

文章编号:1001-9081(2006)06-1389-03

## 一种基于贝叶斯分类与机读词典的多义词排歧方法

谈文蓉<sup>1</sup>,符红光<sup>2</sup>,刘莉<sup>1</sup>,杨宪泽<sup>1</sup>

(1. 西南民族大学 计算机科学与技术学院,四川 成都 610041;

2. 中国科学院 成都计算机应用研究所,四川 成都 610041)

(tan1781@sina.com)

**摘要:**一词多义是自然语言中普遍存在的现象,词义排歧的成功率是衡量机器翻译、信息检索、文本分类等自然语言处理软件性能的重要指标。提出了一种基于贝叶斯分类与机读词典的多义词排歧方法,通过小规模语料库的训练和歧义词在机读词典中的语义定义来完成歧义的消除。实验表明:基于贝叶斯分类与机读词典的多义词排歧算法在标注语料库规模受限的情况下,能取得较高的排歧准确率。

**关键词:**词义排歧;语料库;机读词典;自然语言处理

**中图分类号:** TP181 **文献标识码:** A

## Method of word sense disambiguation based on bayes and machine readable dictionary

TAN Wen-rong<sup>1</sup>, FU Hong-guang<sup>2</sup>, LIU Li<sup>1</sup>, YANG Xian-ze<sup>1</sup>

(1. College of Computer Science and Technology, Southwest University for Nationalities, Chengdu Sichuan 610041, China;

2. Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu Sichuan 610041, China)

**Abstract:** multi-senses of a word are widespread phenomenon in the natural language. The accuracy rate of sense ambiguity is the most important target of a software on the fields of machine translation, information indexing and text sorting. A method based on the bayes and machine readable dictionary was proposed, which could disambiguate by the training of a small-scale corpus and the definition of semantic in machine dictionary. The experimental results show that it has a high accuracy rate of word sense ambiguity when the scale of markup corpus has been limited.

**Key words:** word sense disambiguation; corpus; machine readable dictionary; natural language processing

### 0 引言

多义词是汉语及其他自然语言中普遍存在的现象,词义排歧成功率的高低将很大程度决定机器翻译、信息检索、文本分类等自然语言处理软件的性能。由于多义词的排歧涉及到上下文因素、语义因素、语境因素,甚至还涉及到日常生活中的常识,而这些因素的处理,恰恰是计算机最感棘手的问题<sup>[1]</sup>,因此词义排歧被认为是自然语言计算机处理中最困难的问题之一。

近年来对词义排歧方法的研究取得了很大的进展,但距机器翻译等实用软件的要求还有相当大的差距。词义排歧方法主要分为基于 AI、基于知识和基于语料库三类,由于统计模型具有鲁棒性和概括性,在含有错误的数据库和新数据中性能优异,因此基于语料库的统计学习方法在词义消歧中逐渐占据了主流的地位。文献[2]提出的逐渐扩大搭配实例的排歧方法、文献[3]提出的基于对数模型的多义词自动消歧方法、文献[4]提出从搭配知识获取最优种子的词义消歧方法等基于统计的词义排歧方法都从不同的角度取得了较好的消歧效果。用于统计学习的语料库是按照某种标准收集的特殊文本材料,通过对大规模语料库的自动或半自动学习来决定单词

不同意义。根据用于学习的训练语料事先是否经过人工标注统计学习方法又可分为有监督学习和无监督学习两种。虽然无监督学习方法不需要人工标注语料库,但由于方法本身的一些局限性,排歧准确率还不够高。有监督学习方法要使用人工标注语料库,理论上具有较高的排歧准确率,但给庞大的文本资源加上标记,本身就是一件极具挑战性的工作。由于收集和处理语料中要耗费相当大的工作量,所以大规模标注语料库的价格非常昂贵。但如果标注语料库的规模过小,又会产生严重的数据稀疏,对大规模真实文本的排歧准确率又将大幅下降。显然,标注语料库的规模问题正逐渐成为解决词义消歧的知识瓶颈。

