# 基于SVM的中文类比检索方法

梁 超 吕 钊 顾君忠
(华东师范大学信息科学与技术学院 上海 200241)

**摘 要** 随着互联网的不断发展,用户因不能准确输入查询关键字而无法准确获取未知领域信息的问题日益严重。作为一种根据已知领域知识获取未知领域知识的全新检索方式,类比检索逐渐成为研究热点。类比检索通过分析词对之间的潜在关系而准确地返回目标信息。例如,给定类比查询请求 $Q=\{A:B,C:?\}$,$A$ 与 $B$ 之间具有某种潜在关系,类比检索的目标是得到?所代表的目标词(集)$D$,其中 $A$ 与 $B$ 的关系和 $C$ 与 $D$ 的潜在关系相似。类比检索的两个难点是潜在关系挖掘和目标词抽取,这两个问题对于中文而言,更具挑战性。提出了基于SVM的中文类比检索方法(SVM based Chinese Analogy Retrieval,SVMbCAR)。该方法的两个主要成分包括基于SVM的关系代表词抽取和目标词确定。基于真实测试数据集(包含源自人立方的600个人物实体对)的实验表明,SVMbCAR方法抽取关系代表词的准确率为82.3%,抽取目标词的准确率为90.5%。

**关键词** 类比检索,SVM,语义相似
**中图法分类号** TP391 **文献标识码** A

## Chinese Analogy Retrieval Using SVM

LIANG Chao LV Zhao GU Jun-zhong
(Department of Computer Science and Technology, East China Normal University, Shanghai 200241, China)

**Abstract** With the development of Internet, the problem of not acquiring information of unknown domains because not exactly import keywords becomes more common. As a new retrieve method of acquiring knowledge of unknown domains using the knowledge of known domains, analogy retrieval gradually becomes one of hot topics. Analogy retrieval first analyzes the potential relationships between pairs of words and then accurately returns target information using these relationships. For example, given an analogy query $Q=\{A:B,C:?\}$, here it is assumed that there are some potential relationships between $A$ and $B$. The aim of analogy retrieval is to determine the target(s) $D$ of ?, and the relationships between two pairs of words, $A$ and $B$, $C$ and $D$, are similar. Two key difficulties of analogy retrieval are: (1) mining relationships between two words and (2) extracting target words. Both of them are more challenging in Chinese. This paper proposed a SVM based Chinese Analogy Retrieval (namely SVMbCAR) with two main components, SVM based relation-words extracting and SVM based target words determining. Experiments on a real-life data set (600 person entity pairs from Ren Li Fang) show that the accuracy of extracting relationships between two words is 82.3%, and the accuracy of extracting target words is 90.5%.

**Keywords** Analogy retrieval, SVM, Semantic similarity

## 1 相关工作

随着互联网的发展,网络上的信息越来越多,但用户不能准确描述查询意图而导致查询失败的问题一直是业界研究热点。人类认知特点表明:人们遇到一个新情况时,一般会从过去的经历中寻找一个类似的情景,即通过类比推理将过去经验(源域)中得来的经验知识应用到新场景(目标域)。但现有的搜索引擎缺乏对此类查询请求的有效支持。

例如,苹果公司产品的用户要查询微软的产品,但该用户不知道相关产品的名字或关键字去描述想要的产品。在这种情况下,用户所熟知的苹果产品成为搜索微软产品的重要线索。大多数苹果产品的用户知道iPad是一款苹果销售的电子产品,如果他们想找到一款这样的微软产品,可以利用苹果公司与iPad之间的关系和微软公司与该公司销售的平板电脑之间关系的相似性。这样,用户只需要简单地给一个熟知领域的例子就足够了。即用户给出一个类比查询三元组 Q=(苹果,iPad,微软),用户期望的检索结果Surface将会被搜索引擎返回。在这个类比查询中,(苹果,iPad)和(微软,Surface)之间的关系相似。

这种利用一个已知领域的例子进行查询,从未知领域获

到稿日期:2013-06-14 返修日期:2013-09-16 本文受上海市科学技术基金(11511504002)资助。
梁 超(1988—),男,硕士生,主要研究方向为大数据分析与知识处理,E-mail:lczc1988321@163.com;吕 钊(1970—),女,博士,副教授,主要研究方向为大数据分析与知识处理、文本挖掘及语义分析、本体及其应用,E-mail:zlu@cs.ecnu.edu.cn(通信作者);顾君忠(1949—),男,教授,博士生导师,主要研究方向为分布式数据处理、协同计算和情景感知计算。

