文本信息处理研究述评

袁鼎荣[1,2] 钟 宁[1] 张师超[2]
(北京工业大学国际WIC研究院 北京 100124)[1]
(广西师范大学计算机科学与信息工程学院 桂林 541004)[2]

摘 要 文本信息处理就是通过计算机对文本从表及里、由此及彼的分析处理,不仅仅抽取包含其中的信息,更需要分析推理蕴涵其中的意义。全面地分析探讨了文本信息处理研究现状,概述了文本信息处理的发展历史,将文本信息处理研究归纳为分词研究、文本信息抽取、文本分类、文本信息检索、文本自动摘要等方面,并分别对各领域的研究现状做了概述,指出了各研究领域中存在的问题。讨论了文本信息处理的关键技术问题及其挑战,指出了文本信息处理的远景目标就是文本信息的语义分析、归纳推理和文语转换。

关键词 分词研究,文本分类,信息抽取,信息检索,文本自动摘要
中图法分类号 TP391.1 **文献标识码** A

Research on Text Information Processing: Review

YUAN Ding-rong[1,2] ZHONG Ning[2] ZHANG Shi-chao[1]
(The International WIC Institute, Beijing University of Technology, Beijing 100124, China)[1]
(College of Computer Science and Information Technology, Guangxi Normal University, Guilin 541004, China)[2]

Abstract The object of text information processing is to analyze and transact information about text from superficial to interior, here to there. It includes extracting information from the text file, getting some meaning hidden in the text by analyzing and mining it. We discussed the work about text information processing in detail, and summarized the history of text information processing. We summed up text information transaction into word segmentation, information extracting from text, text classification, text information retrieval, automatically summary from text etc, and summarized the correlation research work and pointed out the problem in the fields. At last we discussed the key technology and challenge in the text information processing and pointed out that the object is text semantics analysis, induction, deduction and conversion between text and language.

Keywords Word segmentation research, Text classification, Information extracting, Information retrieval, Automatically abstracting

1 引言

信息是帮助人们了解、理解、掌握客观事物或对象的知识,信息的载体有文本、图形、图像、声音、视频和动画。随着信息技术的发展,存储在各种媒介上的信息迅速膨胀。所有这些信息中,文本信息最通用,即使图形、图像,甚至音频、视频都辅以一定文字说明,所以文本信息处理研究显得尤其重要。近50年,文本信息处理吸引了许多机构及其研究工作者的高度重视。我们从文本信息处理的发展历史、研究现状展开论述,将文本信息处理归纳为文本分词、文本信息抽取、文本分类、文本信息检索、文本自动摘要等方面。最后讨论了文本信息处理的挑战、前景和关键技术。

2 文本信息处理的发展史

20世纪60年代开始直到80年代,美国纽约大学开展的Lingulsticstring项目[1-3]的主要研究内容是建立一个大规模的英语计算语法。与该研究相关的应用是从医疗领域的X光报告和医院出院记录中抽取满足一定格式的信息(InformationFormatS),这种格式化信息就是文本信息抽取的最终输出结果,现在被称为模板(TemPlates)。它开创了文本信息处理的先河。

20世纪70年代,美国耶鲁大学的Roger Schank及其同事所进行的长期项目是有关故事理解的研究,其学生Gerald De Jong设计实现的FRUMP系统[1,4]是根据故事脚本理论建立的一个文本信息抽取系统。该系统从新闻报道中抽取信息,内容涉及地震、工人罢工等很多领域或场景。该系统采用了期望驱动(top-down,脚本)与数据驱动(bottom-up,输入文本)相结合的处理方法。这种方法被后来的许多信息抽取系统采用[2,3]。

20世纪80年代,信息处理技术得到蓬勃发展,它主要得

到稿日期:2010-03-26 返修日期:2010-08-27 本文受国家自然科学基金重大研究计划培育项目(90718020),澳大利亚ARC项目(DP0667060),广西自然科学基金资助。

