文章编号: 1003-0077(2008)01-0116-07

情感语料库的构建和分析

徐琳宏,林鸿飞,赵 晶

(大连理工大学 计算机科学与工程系,辽宁 大连 116024)

摘 要: 本文介绍了情感语料库构建方面的一些经验,讨论了在设计和建设情感语料库中的几个基本问题: 制定标注规范、选择标注集、设计标注工具以及标注过程中的质量监控。目前已经标注完成近 4 万句,100 万字的语料。在完成这些已标注语料的基础上,进一步给出了语料库的情感分布,情感迁移规律等统计数据,分析了情感语料库的特点及应用。它的建成将为文本情感计算提供更加强大的资源支持。

关键词: 计算机应用;中文信息处理;情感语料库;文本编码规范;一致性检查;情感迁移

中图分类号: TP391 **文献标识码**: A

Construction and Analysis of Emotional Corpus

XU Lin-hong, LIN Hong-fei, ZHAO Jing

(Department of Computer Science and Engineering, Dalian University of Technology, Dalian, Liaoning 116024, China)

Abstract: This paper introduced some experiences on constructing emotional corpus, and discussed several basic questions which included the tagging criterion, tagging set, tagging tools and quality monitoring. There were about 40 000 sentences in the corpus. Moreover based on these, statistical data about emotional distribution and rules of emotional transference were available, and characters and applications of corpus were analyzed, so emotional corpus provide support for text affective computing.

Key words: computer application; Chinese information processing; emotional corpus; text coding initiative; consistency checking; emotional transference

1 引言

情感计算目前是人工智能领域的研究热点,它的主要目标是使计算机能识别人类的情感,也就是需要建立完善的情感识别模型。要使训练的模型准确,容错能力强,就必须有大规模的情感语料支撑。

在国外,语料库的研究很早就已经开始了,也建设完成了许多大规模的语料库,如 Brown 语料库等。汉语语料库的建设开始于 20 世纪 80 年代,现有的大规模语料有国家现代汉语语料库[1]、台湾中研院平衡语料库[2]、中港台汉语语料库[3]、北京大学和富士通公司共同制作的《人民日报》语料库[4]等。上述大规模语料库的建设在收集语料,制定标注规范和质量监控等方面积累了宝贵的经验。文本情感语料库的建设方面,目前已有的语料库包括 Pang 语料库[5],Whissell 语料库[6],Berardinelli 电影评论语料库[7],产品评论语料库[8]。汉语情感语料库标注方面的资源则较少,清华大学标注了部分旅游景点描述的情感语料[9],用来辅助语音合成,但规模也较小。总之,国内情感计算刚刚兴起,这方面还没有比较大规模、权威的汉语文本情感语料库。

大部分语料库的建设分为语料的收集和预处理、标注规范的制定,质量监控等几方面,下面的论

收稿日期: 2007-05-20 **定稿日期**: 2007-12-01

基金项目: 国家自然科学基金资助项目(60373095,60673039);国家 863 高科技计划资助项目(2006AA01Z151);教育部留学回国人员科研启动基金资助项目

作者简介: 徐琳宏(1979—),女,硕士生,研究方向为文本分类和文本倾向性识别;林鸿飞(1962—),男,博导,教授,研究方向为文本过滤,文本挖掘和自然语言理解;赵晶(1961—),女,硕士,讲师,研究方向为文本可视化和图形图像处理。

文将分别阐述语料库建设的各个步骤。第 2 节概略的介绍了目前选择语料的类型和规模,第 3 节详细地介绍了情感语料库的标注体系,第 4 节介绍了语料建设中质量监控的方法,包括正确性和一致性检查的方法。第 5 节阐述了语料库的一些统计数据及应用,最后,第 6 节总结语料库的优点和不足,并进一步提出改进的措施。

