中文信息处理中自动分词技术的研究与展望

刘　迁　贾惠波
（清华大学精密仪器与机械学系，北京 100084）
（清华大学光盘国家工程研究中心，北京 100084）

摘　要　汉语自动分词是中文信息处理的关键技术，已经成为中文信息处理发展的瓶颈。文章介绍了当前自动分词技术的研究状况，对各种分词算法进行了介绍，并对各种算法进行了比较和讨论。最后，对汉语自动分词技术的发展进行了展望。

关键词　中文信息处理　汉语自动分词　分词算法

文章编号 1002-8331-(2006)03-0175-03　文献标识码 A　中图分类号 TP391

A View of Chinese Word Automatic Segmentation Research in the Chinese Information Disposal

Liu Qian　Jia Huibo
(Department of Precision Instruments and Mechanology, Tsinghua University, Beijing 100084)
(Optical Memory National Engineering Research Center, Tsinghua University, Beijing 100084)

Abstract: Chinese word automatic segmentation is the key technology in the Chinese information disposal.In this paper we present the work of the Chinese word automatic segmentation.Through analyzing and comparing the segmentation methods in existence, we point out the developing directions of future Chinese word automatic segmentation.

Keywords: Chinese information disposal, Chinese word automatic segmentation, segmentation methods

1　引言

随着信息时代的到来，计算机在人们生产生活的各个方面将起着越来越大的作用。而对于以汉语为母语的我国来说，中文信息处理技术已经在我国信息化建设中占据了一个非常重要的地位。

中文信息处理是一门以计算机对中文（包括口语和书面语）进行转换、传输、存贮、分析等加工的科学。中文信息处理可概括地分为三个平台：字处理平台、词处理平台和句处理平台，其中的每个平台都是以前者为基础的。字处理平台技术是中文信息处理的基础，经过近 20 年的研究，字平台技术已经达到了一个比较成熟的阶段；词处理平台技术是中文信息处理的中间环节，它是连接字平台和句平台的关键纽带，因此也是关键环节；句处理平台技术是中文信息处理的高级阶段。它的研究主要包括：机器翻译、汉语的人机对话。这方面的研究虽然已取得了一定的成果，但是目前还处于初步阶段。

中文信息处理中词平台以上的技术都要以“词”为基础，但汉语书面语不是像西方文字那样通过天然的切分标志——空格分开的，而是在语句中以汉字为单位一个挨一个地连写，词与词之间没有明显的界限，进入计算机后是等距排列的汉字串序列，没有词的切分标志。因此，如果将句子中的“词”自动切分出来，即让计算机能够辨别哪几个汉字结合起来是一个词，让计算机将等间距排列的汉字字串序列按词切分开，并打上切分标志，这就是中文文本自动分词问题。中文文本自动分词成为了中文信息处理中特有的基础性问题。

2　自动分词技术的研究现状及面临的困难

随着中文信息处理研究的深入，中文文本自动分词问题已经引起相当程度的重视，成为中文信息处理的一个前沿课题。经过十几年的研究，中文文本自动分词技术取得了令人瞩目的成果，出现了一些实用的自动分词系统[1-3]，这些系统在分词的精确度（精度达到 99%以上）和分词速度（速度达到千字/s）方面都具有相当的水平，但是同时在速度和精确度方面都仍然需要进一步的研究。汉语自动分词技术面临着以下三个方面的困难：汉语中“词”的概念缺乏清晰的界定；未登录词的识别；歧义切分字段的处理。

首先，在 1992 年制定的《信息处理用现代汉语分词规范》，作为国家标准为中文信息处理提供了一个实用的分词原则。目前需要制订一个标准化、实例化的分词词表，以便于按这一《规范》进行分词操作，并在运行过程中，对《规范》进行检验、修订和完善。由于汉语的复杂性和特殊性，这一词表的制定过程将是一项艰巨的任务。

由于词典的容量毕竟是有限的，所以文本中必然会存在词典中没有收录的词，如人名、地名、机构名等专有名词及新词语等，我们把这些词称为未登录词。未登录词是自动分词中一个重要的问题，在新闻类文本中十分突出。该问题的解决有赖于我们对汉语结构的进一步认识。标准化分词词表的建立和完善将对该问题的解决有一定的改善。