袁鼎荣(1967-),男,硕士,副教授,主要研究方向为数据挖掘、机器学习,E-mail:dryuan@mailbox.gxnu.edu.cn;钟 宁(1956-),男,教授,博士生导师,主要研究方向为机器学习、数据挖掘、网络智能;张师超(1962-),男,教授,博士生导师,主要研究方向为数据挖掘、机器学习。

益于一系列关于文本信息处理的国际会议。比如系列消息理解会议(MUC,Message Understanding Conference)、文本检索会议(TREC,Text Retrieval Conference)、自动内容抽取(ACE,Atomatic Content Extraction)评测会议等[3]。

MUC 会议从 1987 年开始到 1998 年,共举行了 7 届,由美国国防高级研究计划委员会(DARPA,the Defense Advanced Research Projects Agency)资助。MUC 的任务就是评测信息抽取系统。每次 MUC 会议前,组织者先规定样例、消息文本并给出抽取任务的说明,参加者开发能够按照规定的任务处理这种消息文本的信息抽取系统。在正式会议前,对各参加者开发的系统进行测试,系统的输出结果与手工标注的标准结果相对照,得到最终的评测结果。会议中主要是由参与者交流思想和感受[2,3]。

1992 年开始的 TREC 会议是 MUC 会议模式的扩展。TREC 会议由美国国家标准技术局(National Institute of Standards and Technology,NIST)和国防部高级研究计划局(Defense Advanced Research Projects Agency,DARPA)组织召开。它的主要宗旨是促进技术交流,加速技术的产业化,发展对文本检索系统的评测技术。参加单位包括许多著名的大学和公司,还有不少美国以外的文本检索领域的研究团体。TREC 不仅提供了一个标准文档库,而且提出了一套较为科学的测试评价方法,为各种方法和系统提供了一个公平竞争的舞台,使 TREC 成为文本检索领域最权威的国际评测会议[5]。

1999 开始的、获得美国国家标准技术研究所支持的(NIST)ACE 评测会议,每年举办一次,旨在开发自动内容抽取技术,以支持对 3 种不同来源(普通文本、由自动语音识别 ASR 得到的文本、由光学字符识别(OCR 得到的文本)的语言文本的自动处理,研究的主要内容是自动抽取新闻语料中出现的实体(Entitie)、关系(Relation)、事件(Event)等内容,即对新闻语料中的实体、关系、事件进行识别与描述[6]。

国内对汉语文本信息处理的研究,近年来也受到足够的重视。20 世纪 90 年代北京语言文化大学和清华大学出台了《现代汉语语料库文本分词规范》。教育部语言文字应用研究所出台了《信息处理用现代汉语词类标记规范》,《信息处理用现代汉语分词规范》(国家规范 GB/T13715-92)上升到国家层面[7]。1991 年首次对汉字识别进行测试。2003 年 3 月,由 ACL-Sighan 主办,举行了第一届国际汉语分词评测(The first international Chinese word segmentation bakeoff)[6,8],同年 10 月由 863 中文与接口技术评测组组织对其进行测评。该会议从 2005 年开始每年主办一次,现在受到越来越多的重视。

3 研究现状

文本是历史悠久、应用广泛、使用灵活、认可度最高的信息载体。文本信息处理研究主要包括分词研究、文本信息抽取、文本分类、文本信息检索、文本自动摘要等方面。所有这些研究的研究对象就是文本字串。

3.1 分词

定义 1 文本的最基本要素是字,但单一的字索然无味。符合语言习惯并能表达一定意义的字与字间的连接构成的词才是文本的有机要素。将一篇文档切割成不同的词称为分词。任何词在句子中都具有一定的词性,如字符串 $S=c_1c_2\cdots c_i\cdots c_n$,其中$(c_i,i=1,2,\cdots,n)$为字符,切分后 $S=w_1w_2\cdots w_i\cdots w_p$,其中 $w_i(i=1,2,\cdots,p)$为 S 的切分结果,对应切分的词性表示为 $T=t_1t_2\cdots t_i\cdots t_p$ 其中$(t_i,i=1,2,\cdots,p)$为相应切分词的词性。