2　语料的收集

语料的收集工作,即选择合适的语料,做预处理,为语料的标注提前做好准备。语料选择的方法关系到语料库的覆盖率,所谓覆盖是指语料在各个不同领域的分布或散布,这些不同领域通常是指由时间轴(反映时代特征)、空间轴(反映地域特征)、学科轴(反映知识特征)、风格轴(反映语体特征) 构成的四维模型[10]。

我们的语料包括小学教材(人教版)、电影剧本、童话故事、文学期刊等。从时间轴上看,有童话故事和小学教材等完成较早的经典文章,也有期刊和电影剧本等近一年多的作品。语料以中文的作品为主,但是也有部分电影剧本和童话故事是外文翻译而来,考虑了地域特征的跨度。在风格方面,小学教材等用词比较规范、严谨,而电影剧本等则口语特征比较明显。总的来说,语料的选择偏重于文学色彩比较浓,情感表达丰富多彩的作品,舍弃一些科学说明性的文章。表 1 列出了各类语料的详细信息。

表 1　语料的详细信息

语料来源	详 细 说 明	字数	词数	句子数	篇章数
小学教材	人教版,12 册	129 486	91 032	4 809	171
电影剧本	《狮子王》、《汽车总动员》等 6 个电影剧本	84 118	54 092	5 911	237
童话故事	部分格林童话、安徒生童话	54 066	39 005	2 011	73
文学期刊	《少年文艺》、《青年文摘》、《新青年》等 9 本期刊的 2006 年全年 12 期	6 308 526	4 375 396	237 290	3 754
总计		6 576 196	4 559 525	250 021	4 235

3　情感语料库的标注体系

语料库的标注体系就是指对语料的加工程度,即一个待标注的单元需要填充的信息集合。标注体系决定了语料标注的粒度。如果类别划分过粗,就不能全面、细致地描述语言的复杂现象;但如果类别划分过细、标注信息过于庞大,不但会增加标注难度、降低标注效率,关系之间只有细微差别的情况也会使标注结果呈现严重的不一致性[11]。此外,在语料库规模有限的情况下,类别分的太细,统计数据的稀疏问题越严重,那么训练出来的模型健壮性就越差。可见,语料库的标注体系是构建一个高质量、大规模语料库的关键。

3.1　情感标注体系

理想的情感标注体系是在标注前事先确定,在标注过程中保持不变,这样可以保证标注的一致性。但是由于语料的多样性和复杂性,标注规范也需要多次修正,这就可能导致语料库的质量下降。为了充分考虑各种特殊情况,本文预先标注了部分语料,在总结标注中发现问题的基础上,综合考虑其他类型语料的标注经验和文本情感标注自身特点,制定了如下的标注体系:

$$DocumentModel = (title, author, style, source, persons, sentences, keynote) \quad (1)$$

$$SentenceModel = (origin, sender, [accepter], [rhetoric], emotions, [keywords]) \quad (2)$$

由上面两个公式可以看出本文的情感标注体系的标注粒度分为词汇、语句和篇章。其中语句是主要的情感标注粒度,词汇和语篇的相关信息都是语句情感标注的辅助。方括号内的变量 *accepter*、*rhetoric* 和 *keywords* 是可选的,其他的是不能为空的。语篇和语句标注模型中各变量表示含义和取值范围如表 2 所示。

在上述变量中 *persons*、*sentences*、*emotions* 和 *keywords* 取值都是一个集合,即变量的取值可以表示为一个向量,如 $persons = (persona_1, persona_2 \cdots; persona_i \cdots)$,变量 *sender* 和 *accepter* 分别选择 *persons* 中的一个分量作为变量值。需要说明的是 *persons*