由于汉语的连续书写习惯，无论分词规范多么详细，分词词表多么完善，汉语自动分词中的歧义问题都将始终存在，并且将严重影响着分词系统的切分精度。自动分词中歧义字段的问题成为自动分词系统在实际应用中最大的障碍，因此学者们纷纷将歧义字段的处理作为各自算法研究的重点。歧义切分字段的处理成为中文自动分词技术的关键问题。

3 歧义字段的定义及分类

汉语自动分词过程中出现具有多种切分可能的字段，我们称其为歧义字段。歧义字段可分为[4]：交集型歧义字段(OAS)、和覆盖型歧义字段(CAS)，又称包孕型、多义型或组合型歧义字段。

(1)交集型歧义字段(OAS)：假设 A,B,C 分别代表由一个或多个字组成的字串，如果在 ABC 字段中 A,AB,BC,C 分别都是分词词表中的词，则称该字段为交集型歧义字段。例如：字段“美国会”它可产生“美/国会”和“美国/会”两种切分结果，因此属于交集型歧义字段。据统计[1]，交集型歧义字段占全部歧义字段的 85%~90%，所以交集型歧义字段是自动分词系统需要重点加以解决的问题。

(2)覆盖型歧义字段(CAS)：假设 A,B,分别代表由一个或多个字组成的字串，如果 A,B,AB 分别都是分词词表中的词，则称 AB 为组合型歧义字段。例如：“把/手”、“十/分”等。虽然组合歧义字段在文本中出现的机会并不多，但是潜在的组合歧义字段却不容忽视。例如：“美德”在一般的文本中都是作为一个名词出现的，但是在新闻语料中经常出现字段“中俄美德四强”，这时“美/德”就变成了组合型歧义字段。

4 中文自动分词算法分类及比较

4.1 基于机械匹配的中文自动分词算法

最大匹配法(Maximum Matching Method，简称 MM 法)是在 20 世纪 50 年代末由苏联专家提出的，最早出现的一种自动分词算法。该算法的基本思想是：事先建立词库，其中包含所有可能出现的词。对给定的待分词的汉字串 s，按照某种确定的原则(正向或逆向)取 s 的子串，若该子串与词库中的某词条相匹配，则该子串是词，继续分割剩余的部分，直到剩余部分为空；否则，该子串不是词，则取 s 的子串进行匹配。MM 法是一种得到广泛应用的机械分词方法，这里的“机械”是因为该算法仅仅依靠分词词表进行匹配分词。

根据每次匹配时优先考虑长词还是短词，机械分词法又分为最大匹配法和最小匹配法；根据扫描词表的方向和截取字时的增字还是减字又分为正向 MM 法(FMM)、反向 MM 法(BMM)、增字 MM 法和减字 MM 法。最大匹配法比较常用，它假设自动分词词典中的最长词条所含汉字个数为 i，则取待处理材料当前字符串序列中的前 i 个字作为匹配字段，查找分词词典进行匹配。

机械匹配算法简洁、易于实现，其中的代表算法——最大匹配法，体现了长词优先的原则，在实际工程中应用最为广泛。机械匹配算法实现比较简单，但其局限也是很明显的：效率和准确性受到词库容量的约束；机械匹配算法采用简单机械的分词策略，不涉及语法和语义知识，所以对于歧义切分无法有效地克服，切分精度不高。虽然专家们采用了不少方法来改善机械匹配的性能，但是从整体效果上来看，单纯采用机械匹配式进行分词难以满足中文信息处理中对汉语分词的要求。在机械匹配分词的基础上，利用各种语言信息进行歧义校正，是削弱机械式切分局限性的一种重要手段。

目前，出现了许多机械匹配与其他切分歧义处理方法相结合的中文自动分词算法。其中包括运用规则、语法、语义知识进行歧义处理方法：

(1)基于标记法[5]：利用显式切分标记(标点、数字、西文等其它非汉字符号)和隐式切分标记(出现频率高，结构能力差的单字词)将文本预先切分成汉字短串序列，然后再进行匹配切分。