分词技术是文本信息处理的前提和基础,它大致可分为基于词典分词、无词典分词两种分词方法。

3.1.1 词典分词

依据一定的规范构造一个词典 $D=\{d_1,d_2,\cdots,d_n\}$,d_i 为词典中的词,然后按照某种形式切分文本,切分所得的词记为 w,w 的词性记为 t_w,$t_w\in T$,其中 T 为词性集合。

基于词典分词技术的问题来自两个方面,一方面,文本语料中的词灵活多变,不同的切分方法产生不同的、属于词典的词。比如"搜索引擎",可以分解为"搜索"、"索引"、"引擎"、"搜索引擎",此时就产生切分歧义,文献[10-12]专门对歧义消解做了一定的研究。而汉语中的切分歧义是分词技术的关键,歧义消解是精确把握文本语义的基础,而歧义消解技术依然是至今未很好解决的问题。另一方面,这种基于词典的机械分词法,实现简单,实用性强,但缺点就是词典的完备性不能得到保证。据文献[9]统计,用一个含有 70000 个词的词典去切分含有 15000 个词的语料库,仍然有 30%以上的词条没有被分出来,这样就需要在查找词典的同时还需要充实词典,其计算复杂度指数增加。

3.1.2 无词典分词方法

由于词典的不完备性,近年来在依据统计学原理进行分词方面做了不少工作。文献[13-20]通过统计切分所得词条的频率,依据频率特性确定所得词条是否为真实词条。文献[16,17,22]在统计词频的基础上,进一步考虑词条间的信息熵或互信息确定候选词条是否为真实词条。我们称这种分词技术为基于词频统计的分词技术。文献[23]用多达 9 种不同的统计量比较考察不同情况的分词效果。基于统计学的分词技术的问题主要在于机械地对文本进行切分,得到的候选词条是真实词条的近 10 倍,而且存在许多噪声,即很多高频词是非真实词条。

由于基于统计的无词典分词技术的机械性和噪声,在分词过程中就转而考虑文本的语义语法及规则等特性,希望用文本的语义和语法规则进行分词。文献[24]利用文本的篇章信息及带有频度的边界模板切分可能的人名。文献[25]利用文本中的上下文信息进行分词。考虑字与字间的结合密切程度进行分词。由于语义语法等规则灵活多样,基于语义语法文本分词技术的实用性不强,可扩展性较差。一般情况下,基于规则的分词技术所使用的规则来源于特定语言环境,从而特定了它的使用局限性。

3.1.3 组块分析

分词技术经过数十年的发展,取得了长足的进步,但词典的不完备性、切分歧义性、噪声以及语义模板的局限性一直制约着高效而精确的分词系统的开发。因而就转向更大粒度层次的分词技术研究,比如将句子分成不同的主、谓、宾等主要成分,连接共同构成主语或其它语言成分的词群称为组块,根据句子结构分成不同的组块进行分词[17]。

3.2 文本信息抽取

定义 2 从指定文本或文本集中,抽取满足一定形式或合乎一定内容要求(包括文本中的实体、关系和事件)的信息的过程称为文本信息抽取。

文本信息抽取有指定的抽取对象、明确的抽取内容，早在20世纪60年代美国纽约大学开展的Lingulsticstring项目就是从医疗领域的X光报告和医院出院记录中抽取满足一定格式的信息(Iformation FormatS)。20世纪70年代，美国耶鲁大学的Roger Schank及其同事所进行的长期项目是有关故事理解的研究，希望从新闻报道中抽取涉及地震、工人罢工等很多领域或场景的信息。后面得以蓬勃发展的文本信息抽取和文本内容理解的系列会议旨在推动人们对文字语言使用、理解的人工智能化。经过数十年的研究发展，文本信息抽取从许多方面展开研究。文献[28]根据语料库、文献[29]从统计、文献[30]从角色标注、文献[27]从可信度、文献[31]从边界模板和局部统计分别对中文人名识别与抽取做了一定研究，试图抽取包含在中文文本中的人名。文献[32,33]对中文名词短语的识别做了细致的研究，文献[34,35]分别对网络上新出现的词语识别抽取做了一定工作。文献[36-42]对文本中的关键词的识别做不少工作，其中文献[37]建立了一套如何识别包含在文本中的实体、关系、事件等的关键要素。