本文针对获取大规模标注语料库较为困难的情况,将有监督的统计学习和传统的基于知识的排歧方法结合起来,提出了一种基于贝叶斯分类与机读词典的多义词排歧方法。通过对一些现成的免费文本资源的加工,构建具有一定规模的标注语料库。在对已标注语料库的统计学习基础上,采用贝叶斯和语义同现概率相结合的方式对歧义词进行最大概率排歧。实验表明在标注语料库的规模受限的情况下,该方法仍能取得较高的排歧准确率。

收稿日期:2005-12-20 基金项目:四川省重点科技攻关项目(05SG022-016)

作者简介:谈文蓉(1968-),女,四川广安人,副教授,硕士,主要研究方向:自然语言处理、数据库;符红光(1965-),男,四川成都人,研究员,博士生导师,主要研究方向:计算机自动推理、人工智能软件;刘莉(1965-),女,四川成都人,副教授,硕士,主要研究方向:数据库、自然语言处理;杨宪泽(1954-),男,四川成都人,教授,主要研究方向:自然语言处理、算法与数据结构。

## 1 贝叶斯分类与机读词典相结合的概率模型

根据最优化原则,词义排歧就是根据测试文本中的语法、语义、语用和语境等信息来选择歧义词最大可能的词义项。影响词义消歧特征是多方面的,一般情况下,词类、语法功能、句法关系、语义搭配均是影响词义消歧的主要因素,这些因素往往包含在句子的上下文中。我们将测试文本词语序列的上下文看做是一个无结构词集,通过对上下文窗口中众多词汇信息的整合来消除歧义。本文在一个特定的上下文窗口中考虑歧义词周围词的信息,利用贝叶斯概率和语义同现概率来确定多义词在特定上下文中的含义。

### 1.1 歧义词的贝叶斯概率

若用于训练的小规模语料库中每个歧义词的出现都标记好了其正确的语义,消歧就有了一个统计分类的实例。对于已构建的标注语料库,通过每个词的上下文窗口来收集数据。假定训练时歧义词上下文窗口的大小取  $L$ ,则设  $W$  是一个歧义词,  $s_1, \dots, s_k, \dots, s_K$  是歧义词  $W$  的  $K$  个语义项,  $c_1, \dots, c_l, \dots, c_L$  是歧义词  $W$  在训练样本集中的上下文,则通过标注语料库的训练可得到以下概率取值:

$$P(c_l | s_k) =$$

歧义词  $W$  的词义为  $s_k$  时以  $W$  为中心的窗口  $C_l$  出现的次数  
歧义词  $W$  的词义为  $s_k$  时以  $W$  为窗口中心的次数

$$P(c_l) = \frac{c_l \text{ 在给定的语料库中出现的总次数}}{\text{给定的语料库的总词数}}$$

$$P(s_k) =$$

歧义词  $W$  的词义为  $s_k$  在给定的语料库中出现的总次数  
歧义词  $W$  在给定语料库中出现的总次数

再来看测试语料库,假定上下文窗口的大小为  $J$ , 设  $v_1, \dots, v_j, \dots, v_J$  是歧义词  $W$  上下文窗口中的特征词,  $s_1, \dots, s_k, \dots, s_K$  是歧义词  $W$  的  $K$  个语义项。现在,统计分类的任务就是构建一个分类器,根据上下文窗口中的信息对新的歧义词进行分类。

$$P(s_k | v_1 \dots v_j \dots v_J) = \frac{P(v_1 \dots v_j \dots v_J | s_k)}{P(v_1 \dots v_j \dots v_J)} P(s_k)$$

根据贝叶斯独立性假设,每个词  $v_j$  的出现独立于窗口中的其他词,  $P(v_1 \dots v_j \dots v_J)$  可近似地改写为  $\prod_{j=1 \dots J} P(v_j)$ ,  $P(v_1 \dots v_j \dots v_J | s_k)$  可近似地改写为  $\prod_{j=1 \dots J} P(v_j | s_k)$ 。式中  $P(s_k)$ 、 $P(v_j)$ 、 $P(v_j | s_k)$  是训练语料库通过前面统计学习得到的概率,歧义词的贝叶斯概率的计算公式如下:

$$P_b(s_k | v_1 \dots v_j \dots v_J) = \frac{P(v_1 \dots v_j \dots v_J | s_k)}{P(v_1 \dots v_j \dots v_J)} P(s_k) \\ \approx \frac{\prod_{j=1}^J P(v_j | s_k)}{\prod_{j=1}^J P(v_j)} P(s_k)$$