(2)约束矩阵法[6]：首先建立语法、语义约束矩阵，然后利用相邻词汇之间的约束关系来进行分词。

(3)句模切分法[7]：将汉语句模理论应用到分词算法当中，首先确定待切分字段的动核类型，然后从动核结构表中找到该类型动核所组成的所有可能的句模结构。将所检验的切分结果与可能的句模结构逐一进行比较进行歧义处理。

另外也出现了一些运用统计语言模型进行歧义处理的匹配算法，包括：

(1)利用互信息、t 测试差进行歧义处理[8]：首先，利用词典进行正向及反向最大匹配分词；对正向及反向最大匹配所得出的两种不同的切分方案，分别计算其互信息及 t-信息，然后进行歧义处理。

(2)等同于分类的方法[9,10]：将切分中歧义字段的处理问题形式化为一种分类问题。由于将问题抽象为一种分类问题，因此许多机器学习和模式识别中有关解决分类问题的方法都可以在歧义处理中使用。

4.2 基于统计语言模型(SLM)的中文自动分词方法

随着中文电子文本的增多，越来越多的学者认识到，唾手可得的海量电子文本应成为自动分词的重要资源，利用机器学习手段从生语料库中直接获取分词所需的某些适用知识则应成为自动分词的重要补充手段，因此就产生了基于统计语言模型(SLM，Statistical Language Models)的分词算法[11]，又称为无词表分词算法。该类算法的主要思想是：词是稳定的汉字的组合，在上下文中汉字与汉字相邻共现的概率能够较好地反映成词的可信度，因此对语料中相邻共现的汉字的组合频度进行统计，计算他们的统计信息并作为分词的依据。

假设变量 W 代表一个文本中顺序排列的 n 个词，即 $W=w_1w_2\cdots w_n$，则统计语言模型的任务是给出任意词序列 W 在文本中出现的概率 $P(W)$。利用概率的乘积公式，$P(W)$ 可展开为：

$$P(W)=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_n|w_1w_2\cdots w_{n-1})$$

不难看出，为了预测词 w_n 的出现概率，必须已知它前面所有词的出现概率。从计算上来讲，这是一个十分复杂的过程。如果任意一个词 w_i 的出现概率只同它前面的 $N-1$ 个词有关，问题就可以得到很大的简化。这个简化了的模型就叫作 N 元语言模型(N-gram)，即：

$$P(W)=P(w_1)P(w_2|w_1)P(w_3|w_1w_2)\cdots P(w_i|w_{i-N+1}\cdots w_{i-1})\cdots\approx$$

$$\prod_{i=1,\cdots,n}P(w_i|w_{i-N+1}\cdots w_{i-1})$$

符号 $\prod_{i=1,\cdots,n}P(\cdots)$ 表示概率的连乘。

N 元语言模型是一种常用的统计语言模型。而在中文分词中实际使用的通常是 N=1、2 或 3 的一元模型(uni-gram)、二元模型(bi-gram)或三元模型(tri-gram)。以三元模型为例，近似认为任意词 w_i 的出现概率只同它紧前面的两个词有关，即：

$$P(W)\approx\prod_{i=1,\cdots,n}P(w_i|w_{i-2}w_{i-1})$$

$$P(w_i|w_{i-2}w_{i-1})\approx\frac{count(w_{i-2}w_{i-1}w_i)}{count(w_{i-2}w_{i-1})}$$

式中 $count(\cdots)$ 表示一个特定词序列在整个语料库中出现的累计次数，这些概率参数都是可以通过大规模语料库来估值的。

此类算法的关键是在统计量的选取上，这方面早期的工作包括：文献[12]利用相对简单的汉字串的串频统计信息；而文献[13]利用“互信息”定量描述任意两个汉字之间的结合力。

文献[9]在互信息的基础上引入了汉字间“t 测试差”作为互信息的有益补充，并且将两者线性叠加得到统计量 md[14]，结合所引入的峰和谷的概念，设计了一种无词表的自动分词算法。该算法所需的所有数据均是从生语料中自动获得的，所以对不同领域有着较好的适应。