尽管目前的文本信息抽取取得不少建设性成果，比如医学记录中的信息抽取、文本中人名、地名等关键字的提取都达到一定效果，但所有这些系统都局限在某一特定应用领域，不具备扩展性和广泛的实用性，尤其难以获取文本中字里行间所蕴涵着的信息。这一问题的解决不仅受计算模型的限制，更受到文本分词技术的精确度的限制。这将是今后数年乃至数十年文本信息抽取工作者所面临的问题。

3.3 文本分类

定义3 分类就是在定形或不定形的类别体系框架下，处理客观对象、抽象概念、假设等模式的样例数据，选出与其接近的模型类别归类。文本分类就是将文本对象分成不同的类别。

模式分类是机器学习的一种重要研究领域，它的技术理论几乎可以不加太多修饰地运用到文本分类中。文献[43]从分类的模型、算法和评测等方面对文本分类技术的研究进展进行综述评论。数据挖掘技术是近年来迅速发展的一种专门从海量数据中挖掘事先不知道的、潜在的、有用的规则或知识的技术。文献[44-48]将文本中的词条看成事务项，根据事务项的频繁度对文本进行分类，取得了较好的效果。文献[50]利用信息熵理论对文本进行分类，文献[49]综合利用机器学习的K-Means与数据挖掘中的频繁项集理论对文本进行分类。

不管采用什么技术，文本分类的目的都是便于对文本数据进行组织管理和查询。虽然目前取得了许多进步，但还有许多有待完善和解决的问题，比如数据集偏斜、多层分类、Web页面分类以及算法的扩展性，都是文本分类研究所需要解决的问题。如文献[51]专门研究解决基于KNN文本分类的偏斜问题。

3.4 文本信息检索

广义的信息检索(Information Retrieval)是指信息按一定的方式组织起来，并根据信息用户的需要找出有关的信息的过程和技术。狭义的信息检索是指从信息集合中找出所需要的信息的过程，相当于人们通常所说的信息查寻(Information Search)。

信息的载体有文本、音频、视频、图形、图像等，文本信息检索是指以文本信息为检索对象，不仅仅包括文本中的信息的查找，还应该包括文本的存储。这里所说的文本存储不仅仅是文本本身的存储，还包括经过对文本信息处理后的特征信息存储，比如文本的主题、作者、分类、存储时间等。目前对文本的存储组织研究少见，但它随着网络技术的发展、文本数据的日益积聚，文本的存储组织会引起人们的重视。文本是一种非结构化数据，但通过对文本的处理，比如文本的相关信息抽取、关键词抽取、主题信息提取等将其转换为结构或半结构化数据，从而使得文本从数据量上得到压缩，从存储方式上变成结构化，方便文本信息需求者的查找。文献[52]试图构造一个抽取器，将网络中的文本信息转换成半结构化信息。

3.5 文本的自动摘要

文本的自动摘要就是通过设计一个文本信息处理系统自动提取文本的主题思想。

文本信息处理的目的不仅仅是把握文本中细节信息，更要掌握文本信息的主题思想。文本的主题思想的理解是人的高级智能活动。但人工智能的终极目的就是希望将人的智能活动由机器完成，毕竟文本信息中隐含着众多特征。文本信息处理的自动摘要研究就是希望发现这些字里行间的特征信息，通过特征信息提炼成主题思想。目前虽然在这方面所取得的成果不太如意，但也做了不少工作。如文献[53,54]从文本中词与词同时出现的频率提取文本的主题；文献[55]从文本中词的聚类特征提取文本的主题；文献[56]根据关联规则提取关键词；文献[57]根据文本语义提取主题词。但所有这些工作都还局限在文本信息的表层处理，未能达到文本信息的归纳推理提取主题思想。其主要原因是文本自身灵活多变，从而使得文本信息挖掘的计算模型匮乏。文本分词不精确也在很大程度上制约了文本信息自动摘要技术的发展。