### 1.2 歧义词的语义同现概率

训练用的标注语料库规模较小时,完全的贝叶斯消歧存在较严重的数据稀疏问题,训练结果对大规模真实文本的消歧成功率不高。若在歧义词词义选择时引入除贝叶斯概率外的其他因素,则可弥补训练用标注语料库规模太小给消歧带来的影响。本文将歧义词与上下文特征词在词条定义上的同现词数纳入影响词义选择的因素。

Lesk 认为,词典中词条本身的定义就可以作为判断其语义的一个很好的依据条件。现假设在机读词典中“ash”有 2

个词条定义,即  $S_1(\text{tree})$ : a tree of the oliver family,  $S_2(\text{burned stuff})$ : the solid residue left when combustible material is burned, 则如果 ash 的上下文特征词的词条定义包含 tree 或 burn, 那么有可能在其上下文特征词的词条定义中包含 tree 时选择语义 1, 包含 burn 时选择语义 2。

设  $D_1, \dots, D_k, \dots, D_K$  是测试文本句子中歧义词  $W$  的语义  $s_1, \dots, s_k, \dots, s_K$  的词典定义,在定义中它们被看成是一个可有重复的单词集,  $E_{v_j}$  表示在  $W$  的上下文出现的词  $v_j$  的词典定义,在  $v_j$  的定义中它也被表示为一个可有重复的单词集。为了简单起见,我们忽略了歧义词  $W$  的上下文词  $v_j$  的语义区别,规定如果  $v_j$  的语义定义是  $s_{j1}, \dots, s_{jJ}$ , 那么  $E_{v_j} = \cup_{j1} s_{j1} D_{j1}$ 。

对歧义词  $W$  的每一个语义项  $s_k (k = 1, \dots, K)$  的定义  $D_k$ , 计算出它与  $\cup_{v_j} = 1 \dots J E_{v_j}$  集合中同现的词个数  $tx_{kj} (k = 1, \dots, K; j = 1, \dots, J)$ 。令  $C_k = \sum_{j=1}^J tx_{kj} (k = 1, \dots, K)$ ,  $Count = \sum_{k=1}^K C_k$ , 则歧义词  $W$  在特定上下文窗口  $v_1, \dots, v_j, \dots, v_J$  中选择语义项  $s_k$  的语义同现概率为:

$$P_{tx}(s_k | v_1 \dots v_j \dots v_J) = \frac{C_k}{Count} (k = 1, \dots, K)。$$

### 1.3 消歧概率模型

词义消歧被看成是一个典型的词义分类过程,即对于  $s_1, \dots, s_k, \dots, s_K$  中的每一个语义项  $s_k (i = 1, \dots, K)$ , 计算出  $P(s_k | v_1 \dots v_j \dots v_J)$ , 选择使  $P(s_k | v_1 \dots v_j \dots v_J)$  取最大值的  $s_k$  作为歧义词  $W$  的最终词义。我们利用歧义词  $W$  的贝叶斯概率和语义同现概率对文本进行最大概率词义消歧,词义选择的概率评价公式如下:

$$P(s_k | v_1 \dots v_j \dots v_J) = \alpha * P_b(s_k | v_1 \dots v_j \dots v_J) + (1 - \alpha) P_{tx}(s_k | v_1 \dots v_j \dots v_J)$$

其中  $\alpha (0 \leq \alpha \leq 1)$  是贝叶斯概率的权重,  $P_b(s_k | v_1 \dots v_j \dots v_J)$  为歧义词  $W$  在特定上下文窗口  $v_1, \dots, v_j, \dots, v_J$  时选择语义项  $s_k$  的贝叶斯概率,  $P_{tx}(s_k | v_1 \dots v_j \dots v_J)$  为歧义词  $W$  在特定上下文窗口  $v_1, \dots, v_j, \dots, v_J$  时选择语义项  $s_k$  的语义同现概率。