取信息的搜索方式称为类比检索(Analogy Search,AS),又称为潜在关系检索(Latent Relation Search,LRS)。类比检索能有效地回答用户类比查询请求 $Q=(A:B,C:?)$,它通过从现有搜索引擎(例如,Google)返回的 Web 文档中抽取词对$(A:B)$之间的潜在关系,将获取的潜在关系与第 3 个词 $C$ 相结合,再次从搜索引擎返回的 Web 文档中查找并返回目标词?。这里的 $A$,$B$ 和 $C$ 可以是词语或者命名实体。类比检索是随着语义相似度和关系相似度等领域的发展于近几年才开始兴起且受到关注的,是新一代知识搜索引擎的一个体现,目前已经成为业界新的研究方向和研究热点。

文献调查发现,现有相关研究主要面向英语和日语。日本东京大学的 Kato 等[1]在 2009 年详细阐述了潜在关系搜索的概念和应用,提出了基于词语共现的方法和基于语义模式的方法。基于词语共现的方法使用词袋(bag-of-word)表示两个词语之间的关系,基于语义模式的方法则是将抽取出来的含有词对的一个短语中的 0 个或多个单词用 * 来代替之后构成模式。Duc 等[2]在 2010 年提出了日文的潜在关系搜索方法。这种方法从 Web 语料库中抽取词对,利用语义模式来代表词对之间的关系,获得了较高的准确率和召回率。但这种方法需要抓取大规模的语料并构建超大型的索引库,并且测试集限定在有限的领域内,没有测试其他领域的准确性。Goto 等[3]于 2010 年提出了一种利用关系相似对称性和辅助排名来改进关系搜索结果排名的方法。2011 年 Duc 等[4]提出的 CLRS 方法初步探索了跨语言、跨领域的知识映射。

国内也开展了针对类比检索的初步研究。2012 年,梁等[5]提出了一种中文潜在关系搜索方法。该方法采取词袋来描述词对之间的语义关系,即首先使用预定义的句法模式来抽取能描述词对之间的潜在关系的关系代表词(集),然后利用关系代表词来搜索满足相关预定义句法模式的目标词语(集)。但这种算法基于两种假设,且从概率统计上对候选关系词和目标词进行筛选,准确率相对较低。

现有基于概率统计的相关方法大都基于二种假设,但这二种假设却不适用于中文。通过观察大量实验数据,笔者发现:能准确地抽取出词对之间的潜在关系的句子一般满足某种特殊结构,本文中把这类句子叫做关系句。基于以上的分析,本文提出了基于 SVM 的类比检索方法(SVMbCRA),即首先利用 SVM 将所有包含词对的句子分类,即划分为"潜在关系句"和"非潜在关系句"两类。然后再从关系句中抽取词对之间的潜在关系并进行目标词确定。从而有效地提高了关系词抽取的准确率和目标词识别的准确率。

## 2 基于 SVM 的类比检索方法

基于 SVM 的中文类比检索方法的主要处理步骤如图 1 所示。该方法的主要步骤包括确定关系代表词 $R$ 和抽取目标词(集)$D$。图 1 中,实线描述了关系代表词 $R$ 的抽取过程,即通过这一步骤可以得到描述词对$(A,B)$之间潜在关系的关系代表词(集)$R$;虚线描述了目标词的抽取过程,即通过这一步骤可以得到目标词(集)$D$。

关系代表词的抽取过程如下:(1)输入词 $A$ 和词 $B$,预处理模块通过调用搜索引擎获取包含词 $A$ 和词 $B$ 的网页及句子;(2)提取句子特征生成特征向量,通过 SVM 训练后生成训练模型;(3)利用生成的训练模型对步骤(1)中得到的句子进行测试,识别出关系句;(4)抽取出关系代表词并进行排序。

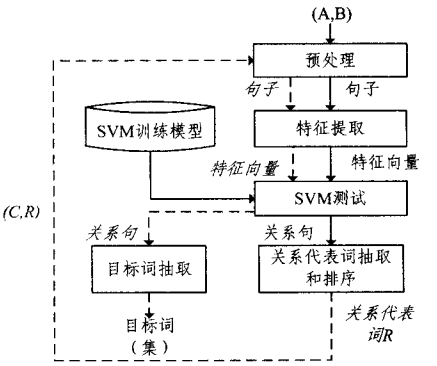


图 1 SVMbCAR 的流程图

目标词的抽取过程类似于关系代表词的抽取,区别在于输入为 $C$ 和关系代表词 $R$ 组成的新词组$(C,R)$。

### 2.1 确定关系代表词

#### 2.1.1 预处理

预处理模块的主要功能是抽取包含输入词对的句子,然后进行分词和词性标注处理。

预处理模块首先通过搜索引擎(例如,Google)抽取包含词 $A$ 和词 $B$ 的搜索结果,即得到包含词 $A$ 和词 $B$ 的一系列网页网址,依次抓取每个网址指向网页中包含词 $A$ 和词 $B$ 的句子。接着将包含词对$(A,B)$的句子从这些文本语料中提取出来。此处以词对$(X,Y)$为例,抽取同时包含 $X$ 和 $Y$ 词的句子。