4 文本信息处理的关键技术

人类社会有史以来，文字是其文明继承和发展的主要载体，尤其在图形、图像、音频和视频作为载体以前，文字是唯一表达人们的思想、情感的有形工具，因为其灵活的表现形式、超强的表达能力以及丰富的思想内容，成为千古不衰的信息传递工具。但是文字作为信息载体是一个高度智能化的过程，高效地驾驭文字需要高层次的智能水平，所以文本信息处理技术要达到人类智能化程度是一个漫长的过程，它是计算机科学家与文字语言学家所共同面临和解决的问题。我们这里从计算技术角度将文本信息处理的关键技术以及挑战归纳为分本分词技术、分本存储技术、语义分析和归纳推理等方面。

4.1 分词技术

词由字构成，词是文本的有机组成部分，对文本信息的任何处理都离不开词。分词是文本信息处理的基石，分词技术的好坏直接影响文本信息的抽取、主题的提取、段落的理解、更进一步的自动摘要和文本中心思想的归纳以及隐藏在文本信息之中的知识挖掘。目前的中文分词技术存在两大难题：词典的不完备性和切分歧义性。

对于歧义消解，虽然做了许多工作，但效果依然不理想，比如将包含组合型歧义字段“将来”、“才能”、“中长期”的句子“市长将来我们学校考察工作”、“人才能推动科技进步”、“这是国际共产主义运动中长期没有解决的一个重大理论问题”在北大计算机语言研究所的分词测试平台上[58]和猎兔分词平台测试[59]都不能获取正确的分词结果。对包含交集型歧

义字段的句子“每前进一步都要付出一定代价，避免暴力活动在大选前进一步升级”、“我看主要是你的问题，主要是再不显灵我们就没救了”、“食品加工厂负责人参加了会议，食品加工厂负责人参加密集的工序”进行测试依然不能获得正确的切分结果。

文献[9]统计，用一个含有70000个词的词典去切分含有15000个词的语料库，仍然有30%以上的词条不在词典之中，基于词典的分词方法的效果因此受到极大的限制。

另外，深层的、能体现文本语意的分词技术研究还比较欠缺。比如“信息抽取研究综述”切分成“信息”、“抽取”、“研究”、“综述”4个词，还是切分成“信息抽取研究”、“综述”两个词或切分成“信息抽取”、“研究综述”，在不同语境中有不同的切分方式。如何使得切分结果最贴近语意，这就存在一个切分方式与语意贴近距离度量问题的研究。如果找到一个能评价分词方式与语意贴近度度量的计算模型，使得切分方式最贴近文本语义，随后诸如文本的信息抽取、主题提取、自动摘要，甚至语义理解和归纳推理等工作就随之简单了。

因此分词歧义的消除、词典的完备以及切分方式如何贴近语义，是分词研究中的3大关键技术。目前这3方面的研究工作做了不少，但离文本信息的智能处理还有很大的差距。尤其是贴近语义的分词研究。

4.2 文本的结构化存储处理

计算机对数据处理的强势表现在结构化数据处理，而文本信息一直以来都是一种非结构化的存储数据。随着网络中的文本信息的急剧增长，网络文本信息的有效处理能力日益低下，因此如何有效地将非结构化文本数据转化为结构化存储，成为一种新的应用研究领域。

XML是为将非结构化文档进行结构化处理而开发的一种简单的数据存储语言，但它只能对文档做半结构化处理存储在网络中，还不能将现有的文本文档做结构化转化处理存储。由于任何事物、事件或对象都具有依附其上的某种特性，比如文本中的人名、地名、公司名称、组织名称等命名实体就拥有实体的编号、时间、住址、电话等属性，实体跟实体间存在一定的关系，单个或数个实体间会发生一系列的事件及实体发生需要的场景。在贴近语义的分词基础上从文本中提取实体及其实体间的关系，并提取相应实体的属性，再通过句法分析得到某种结构表示，比如树、图、表等结构化形式，结构化存储文本中的事物、事件或对象。这些外在和内含属性是文本内在含义的反遇，将非结构化的文本数据从语义上肢解成具有一定结构化形式的事物、事件或对象等，从而达到将非结构化的文本数据转换为结构化存储的目的，进一步提高文本信息的有效处理能力。