$\alpha$  的取值可根据经验进行选择,原则上  $\alpha$  的大小与训练中使用的标注语料库的规模成正比。标注语料库规模越大  $\alpha$  的值就越大,反之亦然。这是因为在标注语料库的规模足够大时,完全的贝叶斯消歧方法本身就具有较高的消歧正确率。

## 2 词义消歧算法的实现

### 2.1 贝叶斯学习算法

首先,选定用于词义排歧的机读词典,按照机读词典的词条定义对训练用语料库进行词义手工标注,得到具有一定规模的标注语料库供统计学习使用。

其次,对标注语料库进行逐词扫描。设统计学习时歧义词上下文窗口大小为  $L$ , 现利用贝叶斯学习算法计算出  $P(s_k)$ 、 $P(c_l)$ 、 $P(c_l | s_k)$  的取值,其算法如下:

1) 变量初值:给定语料库的总词数  $T = 0$ , 每个词  $w$  在语料库中出现的总次数  $C(w) = 0$ , 每个词  $w$  的歧义标记  $Q(w) = 0$ , 词  $w$  为歧义词且词  $c_l$  出现在  $w$  的前  $\frac{L}{2}$  与后  $\frac{L}{2} - 1$  位置的次数  $T(w, c_l) = 0$ , 词  $w$  为歧义词且  $w$  的词义为  $S_k$  的次数  $M(w, S_k) = 0$ , 词  $w$  为歧义词且  $w$  的词义为  $S_k$ , 且词  $c_l$  出现在  $w$  的前  $\frac{L}{2}$  与后  $\frac{L}{2} - 1$  位置的次数  $Z(w, S_k, c_l) = 0$ ;

2) 读入词  $w$ , 读到的是否训练文本的结尾, 是: 转向步骤 5); 否:  $T = T + 1, C(w) = C(w) + 1$ ;

3)  $w$  词是歧义词吗? 是:  $Q(w) = 1$ , 对  $w$  的前  $\frac{L}{2}$  与后  $\frac{L}{2} - 1$  每一个词  $c_i$  有  $T(w, c_i) = T(w, v) + 1$ , 若该歧义词的词义为  $S_k$ , 则  $M(w, S_k) = M(w, S_k) + 1$ , 对  $w$  的前  $\frac{L}{2}$  与后  $\frac{L}{2} - 1$  每一个词  $c_i$  有  $Z(w, S_k, c_i) = Z(w, S_k, c_i) + 1$ ;

4) 转向步骤 2);

5) 根据上述步骤的结果  $T, C(w), T(w, c_i), M(w, S_k)$  和  $Z(w, S_k, c_i)$  计算出每一个歧义词  $W$  的  $P(s_k) = M(w, s_k) / C(w)$ , 歧义词  $W$  的词义为  $S_k$  时以  $W$  为窗口中心的每一个词出现的概率  $P(c_i | s_k) = Z(w, c_i, s_k) / M(w, s_k)$ , 每一个词出现在语料库中的概率  $P(c_i) = C(w) / T$ ;

6) 算法结束。

## 2.2 多义词消歧算法

设测试时歧义词上下文窗口的大小为  $J$ , 则在给定的测试文本上完成多义词的消歧的算法如下:

1) 读入词  $w$ , 读到的是测试文本的结尾, 则转向步骤 10);

2) 否: 在机读词典查找  $w$  的定义;

3)  $w$  词是歧义词吗? 否: 标注  $w$  词义, 转步骤 1);

4) 是: 提取测试言语歧义词  $w$  的前  $\frac{J}{2}$  与后  $\frac{J}{2} - 1$  个词  $v_1, \dots, v_j, \dots, v_j$  作为歧义词  $w$  的上下文特征词;

5) 根据贝叶斯学习的结果计算歧义词  $w$  的每一个词义  $S_k$  的贝叶斯概率  $P_b(s_k | v_1 \dots v_j \dots v_j)$ ;

6) 对  $v_1, \dots, v_j, \dots, v_j$  中的每一个  $v_j (j = 1, \dots, J)$  通过查找机读词典计算出  $E_{vj} = \cup_{ji} D_{ji}$ ;