VSMbCRA 方法抽取满足以下模式的句子:

$$m*X*Y*m \text{ 或 } m*Y*X*m \quad (1)$$

其中,$m$ 为断句的标点符号或空格,* 代表除 $m$ 外的任意字符。采取此方法,我们可以抽取词 $X$ 之前,词 $X$ 与词 $Y$ 之间及词 $Y$ 之后的词语(或短语)。

表 1 包含词对$(X,Y)$的句子示例表

| 序号 | X | Y | R | 句子示例 |
|---|---|---|---|---|
| 1 | 刘翔 | 孙海平 | 教练 | 中国田径选手刘翔的**教练**孙海平在赛后新闻发布会上落泪。 |
| 2 | 冯巩 | 郭冬临 | 搭档 | 央视春晚:冯巩与郭冬临**搭档**。 |
| 3 | 舒乙 | 老舍 | 父亲 | 我的**父亲**老舍由老舍先生的**儿子**舒乙所作。 |
| 4 | 彭莹玉 | 周子旺 | 弟子 | 一三三八年(后至元四)彭莹玉和他的**弟子**周子旺以寅年、寅月、寅日、寅时率众起义。 |
| 5 | 章孝严 | 黄美伦 | 妻子 | 章孝严的**妻子**黄美伦。 |
| 6 | 牛顿 | 巴罗 | 老师 | 牛顿在**老师**巴罗的指导下,在钻研笛卡尔的解析几何的基础上,找到了新的出路。 |
| 7 | 马原野 | 齐华 | 同事 | 在研究阿尔茨海默病、抑郁症的同时,马原野及其**同事**齐华副研究员一起探讨了毒品成瘾患者吸食毒品前后瞳孔变化的特点。 |
| 8 | 陶铸 | 曾希圣 | 战友 | 一是节假日老朋友相聚,二是中央开会期间,老**战友**如陶铸、曾希圣、张平化等约齐到关东店看望。 |
| 9 | 王定标 | 林栋梁 | 合伙人 | 消息称大旗董事长王定标在 IDG **合伙人**林栋梁和过以宏的引荐下联系到了 SIG。 |
| 10 | 贺子珍 | 贺敏学 | 哥哥 | 不仅贺子珍,她的**哥哥**贺敏学、妹妹贺怡在学校里都没有"安安分"读书,也都去闹学潮搞斗争。 |

表 1 中列出了一些抽取出的句子。其中,第 2 列和第 3 列表示词 $X$ 和词 $Y$,第 5 列表示抽取的包含词 $X$ 和词 $Y$ 的句

子示例。

其次，将所抽取的包含词对的句子进行分词，抽取每个分词后序列的主干。本文利用中文分词工具 ICTCLAS 将每个句子分割成独立的词语[6]。每一个词都有自己的词性标注。例如，“章孝严/nr 的/uedl 妻子/n 黄美伦/nr”。

预处理后，将得到大量分词标注后的句子。现在的问题是如何确定能代表两个词语之间潜在关系的词语(即，如何确定关系代表词)。

2.1.2 模型建立

包含两个输入词语的句子可以分为两类，第一类被称为潜在关系句，即句子中有明确表示词对之间关系的词语，如表 1 中的前 10 个句子(其中表征关系的词用粗体表示)；第二类是普通句子，即不含有表示词对关系的词语。我们把表征词对之间关系的词语称为关系代表词，把包含关系代表词的句子称为潜在关系句。因此，如果能够识别出潜在关系句，则可以精确地找到关系代表词。

如前所述，本文把潜在关系句识别看作是一个二分类问题，即“潜在关系句”与“非潜在关系句”两种类别。SVM (Supported Vector Machine)是一种建立在统计学习理论基础之上的机器学习方法。该方法建立了一套有限样本下机器学习的理论框架和通用方法，能够较好地解决小样本、线性和非线性等实际分类问题。因此在本文中采取 SVM 分类器来识别潜在关系句。

潜在关系句识别可以总结为：估计类别 $y$ 和潜在关系句特征 $x$ 的共现概率 $p(x,y)$，这个共现概率是用来衡量一个潜在关系句在特征 $x$ 下成为 $y$ 类别的置信度，从而实现对潜在关系句的分类。在这里类别 $y$ 主要表示“潜在关系句”和“非潜在关系句”两类，而 $x$ 就是潜在关系句的特征向量。

潜在关系句识别问题可以转化成如下数学模型，即：

给定训练样本集合为 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$，其中 $x_i\in R^n$，$y_i\in\{-1,1\}$。这里的 $x_i$ 即为潜在关系句样本的特征向量，而 $y_i$ 即为潜在关系句的类别。那么目标问题就转化成在 $R^n$ 上寻找一个实函数 $g(x)$，使得决策函数 $f(x)=\mathrm{sgn}(g(x))$能够成立，这样就使得每个样本都能够得出它所对应的 $y$ 值。