在网络环境下的信息时代中，针对计算机的特点，为提高计算机对文本数据处理的能力，在努力提高计算机处理非结构化数据能力的同时，将非结构化的文本数据转化为结构化存储，将是文本信息处理的一个具有挑战性的技术领域，对Web信息检索技术具有革新性的重要意义。

4.3 语义分析

自然语言理解的高级智能活动就是语义分析。不同智能水平的人对同一段文字的理解差异很大，因此文本信息的语义分析是人工智能的高级境界。没有文本语义的高级智能分析，不仅仅严重制约文本理解的深入和透切，更是制约诸如Web语义、图像、图形等信息的智能理解，因为所有这些信息载体中都或多或少附带一定文字说明。

目前，文本语义分析研究从国内外的文献资料来看显得很匮乏。现有的关于基于语义分析的文献资料仅仅停留在一定文档模型基础上，对某些满足一定结构的字、词做一定的统计处理，并不是从文本本身的内在定义和外延进行分析处理，更未从文档中所涉及的对象及其对象属性和对象间的关系入手进行分析处理。当然文本的语义分析是计算机学家和语言学家所共同面临的问题，而这一问题的解决对信息处理的智能化程度有着不可估量的促进作用。

从文本所涉及的字词入手，分析其内涵和外延，从分文所涉及的事物、事件或对象入手，分析其关联和变迁，这样的文本语义分析将是文本信息处理的一个新的具有挑战性的综合研究领域，它的实现将是对语言文字人工智能化理解的具体体现。

4.4 归纳推理

归纳推理包含归纳和推理两个过程。所谓归纳就是一种由个别到一般的概括，包括句子的归纳、段落大意的归纳和短文主题的归纳等。推理就是在正确理解文本的语义前提下，透过字里行间推理言外之意及作者的语气、态度。文本信息处理中，计算机的推理像人工一样，能由表及里、由此及彼，从字里行间推理出合乎作者本意的言外之意。推理的结果既来自于字里行间，又高于字里行间；既符合原文实际内涵，又超越实际内涵；既基于已知事实，又不仅仅是已知事实。归纳推理是文本信息理解的最高境界，从人工对文字理解的角度来说依然是一个高难的智能过程。但我们认为，计算机在对文本做正确的分词和文本结构化处理的基础上，进行正确的语义分析，然后从文本信息中归纳出一些有广泛共识的结果是可行的，也是可能的。在归纳的基础上借助机器学习的技术进行推理不是不可能的，毕竟计算机有强大的信息处理能力以及丰富的计算模型，还有人们对科学的不断进步的追求精神。

我们认为文本的语义分析和归纳推理是文本信息处理的最高目标也是终极目标，是文本信息处理中高级人工智能的技术体现，应进一步为文本和语言间的转换提供良好的接口。

结束语 文本是历史悠久、应用广泛、使用灵活、认可度最高的信息载体，即使在信息时代的今天它依然是人类社会文明的继承与文明的发展的主要载体。然而信息的其它载体，比如图形、图像、音频和视频也离不开文本说明。文本信息处理就是通过计算机对文本从表及里，从此及彼的分析处理。本文对文本信息处理研究进行了全面的分析探讨，首先对文本信息处理发展历史做了一定的概述，然后将文本信息处理研究归纳为分词研究、文本信息抽取、文本分类、文本信息检索、文本自动摘要等方面，并分别对各自的研究现状做了概述，指出了各研究领域中存在的问题。最后讨论了文本信息处理的关键技术问题及其挑战，指出文本信息处理的远景目标就是文本信息的语义分析、归纳推理以及文本和语言间的转换。

参考文献

[1] Dejong G. An Overview of the FRUMP System[C]// Proc. of the 5th International Joint Conference on Artificial Intelligence. Cambridge, MA, 1982: 149-176