7) 计算歧义词  $w$  的每一个语义  $s_1, \dots, s_k, \dots, s_K$  的词典定义  $D_1, \dots, D_k, \dots, D_K$  与  $\cup_{vj} = 1 \dots JE_{vj}$  的同现词数  $tx_{kj} (i = 1, \dots, K; j = 1, \dots, J)$ 。则歧义词  $w$  在特定上下文特征词  $v_1, \dots, v_j, \dots, v_j$  时选语义项  $s_k$  的语义同现概率为:

$$P_{wv}(s_k | v_1 \dots v_j \dots v_j) = \frac{\sum_{j=1}^J tx_{kj}}{\sum_{j=1}^J \sum_{k=1}^K tx_{kj}}$$

8) 根据概率评价公式计算出  $P(s_k | v_1 \dots v_j \dots v_j) (k = 1, \dots, K)$ ;

9) 求  $P(s_k | v_1 \dots v_j \dots v_j)$  的最大值  $(k = 1, \dots, K)$ , 用最大值的  $s_k$  作为歧义词  $w$  的最终词义。转步骤 1);

10) 算法结束。

## 3 实验与讨论

为了对消歧效果进行比较, 本文对 12 个多义词进行了测试。实验用语料部分选自于《人民日报》语料库, 多义词的词条定义来源于《现代汉语词典》。我们采用本文的学习算法和消歧算法进行了词义消歧的实验, 并与单纯的贝叶斯消歧方法进行了对比实验。实验结果的分布曲线如图 1 所示。

图 1 中有两条曲线, 上面的一条曲线代表本文方法在训练语料库不同规模时消歧的正确率, 下面的一条曲线代表完全贝叶斯消歧方法在训练语料库不同规模时消歧的正确率。从两条曲线的走势可以看出, 两种模型的测试正确率都随着训练语料库规模的增大而增加。完全的贝叶斯分类方法在语料库规模较小时, 消歧正确率明显低于本文方法。随着训练

语料规模的增大完全贝叶斯分类方法消歧正确率有较大的提高, 说明完全贝叶斯消歧方法在训练用的标注语料库规模较大时, 学习到的特征值较多, 消歧效果较好, 而本文方法无论在训练语料库规模较大还是较小时, 均能取得较高的消歧正确率。随着训练语料规模的增加, 本文方法消歧正确率的上升幅度开始趋于平缓, 与完全贝叶斯分类在消歧正确率上的差距逐步缩小, 说明本文方法主要用于弥补贝叶斯分类方法在标注语料库规模受限时因数据稀疏造成的学习到的特征不足的情况, 更适合标注语料库规模较小时的排歧。