表 2 标注体系中各变量的说明

类别	变量	说明	取值范围
语篇标注模型 (documentModel)	*title*	文章题目	
	author	作者	姓名,国籍,作品写作年代
	style	类别	散文| 诗歌| 小说| 戏剧
	source	来源	小学教材| 格林通话| 电影剧本| 文学期刊
	persons	情感主体	主人公$_1$| 主人公$_2$ …| 主人公$_i$ …
	sentences	所有语句的标注集合	详见 sentenceModel
	keynote	情感基调	o| h| e| i| m| f| d| s
语句标注模型 (sentenceModel)	*origin*	原始语句	
	sender	本句的情感主体	主人公$_i$
	accepter	情感的接受者	主人公$_i$
	rhetoric	修辞类别	比喻| 比拟| 借代| 夸张| 对偶| 排比| 设问| 反问| 重复
	emotions	本句包含的所有情感	o| h| e| p| r| b| l| k| c| i| s| w| g| m| u| f| x| t| d| a| j| y| q
	keywords	确定情感的关键词	词$_1$| 词$_2$ …| 词$_i$ …

中包含两个特殊的情感主体,“旁白”和“其他”。“旁白”表示该句是作者的叙述,没有鲜明的情感发出人,而“其他”是为了处理当一篇文章中涉及的任务较多时,所有非主要人物发出的情感都用它代替,这样可以减轻标注者的负担,又能防止某个情感主体出现次数较少的数据稀疏问题。变量 *sentences* 是所有语句情感标注的集合,每个语句标注的内容就是语句标注模型中声明各个变量。*keywords* 中的值是原始语句中对表达该句情感有决定作用的词,标注 *keywords* 是为了更准确地确定语句中代表情感的词汇。而实验证明,情感词汇的特征在语句的情感自动标注中是一个区分度较大的特征[12]。另外,否定词和程度副词对句子情感色彩影响也较大,特别是对语句的褒贬倾向性影响较大[13],但是本文的标注体系没有标注这两方面的信息,这主要是为了提高标注效率,所以没有列入标注体系。变量 *keynote* 取图 1 中的所有分支节点。变量 *emotions* 的取值是由图 1 叶子节点中的一个或几个组成的向量。

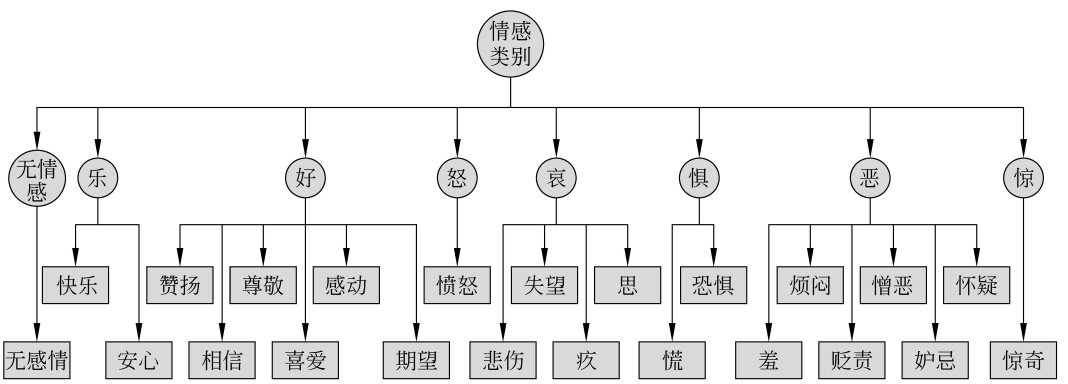

图 1 情感分类树

3.2 基于 TEI 的标注集选择

选择标注集就是选择合适的标注附码和便利的表示方式来存储标注后的语料。英国著名语言学家 Leech 是当今语料库语言学的代表人物之一,他认为(1993) 语料的标注应该遵循标注附码可以删除;所作的标注可以单独抽出;任何标注模式都不能作为第一标准等七个基本原则。本文在综合考察已有的各种标注集优缺点的基础上,结合自己语料库的实际应用情况,以半结构化的方式表示已标注的文本。

本文标注集的选择是在 TEI(Text Encoding Initiative)的基础上,结合情感标注的特殊需求制定的。TEI 是机读语篇的国际信息编码规范。TEI 标注模式是由计算语言学学会(ACL, Association for Computational Linguistics)、文学与语言学计算协会(ALLC, Association for Literary and LinguisticComputing) 和计算机与人文科学学会(ACH, Association for Computers and