[2] 李保利，陈玉忠，俞士汉. 信息抽取研究综述[J]. 计算机工程与应用，2003，39(10)：1-5

[3] 周顺先. 文本信息抽取模型及算法研究[D]. 长沙：湖南大学，

2007
[4] Sager N. Natural Language Information Processing. Reading [M]. Addison, Massaehusetts: Wesley, 1981
[5] 吴立德,黄萤菩. 文本检索会议简介[J]. 计算机科学,2002,29(12):89-92
[6] http://projects.ldc.upenn.edu/ace/intro.html
[7] 杨尔弘,方莹,等. 汉语自动分词和词性标注评测[J]. 中文信息学报,2005,20(1):44,49
[8] http://www.sighan.org/
[9] Chien Lee-Feng. PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval[J]. Information Processing and Management,1999,35:501-521
[10] 张恒,杨文昭,屈景辉,等. 基于词典和词频的中文分词方法[J]. 微计算机信息,2008,12:239-241
[11] 孙茂松,左正平,邹嘉彦. 高频最大交集型歧义切分字段在汉语自动分词中的作用[J]. 中文信息学报,1999,13(1):27-34
[12] 孙茂松,黄昌宁,邹嘉彦. 利用汉字二元语法关系解决汉语自动分词中的交集型歧义[J]. 计算机研究与发展,1997,34(5):332-339
[13] Zhou Qiang. Chinese POS tagging method with rules and statistics combined[J]. Journal of Chinese Information Processing, 1995,9(3):1-10
[14] 傅赛香,袁鼎荣,黄柏雄,等. 基于统计的无词典分词方法[J]. 广西科学院学报,2002,18(4):252-256
[15] 张恒,杨文昭,屈景辉,等. 基于词典和词频的中文分词方法[J]. 微计算机信息,2008,12:239-241
[16] 费洪晓,康松林,朱小娟,等. 基于词频统计的中文分词的研究[J]. 计算机工程与应,2005,7:67-69
[17] 李素建,刘群,杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报,2003,26(12):1722-1726
[18] 姜韶华,党延忠,宣照国. 无词典抽词的 RMMFS 和 BMMFS 方法及其比较研究[J]. 情报学报,2006,25(4):499-503
[19] 韩客松,王永成,陈桂林. 无词典高频字串快速提取和统计算法研究[J]. 中文信息学报,2000,15(2):23-30
[20] 王大玲,于戈,鲍玉斌. 基于最长顺序频繁词组的 Web 文献检索结构[J]. 软件学报,2006,17(10):2096-2106
[21] 金翔宇,孙正兴,等. 一种中文文档的非受限无词典抽词方法[J]. 中文信息学报,2001,15(6):33-39
[22] 李素建,刘群,杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报,2003,26(12):1722-1726
[23] 罗盛芬,孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报,2003,17(3):9-14
[24] 李中国,刘颖. 边界模板和局部统计相结合的中国人名识别[J]. 中文信息学报,2006,20(5):44-50
[25] 洪铭材,张阔,唐杰,等. 基于条件随机场(CRFs) 中文词性标注方法[J]. 计算机科学,2006,33(10):245-151
[26] 罗盛芬,孙茂松. 基于字串内部结合紧密度的汉语自动抽词实验研究[J]. 中文信息学报,2003,17(3):9-14
[27] 罗智勇,宋柔. 一种基于可信度的中国人名识别方法[J]. 中文信息学报,2005,19(3):67-72,86
[28] 郑家恒,李鑫,谭红叶. 基于语料库的中文姓名识别方法研究[J]. 中文信息学报,2000,14(1):7-12
[29] 张锋,樊孝忠,许云. 基于统计的中文姓名识别方法研究[J]. 计算机工程与应用,2004(10):53-54,77
[30] 张华平,刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报,2004,27(1):85-91
[31] 李中国,刘颖. 边界模板和局部统计相结合的中国人名识别[J]. 中文信息学报,2006,20(5):44-50
[32] 周强,孙茂松,黄昌宁. 汉语最长名词短语的自动识别[J]. 软件学报,2000,11(2):195-201
[33] 周雅倩,郭以昆,黄萱菁,等. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展,2003,40(3):440-446
[34] 韩洁,周勇,刘少辉,等. 基于 WWW 的未登录词识别研究[J]. 计算机科学,2002,29(12):155-156
[35] 邹纲,刘洋,刘群,等. 面向 Internet 的中文新词语检测[J]. 中文信息学报,2004,18(6):1-9