此外，文献[15]借助 $\chi 2$ 统计量和广义似然比方法计算汉字之间的相关度，对长度分别为 2、3 和 4 的任意汉字串做内部关联性分析，从语料库中获得未登录词，用于自动分词，同时也可用于机器自动编制词典。

综上所述，基于统计模型的自动分词算法的优点在于：该类算法所需的一切数据均由机器从生语料中自动获得，无须人工介入。能够有效地自动排除歧义，能够识别未登录词，解决了机械匹配分词算法的局限。但是由于该类算法不使用分词词表，所以对常用词的识别敏感度较低，时空开销较大，并且会抽出一些共现频度高但并不是词的常用词组，例如“这是”、“有的”。

4.3 基于人工智能技术的中文自动分词方法

应用人工智能中的神经网络和专家系统来进行中文自动分词，以实现智能化的中文自动分词系统是近年来中文自动分词领域中的一个研究热点。该类算法的分词过程是对人脑思维方式的模拟，试图用数字模型来逼近人们对语言认识的过程。

(1)神经网络分词算法[16]

该类分词算法是以模拟人脑运行，分布处理和建立数值计算模型工作的。它将分词知识的隐式方法存入神经网内部，通过自学习和训练修改内部权值，以达到正确的分词结果。

神经网络分词法的关键在于知识库(权重链表)的组织和网络推理机制的建立。算法的分词过程是一个生成分词动态网的过程，该过程是分步进行的：首先以确定的待处理语句的汉字串为基础，来确定网络处理单元；然后，根据链接权重表激活输入/输出单元之间的链接，该过程可以采用某种激活方式，取一个汉字作为关键字，确定其链接表，不断匹配。

神经网络分词法具有自学习、自组织功能，可以进行并行、非线性处理，并且反应迅速、对外界变化敏感；但是目前的基于神经网络的分词算法存在着网络模型表达复杂，学习算法收敛速度较慢，训练时间长，并且对已有的知识维护更新困难等不足。

(2)专家系统分词算法[17]

专家系统分词算法从模拟人脑功能出发，构造推理网络，将分词过程看作是知识推理过程。该方法将分词所需要的语法、语义以及句法知识从系统的结构和功能上分离出来，将知识的表示、知识库的逻辑结构与维护作为首要考虑的问题。知识库按常识性知识与启发性知识分别进行组织。知识库是专家系统具有“智能”的关键性部件。

专家系统分词算法是一种统一的分词算法，不仅使整个分词处理过程简明，也使整个系统的运行效率得到提高。该算法具有显式知识表达形式，知识容易维护，能对推理行为进行解释，并可利用深层知识来处理歧义字段，其切分精度据称可达语法级；其缺点是不能从经验中学习，当知识库庞大时难以维护，进行多歧义字段切分时耗时较长，同时对于外界的信息变化反应缓慢，不敏感。

4.4 三种自动分词方法的比较

上述三类中文自动分词方法分别代表目前分词方法的发展方向。基于机械匹配的分词方法出现较早，该方法简洁、易于实现，在工程上得到了广泛的应用；该方法切分精度不高，对于切分歧义无法有效地克服。单纯采用机械匹配式进行分词难以满足中文信息处理中对汉语分词的要求。因此将机械匹配式和其它切分方法相结合，来提高机械匹配分词对于切分歧义的处理能力，是目前中文自动分词方法研究的一个比较成熟的发展方向。

基于统计语言模型的分词方法由于具有良好的切分歧义处理能力和识别新词的能力，目前受到了越来越多的研究人员的重视，发展较快。而如何将该类分词方法与基于机械匹配的分词机制有机地结合起来，既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文来自动消除歧义的优点，已经成为该类算法的下一步研究课题。

基于人工智能技术的神经网络分词方法和专家系统分词方法是理论上最理想的分词方法，但是由于该类分词方法的研究还处于初级阶段，并且由于汉语自然语言复杂灵活，知识表示困难，所以对于基于人工智能的中文自动分词技术还需要进行更深入和全面的研究。虽然目前还处于起步阶段，但是该类分词方法是未来中文自动分词方法的发展方向。