Humanities) 等三家学术团体共同参与制订的。“英国国家语料库”(The British National Corpus) 等许多大型语料库都采用了 TEI 的标注模式。根据 TEI 标注模式, 一篇语料分为篇头(Header) 和篇体两部分。篇头指与语篇有关的背景信息, 包括作者、标题、日期、语篇来源、标注方式等信息, 而篇体是指语篇本身。在 TEI 标注模式中语言单位可以是词、句子或段落等, 每个语言单位都有起始标记(Start tag) 和结束标记(End tag)。例如, 段落的开始和结束标记分别为〈p〉和〈/ p〉。

```
<document>
  <header>
    <title>冬天的里的父亲</title>
    <author>
      <name>贤　惠</name>
      <country>中国</country>
      <year>2006</year>
    </author>
    <style>散文</style>
    <source>新青年</source>
    <persons>
      <persona>我</persona>
      <persona>父亲</persona>
      <persona>其他</persona>
      <persona>旁白</persona>
    </persons>
  </header>
  <body>
    <p>
      <sect>
        <orgin>那时候，我的家在茫茫林海的大兴安岭北部，那是一个偏远的小镇。</orgin>
        <sender>旁白</sender>
        <emotions>
          <emotion>o</emotion>
        </emotions>
      </sect>
      <sect>
        <orgin>如今已离开那里许多年了，留在记忆之中最难忘的，不是冬天的雪原，不是五;
```

图 2　语料标注示例

在本文的标注集中通用的信息,如篇头,段落等采用 TEI的标记规范。另外定义一些标签来标记情感标注中特有的信息,标签的定义以简洁、易懂为原则。图 2 是一篇语料的部分标注示例。整篇语料在〈document〉和〈/ document〉之间,〈header〉和〈/ header〉之间的是篇头部分,〈body〉和〈/ body〉中的为篇体部分。〈p〉和〈/ p〉分别为段落的开始和结束标记,〈sect〉和〈/ sect〉为语句的标记。上面的标记模式一方面可以从 title 和 orgin 域中还原出原始语料,另一方面也可以从每句的 emotions 域中得到语篇或者段落的标记序列。这基本符合 Leech 的标注附码可以删除和标注可以单独抽出的几个重要原则。另外,这种半结构化的存储方式使每个标注单元都有开始和结束标记,与 xml 格式类似,也为训练模型时解析语料提供了方便。

4　语料库的质量监控

本文的情感语料库的质量监控主要从标注规范、标注系统和纠错机制三个方面完成。

4.1　标注规范及标注系统

标注规范和标注系统都是在语料的标注过程中减少误操作,提高标注速度和增加一致性的有效措施。统一的标注规范,可以有效缩小不同标注者之间的差异,减少语料标注中的错误和不一致性。情感语料标注的规范是在建设的过程中动态更新的,规范的部分内容如下:

· 在前后句情感主体相同的条件下,各句的情感具有连续性。例如,若连续的三句话都是同一个情感主体发出的,而第 2 句有明显的“快乐”类标记,则第 1 句和第 3 句没有太明显的情感类别时,也倾向于快乐。

· 每句的关键词是广义范围的词汇,可以是词汇或者常用短语,但是不能扩大到一个分句。

· 除了关键词、修辞类别和情感接受者,其他内容都是不能为空的。

· 一个句子可以包含多个情感,但是同一个句子不能同时标记为无情感和其他 23 类中的任何一个。

· 当文章没有清楚的说明作者时,填写“不详”代替。

· 每篇文章的情感主体除了主人公外,还有“旁白”和“其他”两类特殊的情感主体。“旁白”表示该句是作者的叙述,没有鲜明的情感发出人,而“其他”是为了处理当一篇文章中涉及的任务较多时,所有非主要人物发出的情感都用它代替。