[36] Chen Keh-jiann, Ma Wei-yun. Unknown Word Extraction for Chinese documents[C]// Proceedings of COL-ING. 2002:169-175
[37] Hobbs J R. Information extraction from biomedical text [J]. Journal of Biomedical Informatics,2002,35(4):260-264
[38] Kongachandra R, Kimpant C, Suwanapong T, et al. Newly-born keyword extraction under limited knowledge resources based on sentence similarity verification[C]// IEEE International Symposium on Communications and Information Technology. 2004: 1183-1187
[39] 高俊波,栾翠菊,王晓峰. 新的关键字提取算法研究[J]. 计算机工程与设计,2008,29(3):765-767
[40] Sproat R, Shih C. A statistical method for finding word boundaries in Chinese text[J]. Computer Processing of Chinese and Oriental Languages,1990,4(4):336-351
[41] Ge Xianping, Pratt W, Smyth P. Discovering Chinese Words from Unsegmented Text[A]// SIGIR[C]. Berkeley: ACM, 1999:271-272
[42] 任禾,曾隽芳. 一种基于信息熵的中文高频词抽取算法[J]. 中文信息学报,2006,22(5):40-45
[43] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1861
[44] Hynek J, Jezek K, Rohlik O. Short document categorization-itemsets method[C]// Lyon, France: PKDD 4th European Conference on Principles and Practice of Knowledge Discovery in Databases. Workshop Machine Learning and Textual Information Access,2000:14-19
[45] 王永恒,贾焰,杨树强. 基于频繁词集聚类的海量短文分类方法[J]. 计算机工程与设计,2007,28(8):1744-1748
[46] Beil F, Ester M, Xu X. Frequent Term-based Text Clustering [C]// Proc of the 8th Int'l Conf on Knowledge Discovery and Data Mining. 2002:436-442
[47] Fung B C M, Wang K, Ester M. Hierarchical Document Clustering Using Frequent Itemsets [C] // Proc of SDM'03. 2003:59-70
[48] Zhuang L, Dai Honghua. A Maximal Frequent Itemset Approach for Web Document Clustering[C]// Proc of the 4th Int'l Conf on Computer and Information Technology. 2004:970-977
[49] 王乐,田李,贾焰,等. 基于频繁词集和 k-Means 的 Web 文本聚类混合算法[J]. 计算机工程与科学,2008,30(8):92-97
[50] 李陆荣,王建会,陈晓芸,等. 使用最大熵模型进行中文文本分类[J]. 计算机研究与发展,2005,42(1):94-101
[51] 郝秀兰,陶晓鹏,徐和祥,等. kNN 文本分类器类偏斜问题的一种处理对策[J]. 计算机研究与发展,2009,46(1):52-62
[52] 黄豫清,戚广志,张福炎. 从 Web 文档中构造半结构化信息的抽取器[J]. 软件学报,2000,11(1):73-78
[53] 马颖华,王永成,苏贵洋,等. 一种基于字同现频率的汉语文本主题抽取方法[J]. 计算机研究与发展,2003,40(6):876-870
[54] 余刚,陈华月,朱征宇,等. 基于词同现频率的文本特征描述[J]. 计算机工程与设计,2005,26(8):2180-2183
[55] 陈炯,张永奎. 一种基于词聚类的中文文本主题抽取方法[J]. 计算机应用,2005,25(4):754-757
[56] 李钝,曹元大,万月亮. 基于关联规则的安全特色关键词提取研究[J]. 计算机工程与应用,2006(S1)
[57] 唐培丽,王树明,胡明. 基于语义的汉语文献主题词提取算法研究[J]. 吉林大学学报:信息科学版,2005,23(5):535-540
[58] http://www.icl.pku.edu.cn/
[59] http://www.lietu.com/SCSeg.jsp