在利用 SVM 识别潜在关系句时，本文提出根据每一个潜在关系句的语言学特征来判断它是否为潜在关系句。主要过程是利用 SVM 结合候选潜在关系句的各种语言学特征进行处理，计算出候选潜在关系句在这种特征下可以成为潜在关系句的置信度，SVM 选择置信度大于某阈值的句子作为潜在关系句。

2.1.3 特征选取

本文从两个方面选取特征，即形态学特征和词对的上下文特征。形态学特征主要包括：(1)一个词对中词语的位置关系；(2)词对的上下文相对词语的位置。词对的上下文特征是表征句子特征的重要信息，词对的上下文主要指上下文中词语构成序列。一般地，可以把一个句子划分为 4 个部分：词语 $A$ 和词语 $B$、词语 $A$ 之前的词、词语 $A$ 与词语 $B$ 之间的词和词语 $B$ 之后的词。

词语周围一定范围的上下文可以为词语的定义提供较为充分的语言信息。如何确定每一部分词语的数量是一个需要解决的问题。在文献[8]中采用信息论中信息增益的计算方法来实现上下文有效范围的量化确定。计算结果反映了这样一个情况：上下文对核心词语的描述能力随着相对位置由近及远而逐渐递减，即通过已知上下文推断空缺词语时，近距离的上下文在推断中所起的作用比远距离的上下文更有价值。这一计算结果与人们的认知过程基本一致。虽然是近似结果，但在一定程度上具有统计意义。本文根据这一原则将上下文范围确定为 $A$ 词前 2 个位置，$A$ 与 $B$ 词中间 3 个位置和 $B$ 词后 2 个位置的范围大小。

根据确定的上下文范围，将句子定义为如下形式：

$$a1,a2,A,c1,c2,c3,B,b1,b2 \tag{2}$$

分词后的词语全部进行了词性标注，词性标注的序列如下所示：

$$a1pos,a2pos,c1pos,c2pos,c3pos,b1pos,b2pos \tag{3}$$

上式称为 N-POS。

对大量抓取的包含词 $A$ 和词 $B$ 的句子进行分析，如表 2 所列。本文确定 $c1$,$c2$,$c3$,$b1$ 和 $b2$ 作为潜在关系句识别的上下文范围。

表 2 潜在关系句的 N-POS 表

| c1pos | c2pos | c3pos | b1pos | b2pos |
|---|---|---|---|---|
| 是 | 0 | 0 | 30058 | 28160 |
| 24832 | 30058 | 28160 | 0 | 0 |
| 30058 | 28160 | 是 | 0 | 0 |
| 30058 | 28160 | 叫 | 0 | 0 |
| 和 | 0 | 0 | 30058 | 28160 |
| 就是 | 0 | 0 | 30058 | 28160 |
| 为 | 0 | 0 | 30058 | 28160 |
| 以 | 0 | 0 | 为 | 28160 |
| 视 | 0 | 0 | 为 | 28160 |
| 30208 | 30058 | 28160 | 0 | 0 |
| 与 | 0 | 0 | 30058 | 28160 |
| 28160 | 28160 | 0 | 0 | 0 |
| 28160 | 28160 | 0 | 0 | 0 |
| 30208 | 30058 | 0 | 0 | 0 |
| 对 | 28160 | 0 | 0 | 0 |
| 之 | 28160 | 0 | 0 | 0 |
| 与 | 28160 | 0 | 0 | 0 |
| 30058 | 28160 | 0 | 0 | 0 |
| 和 | 28160 | 0 | 0 | 0 |

设有$(A,B)$为词对，特征向量的集合为：

$$V(A,B)=(v_1,v_2,\cdots,v_n) \tag{4}$$

向量的每一维是这个词对的各项特征形成的七元组。

$$v_i=(distance,order,c1pos,c2pos,c3pos,b1pos,b2pos) \tag{5}$$

其中：

1)*distance* 为命名实体对中两个命名实体之间的词距。

2)*order* 为两个命名实体的相对位置，分别为 0($A$ 在 $B$ 左边)，1($B$ 在 $A$ 左边)。

3)N-POS 即词语序列中 $N$ 个连续词性的顺序组合，本文认为几种连续词性的组合表征关系信息。其中 *c1pos*、*c2pos*、*c3pos*、*b1pos*、*b2pos* 即为 N-POS 序列组合。

2.1.4 识别潜在关系句

本文的潜在关系句识别主要分两大模块实现：(1)SVM 训练模块；(2)SVM 测试模块。在训练模块中，首先进行特征选取并选取特征数据生成特征向量，然后训练生成目标模型。