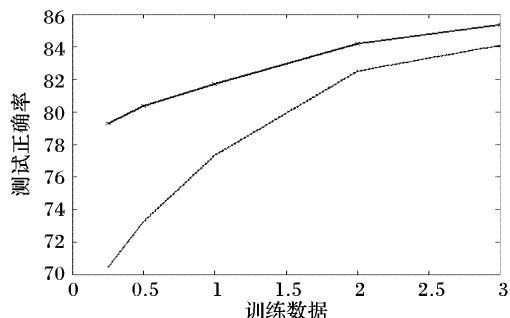


图1 测试上下文窗口 6, 贝叶斯概率权重 0.7

我们还对贝叶斯权重  $\alpha$  的取值进行了词义消歧实验, 实验结果的分布曲线如图 2 所示。

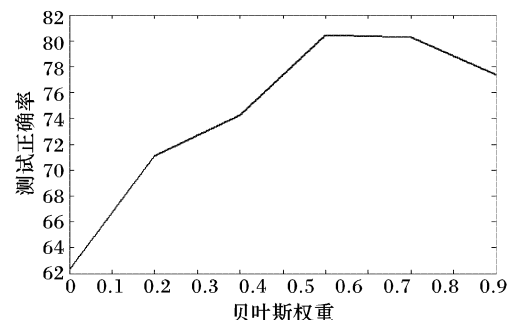


图2 测试上下文窗口=6, 训练数据为 10000

从图 2 中曲线的走势可以看出, 随着贝叶斯权重  $\alpha$  的增加, 本文方法在初始时消歧正确率有上升明显, 但达到峰值后, 正确率又有所降低。实验表明峰值随训练标注语料库规模大小而变化, 当训练语料为 10000 句时, 出现峰值的  $\alpha$  在 0.7 左右; 当训练语料为 5000 句时, 出现峰值的  $\alpha$  大约为 0.6。针对不同规模标注语料库进行的贝叶斯权重  $\alpha$  峰值实验显示贝叶斯概率对词义消歧的贡献在本文方法中仍然占据较为主要的地位。

## 4 结语

本文提出了一种基于贝叶斯分类与机读词典的多义词排歧方法, 建立了相应的概率模型并给出了排歧算法。该方法通过对小规模标注语料库的构建与学习, 结合传统机器词典的词条定义来完成消歧, 降低了获取大规模标注语料库的人工代价。实验表明基于本文方法的消歧算法在用于训练的语料库规模较小时, 消歧效果明显优于贝叶斯方法。

本文选择歧义词上下文窗口中的所有词作为消歧使用的特征词, 但上下文窗口中的虚词等一些无用信息在实际消歧时反而可能会引起噪声, 寻找词条定义的同现词时也存在类似情况, 噪声的存在对排歧准确率将产生较大的影响。如何过滤这些虚词和不具有实际意义的数字及字符, 有效地减少无效特征等噪声的影响, 进一步提高词义排歧的准确率, 有待于我们进一步研究和思考。

(下转第 1395 页)

下层数据源或其他 DISL-Mediator 中间件输出的数据进行集成或合成。应用体系的最高层为客户端应用或高层综合信息应用系统,DISL-Mediator 为客户端应用和其他 DISL-Mediator 提供了一个良好的接口,客户端应用可以通过远程接口引用存取 DISL-Mediator,动态改变它的运行行为(集成语义)或运行配置参数。

### 2.3.2 DISL-Mediator 的客户端应用框架

以下给出了客户端应用的基本框架和重要语句片断:

```
...import usc. DISLmediator. TEST_DISL_mediator. *
//引用特定的 DISL-Mediator 包
import usc. DISLmediator. DISLcore. *;
//引用 DISL 语言核心包
...
class ExampleDISLMediatorTest
{
    public static void main(String[] args)
    {
        //创建 DISL-Mediator 对象接口引用
        TEST_DISL_mediator exampleMediator
            = new TEST_DISL_mediator ();
        //通过对象引用与对象通信
        /* 通信语句格式为:
           exampleMediator. <语言组件名>. <方法名>([参数]);
        */
        ...
    }
}
```

## 3 结语

通过引入数据站、数据通道、操作站和流加工组件(过滤器/映射器/转换器)等抽象的语言成分组件或连接器,数据集成定义说明语言 DISL 支持从一个高度抽象的层次来表达复杂数据集成语义,有效解决异构数据源集成时面临的结构/语义异构冲突,从而为定义数据集成过程、自动生成数据集成中间件 Mediator 奠定了坚实的基础。与传统的 Mediator 方法相比,基于 DISL 定义说明并生成 DISL-Mediator 方法,具有以下一些优势:

1) 集成过程和集成语义定义说明简单直观,支持图形化方式自动生成;所生成的 Mediator 内部结构简明清晰、直观易理解,维护方便,具有良好的伸缩性和适应性,能很好适应演化的异构数据环境,完成复杂的集成任务。

2) 使用标准的简单关系模式作为内部工作的中介模式,对参与集成的数据源要求很低:所有支持 ODBC、OLE DB 或 JDBC 标准编程接口的数据源,以及 EJB 实体 BEAN 或其他遵循 J2EE 标准 CORBA 中间件的输出,都可以直接集成,不需编写专门的 Wrapper。