全面的标注规范可以减少语料的不一致,而方

便、高效的标注系统可以大幅度提高标注的效率和准确性,防止标注者的误操作。图 3 是情感语料标注系统的界面,"情感主体"以上的部分是描述语篇的信息,接下来的部分是标注语句情感的,从最下面的文本框中可以浏览整篇文档。为了减轻标注者的负担,提高标注速度和准确率,该系统采用启发式搜索算法[14]自动分割语句,并根据某些项不能为空的规范自动完成合法性检查,防止错误的语料进入语料库。

图 3 情感标注系统

4.2 纠错机制

标注规范和标注系统是保证语料在录入时的准确率和一致性,而纠错机制是在语料标注完成后统一进行语料的正确性和一致性检查。

为了统一标注者在某些常见情况的标注标准,我们采用了许多大规模语料库常用的方法,即做部分的交叉标注,保证语料标注的正确性。在一致性检查方面本文采用的纠错机制是机器自动检查,人工修正的方法。根据情感语料标注的特点,本文从词汇和情感连续性两个角度分析标注的一致性,为了清楚的介绍这部分内容,首先说明这部分相关的函数和变量,具体见表 3。

表 3 一致性检查的部分函数说明

函数名	自变量	说明	取值	条件
Neg	S_i	第 i 个语句中是否包含否定词	0	没有否定词
			1	包含否定词
larSame	E_i, E_j	第 i 句和第 j 句的情感在大类范围内(情感分类树的分支节点)是否相同	0	不同
			1	相同
wordSame	S_i, S_j	第 i 句和第 j 句是否包含相同的关键词	0	不同
			1	相同
personSame	S_i	第 i - 1, i 和 i + 1 句情感主体是否相同	0	不同
			1	相同
emotionSame	E_i	第 i - 1, i 和 i + 1 句情感是否相同	0	不同
			1	相同

从情感词汇的角度出发考虑一致性,主要以关键词为依据,检查一致性。

$$wordConsistency(i,j) = wordSame(S_i,S_j) \cap \overline{larSame(E_i,E_j)} \cap \overline{Neg(S_i)} \cap \overline{Neg(S_j)} \quad (3)$$

公式中 S_i 和 S_j 分别表示一篇语料中的第 i 句和第 j 句,E_i 和 E_j 分别表示第 i 句和第 j 句的情感。$wordConsistency$ 表示当两句中都不包含否定词时,如果两句的关键词相同,但是所属的情感大类不同时,两句可能存在不一致,此时取值为 1。

从情感的连续性上考虑,当前后句的情感不一致,但是情感主体相同的条件下,该句的情感可能存在错误。具体见公式(4):

$$contextConsistency(i) = personSame(S_i) \cap \overline{emotionSame(E_i)} \quad (4)$$

上述的两个公式分别从词汇和情感连续性两个方面检查情感的一致性,通过机器自动识别出不一致的地方,再人工确认是否需要修改。两种方法虽

然都是进行一致性检查,但是关键词方法的一致性错误级别较高,需要优先确认。而情感连续性方面的一致性检查,则只是说明有出现不一致的可能,但是不一定都是错误。

5　语料库的统计数据及应用

5.1　语料库的统计数据

目前已经标注完的语料有 1 035 601 字,726 605 词次,39 488 句。这是情感语料库第一期计划完成的语料,第二期完成后预计标注的总量将达到一千万字。

5.1.1　语句的情感分布

在 39 488 句中,标注的各类情感所占的比例大致分为三个等级。其中标注为"无情感"类的语句数最多,达到 15 449 句,其次是"快乐"、"赞扬"、"烦闷"和"怀疑"四类情感数较多,都超过 2 000 句,其余各类情感均在 1 000 句左右。

5.1.2　情感迁移规律

情感迁移规律是指在语句的上下文中,情感的接续概率,即由一种情感向另一种情感(包括转移前情感)迁移的可能性。本文通过公式(5)计算情感迁移的概率:

$$transfer(E_a, E_b) = \frac{2 \times \sum_{i}^{n} \sum_{j}^{m_i - 1} (equal(E_j, E_a) \times equal(E_{j+1}, E_b))}{T_a + T_b} \quad (5)$$

