bases[J], 2001, 10(4)：334-350

[2] 张治，车皓阳，施鹏飞. 模式匹配问题的描述框架与算法模型. 模式识别与人工智能[J], 2006, 12(6)：715-721

[3] Ganter B, Wille R. Formal Concept Analysis: Mathematical-Foundation[M]. New York: Springer-Verlag, 1999

[4] Dtintsch I, Gediga G. Algebraic aspects of attribute dependencies in information systems. Fundamenta Informaticae[J], 1997, 29：119-133

[5] Dtintsch I, Gediga G. Approximation operators in qualitative data analysis// de Swart H, Orlowska E, Schmidt G, et al., eds. Theory and Application of Relational Structures as Knowledge instruments[M]. Heidelberg: Springer, 2003：216-233

[6] Pagliani P. From concept lattices to approximation spaces: Algebraic structures of some spaces of partial objects. Fundamenta in Formaticae[J], 1993, 18(1)：1-25

[7] Yao Y Y. Concept lattices in rough set theory//Proceedings of 23rd International Meeting of the North American Fuzzy Information Processing Society. 2004

[8] Krantz D H, Luve R D, Suppes P, et al. Foundation of Measurement[M]. New York: Academic Press, 1971, 1, 1989, 2, 1990, 3

[9] Ganter B, Wille R. Conceptual scaling// Fred S. Roberts, ed. Applications of combinations and graph theory to the biological and social science[M]. Berlin, Heidelber, New York: Springer-Verlag, 1989：139-167

[10] Reurer K, Wille R. Complete congruence relations of concept lattice. Acta Sci. Math[M], 1987, 51：319-327