在测试模块中首先根据选取的特征生成特征向量，然后利用目标模型对特征向量进行识别，最终识别出潜在关系句。

在训练过程中，特征提取主要是针对特征模型提取其相应的特征。在训练语料库中对分词后句子的特征信息进行统计，这些信息主要是词语顺序(order)、词语间距(distance)、词语序列及词性(N-POS)。表3列出了训练样本中提取的一些潜在关系句的特征。对于训练数据，识别出相同数量的潜在关系句和非潜在关系句进行正反例标注并生成特征向量。

表3 潜在关系句的特征示例表

| 潜在关系句 | Order | Distance | c1pos | c2pos | c3pos | b1pso | b2pos |
|---|---|---|---|---|---|---|---|
| 中国田径选手刘翔的教练孙海平 | 1 | 3 | 的 | 28160 | 0 | 0 | 0 |
| 央视春晚：冯巩与郭冬临搭档 | 1 | 2 | 与 | 0 | 0 | 28160 | 0 |
| 我的父亲老舍由老舍先生的儿子舒乙所作 | 0 | 3 | 28160 | 的 | 28160 | 所 | 30208 |
| 马原野及其同事齐华副研究员 | 1 | 4 | 及其 | 28160 | 0 | 28160 | 0 |
| 牛顿在老师巴罗的指导下 | 1 | 3 | 在 | 28160 | 0 | 的 | 28160 |

SVM训练主要是利用之前从训练语料库中提取的正负特征向量。将训练样本向量进行SVM训练，得到支持向量模型。本文采用林智仁开发的支持向量机学习测试工具包LibSVM[7]，LibSVM程序包提供4种核函数供选择：线性、多项式、径向基、sigmoid核函数。程序包作者通过实验指出在一般情况下使用径向基核函数较好，但在特征数量较大时则应该选择线性核函数。本文的研究重点是有效区分潜在关系句和非潜在关系句，选取了程序包默认参数。通过SVM的训练，使正负样本特征向量生成支持向量模型，为后续的测试识别潜在关系句提供测试分类的数据依据。

为了提高SVM训练的准确度，本文将 $c1$、$c2$、$c3$、$b1$、$b2$ 所对应的词性标注映射到连续区间，并将表征潜在关系句的词性标注和其他词性标注划分为两个连续的数值区间。例如将 $b1pos$ 的标注“30058”、“为”、“是”、“28160”分别映射为0、1、2、3。其余的标注依次从4开始映射。

整个测试过程的流程如下：同训练过程，首先按照句子划分的定义，记录下每个位置的上下文词语的词性。同时记录下两个实体间的位置关系，包括实体的前后顺序和两个实体之间的间距。然后生成潜在关系句的正负样本特征向量，最后将生成的特征向量进行SVM训练，从而标记出测试样本的正负实例。测试的过程其实就是把候选潜在关系句分成正和负两类，正表示潜在关系句，负表示非潜在关系句。

#### 2.1.5 候选关系代表词排名

识别出潜在关系句后，我们从这些潜在关系句中提取出候选关系代表词。

候选关系句排名分为4个步骤：(1)识别出潜在关系句后，抽取出其中的候选关系代表词；(2)将潜在关系句中出现的名词和动词进行词频统计；(3)统计完成后，计算词语之间的相似度进行合并；(4)最后对合并后的候选关系代表词排序。具体算法见算法1。

**算法1** 候选关系代表词排名算法

输入：Set S(A,B) //识别出的潜在关系句
输出：Set R(A,B) //排名后的关系代表词集合

```
1. For each s in S(A,B) do //假设性检验
2.   For each w in s //w为s分词后得到的词语
3.     If w is Noun or w is Verb then
4.       If w exist in E(A,B) //E(A,B)临时存储候选关系词
5.         add 1 to f(w) //f(w)为w的词频
6.       Else
7.         add w to E(A,B)
8.         set F(w) to 1
9.       End if
10.    End if
11.  End for
12. End for
13. For each in E(A,B) do
14.   For each in E(A,B) do
15.   Add sim(t_i,t_j) to S[i,j] //S[]为记录词语相似度的矩阵
16.   End for
17. End for
18. For each in E(A,B) do //按照语义相似度进行合并
19.   For each t_j(j>i) in E(A,B) do
20.     If S[i,j]>θ then
21.         combine t_i into t_j
22.         If t_j∈R(A,B) then
23.           update R(A,B)
24.         Else
25.           Add t_j to R(A,B)
26.         End if
27.       End if
28.     End for
29. End for
30. For each t_i in R(A,B) do //按照词频进行排序
31.   key=t_i,k=i
32.     For each t_j(j>i) in R(A,B) do
33.       If t_j<k then
34.         k=j
35.       End if
36.     End for
37.         Swap(t_i,t_k)
38. End for
39. Return R(A,B)
```