5 结论

由于西文文本信息处理天然地就在词平面上。而汉语文本起步是在字平面上，落后西文一个层次。这一个层次的差异是本质上的、全局性的，如果解决不好，中文信息处理将在整体上永远处于低水平，无法向高级形态发展。所以中文文本自动分词问题已经成为制约中文信息处理发展的最大瓶颈。

中文文本自动分词问题的研究主要分为以下三个方面：一是对分词词表的建立和完善。这里包括建立通用的核心词表，构造各个领域的基本专业词表。并且要进行基于上述各分词词表的分词歧义穷尽式调研。这方面的工作需要各个领域的专家同心协力、达成共识；二是关于分词方法的研究，包括加强对汉字串统计性质的研究，分词策略及分词算法的研究，以及如何更加有效地解决切分歧义以及未登录词的问题；三是研究汉语的特点和规律，从汉语的书写规则出发来寻求中文自动分词的突破口。在建立一系列规范的基础上，书写或录入时在词与词之间增加分隔符，使计算机自动识别和切分，这是一种毕其功于一役的自动分词方法，它解决了中文文本信息处理落后于西文一个层次的问题，可以将中文文本信息处理的起步点也建立在词平面上，这种方法将是中文自动分词未来新的方向。

综上所述，解决中文文本自动分词问题已经成为中文信息处理当前的一项战略任务，它已经同让世界了解汉语、中华民族文化的伟大复兴紧密地联系在一起。该任务具有相当的紧迫性和必要性，需要语言学、计算机语言、自然语言处理等多方面的专业人士共同努力来完成。(收稿日期：2005 年 7 月)

(下转 182 页)

息，继续对 MFS_2 中节点应用降维策略进行剪枝和降维，产生后继节点集 MFS_3。在 MFS_3 中，由于非频繁项集的维数不大于频繁项集的维数，即 $card(\{bad\}) \leqslant card(\{bae\}) = card(\{bac\})$，则无需继续搜索 CIE-树，则项集 $\{bae\}$、$\{bac\}$ 为最大频繁项集。

3 算法的分析与比较

CIE-树可以尽快发现二维非频繁项集，并充分利用这些信息，减少扫描数据库的次数、快速生成候选最大频繁项目集；结合数据库的分块，使得必要的数据库的扫描只需在较小的数据块中进行，进一步提高了算法的效率。

Max-Miner 算法无法利用非频繁项目集的信息，以至生成很多不必要的候选项目集。P&M 算法不能在第一时间发现非频繁项目集，使得该信息的利用较晚。如候选最大频繁项目集 $ABCDEF$，假设 $A>B>C>D>E>F$，DE 是二维非频繁项目集，并且 $ABCDF$ 和 $ABCEF$ 是最大频繁项目集，在不考虑其它因素的情况下，Max-Miner 算法要搜索到集合枚举树第 5 层时，才能生成 $ABCDF$、$ABCEF$；P&M 算法在搜索集合枚举树的第 4 层节点时才生成 $ABCDF$、$ABCEF$；而 BDIF-Ⅱ算法在搜索 CIE-树的第 2 层时首先获得 DE 的非频繁信息，并立刻将其应用到生成第 3 层节点的算法中，使得该算法在搜索 CIE-树的第 2 层时就生成了 $ABCDF$、$ABCEF$。

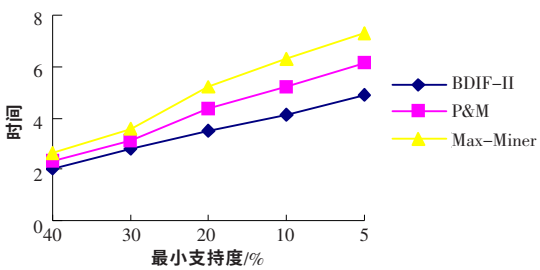

图 4 三种算法的性能比较

我们利用具有 4 000 条记录的数据库，在同一台计算机上实现了 Max-Miner 算法、P&M 算法和 BDIF-Ⅱ算法，图 4 所示为三种算法在不同支持度下算法运行时间的变化，从中可以看出 BDIF-Ⅱ算法性能的优越。