$transfer(E_a, E_b)$ 表示由情感 a 向情感 b 迁移的概率,n 表示语料库中语篇的总数,m_i 表示第 i 篇文档的句子总数,T_i 表示语料库中被标记为 i 类情感的句子总数。当 E_i 与 E_j 相同时,函数 $equal(E_i, E_j)$ 取值为 1,否则取值为 0。将 a 类和 b 类情感的总数作为分母是为了减弱各类情感包含的语句数量不同给情感迁移带来的影响。公式主要计算语篇范围内,上下句之间的情感变化。

图 4 是 23 类情感之间的迁移概率图,因为 23 类情感彼此的迁移可能性比较多,为了表示的更加清楚明晰,本文在图中给出了 $transfer(E_a, E_b)$ 大于等于 0.05 的情感迁移概率。由图可以看出"哀"类情感的内聚性(情感大类内的情感迁移)较弱,而"恶"类情感的内聚性较强,"惧"和"好"类情感的内聚性一般。

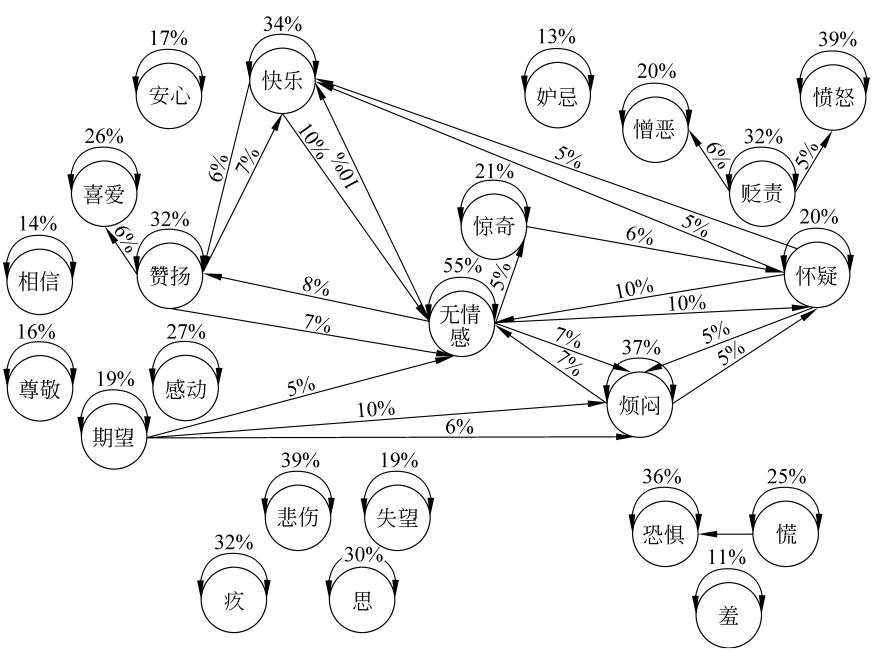

图 4　情感迁移图

5.2　语料库的应用

语料库的标注内容和标注形式决定了它的应用范围。目前情感语料库主要应用在训练文本情感识别模型、情感词汇本体的自动学习和统计情感迁移规律三方面。按句标注的情感不仅给出了情感的类别,而且标注了情感主体、关键词和修辞手法等信息,这些都为情感识别模型的训练提供了丰富和区分度较高的特征,为提高情感识别的准确率奠定了基础。每句在情感标注过程中都尽可能标记了关键

词,这些关键词为情感词汇本体的自动学习提供了第一手的资料。文本情感的迁移规律不同于脸谱和语音的情感迁移,它有其自身的特点。通过统计语料库中种数据,可以得到类似图 4 的情感迁移规律图。

6 结论及改进措施

情感语料库在建设过程中从制定标注规范,选择合适的标注集以及质量监控等多方面提高语料标注的质量和速度。目前已标注完成的语料有 1 035 601 字,39 488 句,第一期标注的语料已经基本完成。在总结第一期标注经验的基础上,计划完成 10 000 句,近千万字的语料。任何语料库的建设都不可能是完美无缺的,肯定会存在一些问题和不足。情感语料库的建设也存在语料在体裁和情感类别上分布不均以及参考的标注建议较少等缺点,我们将在今后的建设中不断改善。

参考文献:

[1] 刘连元. 现代汉语语料库研制[J]. 语言文字应用, 1996, (3):2-9.