### 2.2 抽取目标词

目标词的抽取过程和关系代表词的抽取过程基本相同。主要步骤包括：抽取出的关系代表词 $R$ 同 $C$ 组合成新的词对 $(C,R)$。通过2.1.1节中介绍的预处理过程得到包含 $C$ 和 $R$ 的句子。然后按照2.1.2节、2.1.3节和2.1.4节介绍的方法抽取句子特征生成特征向量，利用训练得到的目标模型对特征向量进行识别，识别出潜在关系句。最后从潜在关系句中提取出满足N-POS特征的候选目标词。

## 3 实验结果及分析

### 3.1 测试数据集和实验方案设计

现有的实验数据集仅涉及日文和英文，没有标准中文实

验数据集。人立方[9]关系搜索是微软亚洲研究院开发设计的一款新型社会化搜索引擎，它能从超过十亿的中文网页中自动地抽取出人名、地名、机构名以及中文短语，并通过算法自动地计算出它们之间存在的可能关系。本文考虑了人立方中人与人的13种关系，即师徒，夫妇，情侣，父子，母子，兄弟，姐妹，合伙，搭档，朋友，同事，战友，经纪人，从人立方中选取了2400人物实例，并构成1200个测试词对。这些测试词对涵盖了上述13种关系。

部分测试词对及组成该词对的两个词之间的关系如表4所列。对于每一种关系大类中的实例，分别取一半用来进行关系代表词确定实验，一半用来进行目标词确定实验。因此关系代表词确定和目标词确定实验的测试词对均为600组。

表4 测试词对举例

| 序号 | 词A | 词B | 关系 |
|---|---|---|---|
| 1 | 方伟 | 汪洋 | 情侣 |
| 2 | 陈少白 | 孙中山 | 朋友 |
| 3 | 刘聪 | 刘义 | 兄弟 |
| 4 | 马原野 | 齐华 | 同事 |
| 5 | 维罗尼卡·帕夫洛维奇 | 张怡宁 | 姐妹 |
| 6 | 冯巩 | 郭冬临 | 搭档 |
| 7 | 张伟平 | 张艺谋 | 搭档 |
| 8 | 老舍 | 舒乙 | 父子 |
| 9 | 马林 | 尼莫 | 父子 |
| 10 | 肖金萍 | 肖云成 | 父子 |
| 11 | 蒋宏伟 | 李娜 | 师徒 |
| 12 | 刘翔 | 孙海平 | 师徒 |
| 13 | 唐学华 | 谢杏芳 | 师徒 |
| 14 | 李永波 | 林丹 | 师徒 |
| 15 | 彭莹玉 | 周子旺 | 师徒 |
| 16 | 刘学宇 | 扬戈维奇 | 师徒 |
| 17 | 巴罗 | 牛顿 | 师徒 |
| 18 | 克林斯曼 | 莱曼 | 师徒 |
| 19 | 吴金贵 | 肖战波 | 师徒 |
| 20 | 诺维茨基 | 约翰逊 | 师徒 |
| 21 | 陈亢 | 孔子 | 师徒 |
| 22 | 萧芸 | 叶小平 | 夫妇 |
| 23 | 黄美伦 | 章孝严 | 夫妇 |
| 24 | 江月华 | 周明 | 夫妇 |
| 25 | 霍光 | 霍显 | 夫妇 |
| 26 | 莫扎特 | 斯特凡 | 夫妇 |
| 27 | 李讷 | 王景清 | 夫妇 |
| 28 | 李毅男 | 王秀佳 | 夫妇 |
| 29 | 滨滨 | 璐璐 | 夫妇 |

SVMbCAR方法有两个重要步骤，步骤1是潜在关系句识别，步骤2是关系代表词抽取和目标词抽取。为此，本文设计了两个实验，第1个实验测试潜在关系句识别的效果，第2个实验测试关系代表词抽取及目标词抽取的效果。这两个实验的机器配置如下：处理器为 Intel® Pentium® CPU G630 @ 2.7GHz，内存4GB，操作系统为Win7 64位。

## 3.2 实验评估

### 3.2.1 潜在关系句识别效果评估

潜在关系句识别实验采用训练测试时间和准确率来衡量实验效果。其中表5中的均方误差为SVM衡量测试准确率的相关指标，均方误差越低，说明准确率越高。

本文选取3组测试数据来训练SVM模型，第1组中的正反例个数均为2000，第2组的正反例个数均为4000，第3组的正反例个数均为8000。各组训练时间和模型大小如表5中第2列和第3列所列。在识别之前对测试数据进行人工的正反例标注，标注方法与训练数据的标注方法相同，即对于N-POS序列符合潜在关系句的句子标注为正例，其余的标注为反例。测试结果如表5的第4列所列。表5显示，随着训练正反例个数的增加，训练模型时间增加，但测试时间整体变化不大。