综上所述，本文提出的快速生成最大频繁项目集的算法 BDIF-Ⅱ能快速发现并有效地利用非频繁项目集的信息生成候选最大频繁项目集，并结合数据块的划分，进一步减少了扫描数据库的次数。理论和实验结果表明，相对于其它发现频繁项目集的算法，BDIF-Ⅱ算法有较优越的性能，为相关发现最大频繁项目集的数据挖掘应用提供了一种有效而快速的算法。

（收稿日期：2005 年 4 月）

参考文献

1.Agrawal R，Srikant R.Fast algorithms for mining association rules in large databases[C].In：Bocca J B，Jarke M，Zaniolo C eds.VLDB'94，Proceedings of 20th International Conference on Very Data Bases，Santiagode Chile：Morgan Kaufmann，1994：487~499

2.Jong Soo Park，Ming-Syan Chen，Philip S Yu.An effective Hash-Based Algorithm for Mining Association Rules[C].In：Proc of the ACM SIGMOD Int'1 Conf on Management of Data，San Jose，1995：175~186

3.A Savasere，E Omiecinski，S Navathe.An Efficient Algorithm for Mining Association Rules in Large Databases[C].In：21st VLDB Conf，Zurich，Switzerland，1995-09：432~444

4.Bayardo R.Efficiently mining long patterns from databases[C].In：Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data，New York：ACM Press，1998：85~93

5.刘大有，刘亚波，尹治东.关联规则最大频繁项目集的快速发现算法[J].吉林大学学报（理学版），2004；42（2）

6.李红，黄晓杰，胡学钢.一个改进的关联规则挖掘算法[C].见：全国第 16 届计算机科学与技术应用学术会议，2004

（上接177 页）

参考文献

1.梁南元.书面汉语自动分词系统——CDWS[J].中文信息学报，1987；（2）

2.沈达阳，孙茂松，黄昌宁.汉语自动分词和词性标注一体化系统[J].中文信息，1996；（5）

3.北京大学计算语言学研究所.http：//icl.pku.edu.cn/icl_res/segtag98/，1998

4.孙茂松，邹嘉彦.汉语自动分词研究中的若干理论问题[J].语言文字应用，1995；（4）

5.亢临生，张永奎.基于标记的分词算法[J].山西大学学报（自然科学版），1995；17（3）

6.雷西川，余靖维，卢晓铃.基于相邻知识的汉语自动分词系统研究[J].情报科学，1994；（2）

7.张滨，晏蒲柳，李文翔等.基于汉语句模的中文分词算法[J].计算机工程，2004；（1）

8.孙茂松，黄昌宁，邹嘉彦等.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）

9.李蓉，刘少辉，叶世伟等.基于 SVM 和 K-NN 结合的汉语交集型歧义切分方法[J].中文信息学报，2001；（6）

10.湛燕，陈昊，袁方等.基于文本分类的分词方法研究[J].计算机工程与应用，2003；39（23）：87~88

11.黄昌宁.统计语言模型能做什么[J].语言文字应用，2002；（1）

12.刘挺，吴岩，王开铸.串频统计和词形匹配相结合的汉语自动分词系统[J].中文信息学报，1997；（1）

13.Sproat R，Shih C L.A statistical method for finding word boundaries in Chinese text[J].Computer processing of Chinese and Oriental Language，1993；4（4）

14.孙茂松，肖明，邹嘉彦.基于无指导学习策略的无词表条件下的汉语自动分词[J].计算机学报，2004；（6）

15.黄萱菁，吴立德，王文欣等.基于机器学习的无需人工编制词典的切词系统[J].模式识别与人工智能，1996；9（4）

16.尹锋.基于神经网络的汉语自动分词系统的设计与分析[J].情报学报，1998；（2）

17.王彩荣.汉语自动分词专家系统的设计与实现[J].微处理机，2004；（6）