[2] http://www.sinica.edu.tw/SinicaCorpus/[DB/OL].

[3] 胡百华, 李行得, 汤志祥. 香港的语料库和相关研究概况[J]. 语言文字应用, 1997, (2):49-54.

[4] http://www.icl.pku.edu.cn/icl_groups/corpustagging.asp[DB/OL].

[5] http://www.cs.cornell.edu/People/pabo/movie-review-data/[DB/OL].

[6] Theologos Athanaselis, Stelios Bakamidis, and Ioannis Dologlou. Recognizing Verbal Content of Emotionally Colored Speech [A]. European Signal Processing Conference[C]. 2006.

[7] http://www.reelviews.net/[DB/OL].

[8] http://epinions.com/[DB/OL].

[9] Hongwu Yang, Helen M. Meng, Zhiyong Wu and Lianhong Cai. Modeling the Global Acoustic Correlates of Expressivity for Chinese Text-to-Speech Synthesis [A]. IEEE/ACL 2006 Workshop on Spoken Language Technology[C]. Aruba, 2006. 10-13.

[10] 张普. 关于大规模真实文本语料库的几点理论思考[J]. 语言文字应用 1999, (1):34-43.

[11] 周明. 面向语料库标注的汉语依存体系的探讨[J]. 中文信息学报,1994,8 (3):35-51.

[12] 徐琳宏,林鸿飞. 基于语义特征和本体的语篇情感计算[J]. 计算机研究与发展,2007, 44(S2):356-360.

[13] 徐琳宏,林鸿飞,杨志豪. 基于语义理解的文本倾向性识别机制[J]. 中文信息学报,2007,21(1):96-100.

[14] Christopher D. Manning, Hinrich Schutze. 统计自然语言处理基础[M]. 北京: 电子工业出版社, 2005. 82-83.

(上接第 66 页)

[8] Berger, A. and Lafferty, J. D.. Information Retrieval as Statistical Translation [A]. In: proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval [C]. 1999, 222-229.

[9] Andreas Hotho, Steffen Staab, and Gerd Stumme. Ontologies improve text document clustering [A]. In: Proc. of the ICDM 03, The 2003 IEEE International Conference on Data Mining, 2003. 541-544.

[10] Zhou, X., Zhang, X., and Hu, X., The Dragon Toolkit, Data Mining & Bioinformatics Lab, iSchool at Drexel University, http://www.ischool.drexel.edu/dmbio/dragontool [CP/OL].

[11] Zhou X., Hu X., Zhang X.. Using Concept-Based Indexing to Improve Language Modeling Approach to Genomic IR [A]. ECIR 2006 [C]. LNCS 3936, 2006. 444-455.

[12] J.C. Bezdek. Pattern Recognition with Fuzzy Objective Function Algorithms [M]. Plenum Press, New York 1981.

[13] Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society, 1977, 39: 1-38.

[14] http://ir.ohsu.edu/genomics/data/ [DB/OL].

[15] Hersh W, et al. TREC 2004 Genomics Track Overview [A]. the thirteenth Text Retrieval Conference [C]. 2004.

[16] Hersh W, et al. TREC 2005 Genomics Track Overview [A]. the fourteenth Text Retrieval Conference [C]. 2005.

[17] 张俊林,孙乐,孙玉芳. 基于主题语言模型的中文信息检索系统研究[J]. 中文信息学报,2005,19(3):14-20.