表5 SVM识别效果

| 组号 | 训练时间/s | 测试时间/s | 测试准确率 | 均方误差 |
|---|---|---|---|---|
| 1 | 10.7862459 | 379.941351 | 98.411126515% | 0.063554939 |
| 2 | 29.51294 | 384.8341788 | 99.281827854% | 0.028726886 |
| 3 | 122.2432039 | 434.3740862 | 99.455710720% | 0.021771571 |

从表5中可以看出，随着训练数据的增加，准确率有所提高。实验结果表明，SVMbCAR算法识别潜在关系句是十分有效的。

### 3.2.2 关系代表词和目标词抽取效果评估

关系代表词抽取从3个方面进行评估，即准确率($P$)、期望目标词排名的倒数平均值($MRR$)和期望目标词排名在前$k$位的百分比($Rank(i)$)。目标词抽取采用准确率($P$)进行评估。准确率指的是期望词排名在第一位的测试词对占所有测试词对的比例。$MRR$是期望词排名的倒数平均值。$Rank(i)$是指期望排名在前$i$位的测试词对所占的比例。

$$P=\frac{\text{期望词排名在第一位的测试词对}}{\text{测试词对总数}}*100\% \tag{6}$$

$$MRR=\frac{\sum_{i=1}^{n}1/\text{第}i\text{组期望词排名}}{\text{测试词对总数}} \tag{7}$$

$$Rank(i)=\frac{\text{期望词排名在前}i\text{位的词对}}{\text{测试词对总数}} \tag{8}$$

文献调查发现，现有的类比检索相关研究围绕着日语和英语展开，鉴于中文的语言特点，面向日语和英语的类别检索方法不能直接处理中文。文献[5]提出了面向中文的基于统计的类比检索方法(简称为STAbCAR方法)。因此本文将SVMbCAR方法和STAbCAR方法进行了比较。

表6显示了两种方法针对600组测试数据在关系代表词抽取方面的实验效果，即两种方法得到的关系代表词排名在每个区间中的实例个数。表6中，第1列为两种方法，第2列至第7列分别表示两种方法抽取出的关系代表词的个数的排名情况，即第1位的个数，第2—5位的个数，第6—10位的个数，依次类推。从表6中可以明显看到代表STAbCAR算法关系代表词排名在第1位的个数少于SVMbCAR算法，越往后，实例个数分布越少。SVMbCAR算法抽取的关系代表词排名前10的占到了99.3%，STAbCAR算法占到了97.3%。

表6 600组关系代表词排名

| | 1 | 2—5 | 6—10 | 11—15 | 16—25 | 26—30 | >30 |
|---|---|---|---|---|---|---|---|
| STAbCAR | 458 | 97 | 29 | 8 | 2 | 4 | 2 |
| SVMbCAR | 523 | 52 | 16 | 5 | 0 | 2 | 2 |

表7显示，SVMbCAR方法抽取关系代表词的准确率为82.3%，比使用STAbCAR方法抽取关系代表词的准确率提高了10.2%。SVMbCAR方法的MRR值由78.7%提高到了85.9%。

表7 两种方法抽取关系代表词的结果比较

| | 正确个数 | 准确率 | MRR |
|---|---|---|---|
| STAbCAR | 433 | 0.721667 | 0.787447 |
| SVMbCAR | 494 | 0.823333 | 0.859232 |

(下转第149页)

[14] Li Tie-yan. Employing lightweight primitives on low-cost rfid tags for authentication[C]// Vehicular Technology Conference, 2008(VTC 2008-Fall), IEEE 68th. IEEE, 2008:1-5

[15] Peris-Lopez P, Hernandez-Castro J C, Tapiador J M E, et al. Advances in ultralightweight cryptography for low-cost RFID tags: Gossamer protocol[C]// Proc. International Workshop on Information Security Applications(WISA'08), 2009. Berlin: Springer Berlin Heidelberg, 2009:56-68

[16] Tian Yun, Chen Gong-liang, Li Jian-hua. A New Ultralightweight RFID Authentication Protocol with Permutation[J]. IEEE Communications Letters, 2012, 16(5):702-705

[17] Gurubani J B, Thakkar H, Patel D R. Improvements over extended LMAP+: RFID authentication protocol[C]// 6th International Conference on Trust Management-FIPTM, 2012. Surat: Springer Boston, 2012:225-231

[18] Peris-Lopez P, Hernandez-Castro J C, Estevez-Tapiador J M, et al. M2AP: A minimalist mutual-authentication protocol for low-cost RFID tags[C]// Proc. of UIC, 2006. Berlin: Springer-Verlag, 2006:912-923

[19] Li Tie-yan, Wang Gui-lin. Security analysis of two ultra-lightweight RFID authentication protocols[C]// Proc. of IFIP-SEC'07, 2007. Sandton: Springer US, 2007:109-120

[20] Phan R C W. Cryptanalysis of a new ultralightweight RFID authentication protocol-SASI[J]. IEEE Transactions on Dependable and Secure Computing, 2008, 6(4):316-320