3) 提供了一种轻量级的数据集成方法,允许仅对部分数

据源或数据源中的部分数据子集进行集成,当数据源改变或添加新数据源时,维护和修改已有的 DISL-Mediator 简单容易。

4) 由 DISL 编译器生成的 DISL-Mediator 中间件对象(源码),是标准的 EJB 组件 Java Package,能平滑地应用于任何基于 Java 的中间件开发环境。DISL-Mediator 提供一套对其内部语言成分组件进行存取的良好接口,通过它,可以方便改变 mediator 在运行时的行为,或进行动态的参数配置。

目前,我们已在实际工程项目中,将该方法和数据仓库方法结合:通过将基于 DISL 图形平台生成的一组 DISL-Mediator 中间件对象,组装构建为一个数据集成工具,作为数据仓库数据增量更新的一个工具。实践收到了良好的效果。

致谢:山西省电力公司科技信息中心和太原市供电局,为本研究工作和相关辅助平台的应用提供了大力支持,提出了许多宝贵建议,在此表示感谢。

### 参考文献:

- [1] GRANT J, LITWIN L, ROUSSOPOULOS N, *et al.* An algebra and Calculus for Relational Multidatabase Systems[ A]. IEEE IMS'91 [ C]. 1991(4): 118 - 124.
- [2] WANG LC, ZHOU LZ, *et al.* A Multidatabase Integration Environment [J]. TSINGHUA Science and Technology, 1996, 1(2): 146 - 152.
- [3] BOUGUETTAYA A, PAPAOGLOU M, KING R. On building a hyperdistributed database[ J]. Informaion systems, an International Journal, 1995, 20(7) : 251 - 257.
- [4] MAH PS, CHUNG SM. Schema integration and transaction management for multidatabase [ J]. Information schience, 1998, 2(111) : 153 - 188.
- [5] SHETH AP, LARSON JA. Federated Database System for Managing Distributed, Heterogeneous, and Autonomous Database[ J]. ACM Computing Surveys( 1993 Special issues). 1993: 175 - 236.
- [6] BUSSE S, KUTSCHE R, LESER U, *et al.* Federated Information systems: Concepts, terminology and architectures[ R]. Berlin: Technishee University, 1999. 9.
- [7] WIDERHOLD G. Mediators in the architecture of future information systems[ J]. IEEE computer, 1992, 25(3) : 38 - 49.
- [8] CHAWATHE S, GARCIA-MOLINA H, HAMMER J, *et al.* The TSIMMIS project: Integration of heterogeneous information sources [ A]. In: Proc. Of the 10th Meeting of the Information Processing Society of Japn[ C]. 1994. 7 - 18.
- [9] 石祥滨, 张斌, 于戈, 等. 基于 CORBA 的异构分布信息系统 [J]. 小型微型计算机系统, 1997, 18(2): 37 - 43.
- [10] 王宁, 陈滢, 俞本权, 等. 一个基于 CORBA 的异构数据源集成系统的设计[ J]. 软件学报, 2000, 9(5): 178 - 392.
- [11] 孙志挥, 白义传, 等. 一种解决联邦数据库系统查询的模式转换方法[ J]. 计算机研究与发展, 1995, 32(2): 46 - 50.

(上接第 1391 页)

### 参考文献:

- [1] 冯志伟. 词义排歧方法研究[ J]. 术语标准化与信息技术, 2004, 22(01): 31 - 37.
- [2] YAROWSKY D. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods[ A]. In: Proceed Annual Meeting of ACL[ C]. Cambridge, Massachusetts, USA, 1995. 181 - 188.
- [3] 朱靖波, 李珩, 张跃, 等. 基于对数模型的词义自动消歧[ J]. 软件学报, 2001, 21(09): 1405 - 1408.

- [4] 全昌勤, 何婷婷, 姬东鸿, 等. 从搭配知识获取最优种子的词义消歧方法[ J]. 中文信息学报, 2005, 19(1): 30 - 35.
- [5] MICHAEL L. Automatic sense diaambiguation: How to tell a pine cone from an ice cream[ A]. In Proceedings of the 1986 SIGDOC Conference[ C]. New York, Association for Computing Machinery, 1986. 24 - 26.
- [6] MANNING CD, SCHUTZE H. 统计自然语言处理基础[ M]. 苑春法, 等译. 北京: 电子工业出版社, 2005.