[21] Cao Tianjie, Bertino E, Lei Hong. Security Analysis of the SASI Protocol[J]. IEEE Trans. Dependable and Secure Computing, 2009, 6(1):73-77

[22] Yeh K H, Lo N W. Improvement of two lightweight RFID authentication protocols[J]. Information Assurance and Security Letters, 2010, 1:6-11

[23] Ahmadian Z, Salmasizadeh M, Aref M R. Desynchronization attack on RAPP ultralightweight authentication protocol[J]. Information Processing Letters, 2013, 113(7):205-209

[24] Wang Shao-hui, Han Zhi-jie, Liu Su-juan, et al. Security analysis of RAPP: an RFID authentication protocol based on permutation [R]. Cryptology ePrint Archive, Report 2012/327. 2012

[25] Clark J A, Jacob J L. Fault Injection and a Timing Channel on an Analysis Technique[C]// International Conference on the Theory and Applications of Cryptographic Techniques, 2002. Berlin: Springer-Verlag, 2002:181-196

[26] Juels A, Weis S A. Defining Strong Privacy for RFID[J]. ACM Transactions on Information and System Security, 2009, 13(1): 342-347

(上接第 115 页)

在目标词的确定过程中，我们可以明确目标词 $D$ 一定是名词并且目标词的出现相对于关系代表词的出现更加集中和准确。因此，在目标词的确定实验部分不再使用 *MRR* 指标，只要出现在排名第 1 位则认为准确，否则错误，因此这里只采用正确率进行衡量。从表 8 可以看出，SVMbCAR 抽取目标词的准确率达到了 90.5%，而 STAbCAR 的准确率为82.8%。

表 8 两种方法抽取目标词的结果比较

| | 正确个数 | 准确率 |
|---|---|---|
| STAbCAR | 497 | 0.828333 |
| SVMbCAR | 543 | 0.905 |

**结束语** 类比检索是一种根据已知领域知识查询未知领域知识的全新检索方式。通过分析词对间的潜在关系，类比检索可以准确地返回目标信息。即，给定查询请求 $Q=\{A:B,C:?\}$，类比检索的目标是得到?所代表的目标词 $D$，其中 $A$ 与 $B$ 的关系和 $C$ 与 $D$ 的关系相似。本文提出了基于 SVM 的中文类比方法。该方法首先利用 SVM 识别潜在关系句，然后抽取出潜在关系句中的相关词语作为关系代表词或目标词。本文采用从人立方中选取的 600 组人物实例对设计了两个实验，一是针对潜在关系句识别的实验，二是针对关系代表词抽取和目标词抽取的实验。实验结果表明了，本文提出的 SVMbCAR 在潜在关系句识别、关系代表词和目标词抽取两个方面的效果。

本文提出的 SVMbCAR 方法虽然取得了良好效果，但也存在一些需要改进的地方：(1)由于需要实时抓取大量网页语料进行处理，时间和空间消耗较大。接下来的研究考虑通过提高关系词的提取准确率来减少网页访问次数，以实现实时的类比检索。(2)在 SVMbCAR 算法中，特征提取的有效性对识别准确率有较大的影响，未来考虑选取更加有效的特征来构造特征向量，提高潜在关系句的识别准确率，进而提高抽取关系词和目标词的准确率。(3)进一步研究 SVM 的不同参数对实验效果的影响。

## 参 考 文 献

[1] Kato M P, Ohshima H, Oyama S, et al. Query by analogical example: relational search using Web search engine indices[C]// Proceedings of the 18th ACM Conference on Information and Knowledge Management. ACM, 2009:27-36

[2] Duc N T, Bollegala D, Ishizuka M. Using relational similarity between word pairs for latent relational search on the web[C]// 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). IEEE, 2010, 1:196-199

[3] Goto T, Duc N T, Bollegala D, et al. Exploiting symmetry in relational similarity for ranking relational search results[C]//PRICAI 2010: Trends in Artificial Intelligence. Springer Berlin Heidelberg, 2010:595-600

[4] Duc N T, Bollegala D, Ishizuka M. Cross-language latent relational search: Mapping knowledge across languages[C]// Proceedings of 25th AAAI Conference on Arificial Intelligence. 2011:1237-1242

[5] Liang Chao, Lu Zhao. Chinese Latent Relational Search Based on Relational Similarity[M]// Data and Knowledge Engineering. Springer Berlin Heidelberg, 2012:115-127

[6] 中科院分词系统ICTCLAS[OL]. http://www.ictclas.org/. 2012

[7] Chang C-C, Lin C-J. LIBSVM: a library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3):27

[8] 鲁松，白硕，黄雄. 基于向量空间模型中义项词语的无导词义消歧[J]. 软件学报，2002，13(6)：1082-1089

[9] 人立方关系搜索[OL]. http://renlifang.msra.cn/